

The Pioneer Research Journal

*An International Collection of Undergraduate-Level
Research*

Volume 4

2017

Pioneer[®]
academics

The Pioneer Research Journal

*An International Collection of Undergraduate-Level
Research*

Contents

Contributing Readers	iii
Foreword	xiii
Selection Process	xv
Are Hedge Funds a Veblen Good? (Economics)	1
Author: Georgia McKirgan	
School: The Portsmouth Grammar School, Portsmouth, UK	
Pioneer Seminar Topic: Financial Market Behavior and Misbehavior	
Brain Plasticity: The Effects of BrainPort Sensory Substitution (Neuroscience)	25
Author: Elexis Hernandez Sanchez	
School: Dr. Kirk Lewis Career and Technical High School, Houston, Texas, USA	
Pioneer Seminar Topic: Understanding the Sense of Touch through Neuroscience	
Descartes on Other Minds: Human and Animal (Philosophy)	51
Author: Hao Yu	
School: Beijing National Day School, Beijing, China	
Pioneer Seminar Topic: Descartes' Meditations	
Deviation and Integration: How Nüshu Serves as a Centrifugal and Centripetal Force for Women in Changing Rural China (Anthropology)	61
Author: Yilin Chen	
School: The Experimental High School Attached to Beijing Normal University, Beijing, China	
Pioneer Seminar Topic: Communication and Culture	

Divided by Discrimination: An Analysis of Racial Profiling during a Shopping Expedition (Anthropology) 73

Author: Gabrielle Battle

School: The College Preparatory School, Oakland, California, USA

Pioneer Seminar Topic: Nonverbal Communication

Ethical Considerations of State Responsibility toward Refugees: Analyzing China's Refugee Capacity from a Socio-Economic Perspective (International Relations) 85

Author: Xiangyu Zheng

School: International Department of the Affiliated High School of South China Normal University, Guangzhou, China

Normal University, Guangzhou, China

Pioneer Seminar Topic: Globalization and International Migration

Exposing Heat Stress Differential within the Urban Heat Island (Environmental Science) 101

Author: Olivia R. Colombo

School: Sacred Heart High School, San Francisco, California, USA

Pioneer Seminar Topic: The Nexus of Buildings, Energy and the Environment

Fast Video Retargeting Based on Seam Carving with Parental Labeling (Computer Science) 111

Author: Chuning Zhu

School: The Experimental High School Attached to Beijing Normal University, Beijing, China

Pioneer Seminar Topic: Computers That See: Exploring New Techniques of Computer Vision

Home Alone: How the Lack of Chinese Immigrant Women Caused the Failure of Chinese Immigrants to Integrate into the United States, 1848-1882 (Sociology) 125

Author: Felix Hohne

School: St. George's School, Vancouver, British Columbia, Canada

Pioneer Seminar Topic: Hyphenated Americans

Identification of Reliable Predictor of Primary Spontaneous Pneumothorax Recurrence Risk (Biology) 137

Author: Ruibing Xu

School: Shenzhen Foreign Languages School, Shenzhen, China

Pioneer Seminar Topic: Integrative Physiology

Investigating the Characteristics of the Optimal Point for Magnetic Nanoparticle Hyperthermia (Engineering)	145
Author: Ege Özgüroğlu	
School: Robert College, Istanbul, Turkey	
Pioneer Seminar Topic: Introduction to Nanoscience and Nanotechnology	
Kiezdeutsch: A Youth Dialect in the Face of Linguistic Conservatism (Anthropology)	159
Author: Zehan Zhou	
School: St. George's School, Vancouver, British Columbia, Canada	
Pioneer Seminar Topic: Communication and Culture	
Mangrove Forests Mitigate Tsunami Hazard and Require Conservation (Environmental Science)	175
Author: Qiqing Li	
School: Shenzhen Middle School, Shenzhen, China	
Pioneer Seminar Topic: Understanding Earthquakes and Earthquake Prediction	
Microneedle Transdermal Drug Delivery for Traditional Chinese Medicated Plasters (Engineering)	197
Author: Chendan Luo	
School: WHBC of Wuhan Foreign Languages School, Wuhan, China	
Pioneer Seminar Topic: Mechanical Engineering	
r-bonacci Numbers, r-Lucas Numbers and Their Identities (Mathematics)	217
Author: Yuheng Wu	
School: The Affiliated High School of South China Normal University, Guangzhou, China	
Pioneer Seminar Topic: Fibonacci Numbers and Visual Proofs	
Social Networks in Entrepreneurial Opportunity Recognition through Different Methods of Communication (Business)	233
Author: Yuxin Wu	
School: Shanghai Pinghe Bilingual School, Shanghai, China	
Pioneer Seminar Topic: Examining the Factors of High-Growth High-Performance Entrepreneurship: Firms, Founders and Ecosystems	
Sofonisba Anguissola and Her <i>Self-Portrait at the Easel</i> (Art History) . .	249
Author: Xingzhi Jing	
School: Hangzhou Foreign Languages School, Hangzhou, China	
Pioneer Seminar Topic: Methodologies of Art History	

The Effect of Political and Cultural Ideas on Clothing Reform,
Chinese Hair, and Clothing Fashion, 1890-1910's (History) 271

Author: Yanyu Zhong

School: Shenzhen Middle School, Shenzhen, China

Pioneer Seminar Topic: Qing China and the West: Conflict and Cultural Exchange

The Effect of Viewing Psychopathic Characters on Television on the
Development of Psychopathic Traits on Children (Psychology) 283

Author: Selin Baydar

School: Hisar School, Istanbul, Turkey

Pioneer Seminar Topic: Clinical Psychology

The Impact of Emotions, Ambiguity, and Apathy on Decision-Making
in Individuals with Alzheimer's Disease: A Proposal to Facilitate the
Diagnosis of Alzheimer's Disease (Neuroscience) 299

Author: Sammer M. Marzouk

School: The University of Chicago Laboratory Schools, Chicago, Illinois, USA

Pioneer Seminar Topic: The Decision-Making Brain

The Neural Mechanism, Genetic Basis and Possible Treatment of
Procrastination (Neuroscience) 321

Author: Yuyang Sun

School: Shenzhen Middle School, Shenzhen, China

Pioneer Seminar Topic: The Decision-Making Brain

The Power and Significance of Agency in *The Iliad* (Literature) 333

Author: Yumeng Li

School: Beijing National Day School, Beijing, China

Pioneer Seminar Topic: The Hero in Literature and Culture

The Existence of a Knight's Tour on the Surface of Rectangular Boxes
(Mathematics) *

Author: Shengwei Lu

School: Hwa Chong Institution, Republic of Singapore

Pioneer Seminar Topic: Combinatorics and Graph Theory

* *Shengwei Lu has elected to forego publication of his paper in The Pioneer Research Journal, Volume 4, at this time because his paper is under consideration for publication in another research journal.*

Are Hedge Funds a Veblen Good?

Georgia McKirgan

Author Background: Georgia McKirgan grew up in the United Kingdom and currently attends The Portsmouth Grammar School in Portsmouth, UK. Her Pioneer seminar topic was in the field of economics and titled "Financial Market Behavior and Misbehavior."

1. Introduction

Hedge funds have grown to be an important part of the investment universe. Initially they promised outperformance over traditional funds, which was used to justify their high fees. As the hedge fund industry has grown, performance has deteriorated but assets under management have continued to grow. In economics, a ‘normal’ good is one where demand decreases as the price increases. The *Financial Times* lexicon defines a ‘veblen’ good as “*a luxury item whose price does not follow the usual laws of supply and demand. Usually, the higher the price of a particular good the less people will want it. For luxury goods, such as very expensive wines, watches or cars, however, the item becomes more desirable as it grows more expensive and less desirable should it drop in price.*” hedge funds are the highest-priced investment funds in the market. Does growing demand for hedge funds as an asset class and the preference of investors for higher-fee funds suggest that hedge funds are veblen goods or are the high fees justified by superior investment performance?

To answer this question, I analyze the development and growth of the hedge fund industry and try to understand the reasons investors choose this kind of asset over other assets. I then use data from one of the leading databases of hedge fund performance, EurekaHedge, to test a number of hypotheses relating to the decision by institutional investors to invest in hedge funds. I then conclude that while movements in other financial markets have been the main driver of hedge fund asset growth, investors have still tended to favor higher-priced hedge funds over cheaper alternatives, suggesting that to some extent, hedge funds are ‘veblen’ goods.

A hedge fund is a largely unregulated investment pool that is distinguished by a number of characteristics. Firstly, as the name suggests, the fund is supposed to be ‘hedged’ against general market risk. This distinguishes them from more general ‘long only’ mutual funds or passive funds (also known as Trackers or ETFs) the value of which generally track the performance of the market in which they are invested.

The manager of a regular mutual fund benchmarks his or her performance against a relevant market index. For equity funds, this is likely to be a broad market index like the S&P 500, Dow Jones Industrial, Russell 2000 or FTSE100. The manager’s job is to use his or her investment skill to outperform the market. If the market is up 10% over a certain period and the fund is up 12%, the manager will be deemed to have performed well. Similarly, if the market is down 10% and the fund is down 8%, the investors in the fund will have lost money, but the manager will be deemed to have performed well. The assumption is that by putting money in this fund, the investor is happy to assume the market risk and is

paying the manager to improve upon this market performance. Typical mutual fund fees will be 1-1.5% per year (100-150 basis points) and this fee is deducted annually from the fund.

A passive fund has no active fund manager making investment decisions. As money is invested in the fund, it is invested in every stock in an index based on the weighting in that index. Accordingly, the performance of the fund will mirror exactly the performance of the underlying index (less the management fees). As there is no fund manager making investment decisions, the fees on these passive products are much lower, usually in the range of 2-10 basis per year. As with a mutual fund investor, the investor is deemed to be happy to assume the market risk and will want to achieve this exposure as cheaply as possible rather than pay higher fees to a mutual fund manager to try (but normally fail) to outperform the market.

The two significant differences that distinguish hedge funds from these other types of funds is that rather than trying to match or outperform a market, hedge funds attempt to make money in absolute terms rather than relative to an index. This characteristic is the reason for the existence of another name for hedge funds, absolute return funds. The other difference is the fee level. The annual management fee charged by the manager of a hedge fund is usually 1.5-2% (150-200 basis points) as well as 20% of the profits of the fund. In most other fund types, there is only an annual management fee with all profits being for the benefit of investors. The justification for these high fees is that the hedge fund managers are more skilled than regular mutual fund managers and can deploy a much wider range of trading strategies. These skills, plus the range of strategies allow the hedge fund managers to always make money in absolute terms and outperform traditional funds.

2. Hedge Funds

2.1 Hedge Fund Performance Terminology

In this report, a number of concepts will be used to describe hedge fund performance. 'Returns' means the percentage change in asset value over a period after fees. This is how much the fund has made as a return on investors' money. The returns from a fund are broken into 'Alpha' and 'Beta'. 'Beta' is the part of the fund's performance that is attributable to the performance of the underlying market. 'Alpha' is the portion of the return that is above (or below) the return from the underlying market. Most hedge funds justify their fees on the basis that they can generate 'Alpha'.

2.1 Growth and Development of Hedge Funds as an Asset Class.

The hedge fund industry has grown from about \$118bn Amount under Management (AuM) in 1997 to over \$3tr AuM in 2017. The size of the industry was \$2.2tr in 2007, but after the 2008 financial crisis, it declined to \$1.5tr in assets. The size of the industry has since grown consistently to \$3tr AuM today. Investors typically look to diversify their investments across equities, bonds, cash, commodities and now allocate a portion of their assets to a category called Alternatives. hedge funds are the most common investment used for this category.

2.3 Why Do Investors Invest in Hedge Funds?

In the early stages of the hedge fund industry, most hedge fund clients were high net worth investors (HNWI), family offices and endowments looking for absolute returns and outperformance over traditional investment products. The fact that hedge funds were unregulated also meant that they were not open to retail investors. The investor base for

hedge funds has changed over the years and according to a report by Prequin Inc., 65% of hedge fund investors are now institutional investors (pension funds, insurance companies, sovereign wealth fund, *etc.*).

These institutional investors have a very sophisticated process for managing their assets. They seek to run a portfolio that diversifies their investment risk and produces attractive risk-adjusted returns. For each type of asset, they forecast expected return, volatility and correlation. Once they have these estimates, an asset allocation optimisation model is run to determine what gives them the best return for a given amount of risk. A diversified portfolio of hedge funds adds value to a portfolio by having a low correlation to major benchmarks with less volatility. Much of the discussion about the attractiveness of hedge funds focuses on their (lack of) performance versus major stock and bond indexes. The approach by institutional investors shows that whether hedge funds can outperform stock and bond markets is not a driver of their decision to invest. Even if hedge funds underperform major markets, they may still add value to a diversified portfolio if they move in a way that is uncorrelated with these markets. According to the Prequin report, only 7% of hedge fund investors are now seeking outperformance over other types of investments.

Ultra-low interest rates may also be a driver of hedge fund investment flows. Since the financial crisis, interest rates have been lowered across the world by central banks looking to help economies recover. These low interest rates have meant that interest rates on bonds are also very low. For pension funds that have to meet fixed pension payments for pensioners, this is a major problem. Pension funds have traditionally used bonds (particularly government bonds) to meet these commitments, but the current low interest rates on bonds are not enough to enable them to meet their pension liabilities. As long as hedge funds perform better than bond yields, they help pension funds to meet their obligations. This explains why most institutional money that has been invested in hedge funds has come from the parts of institutional portfolios that were previously invested in bond portfolios (Prequin, 2014).

So, if hedge funds charge much higher fees than any other type of fund that invests in public securities, do they offer investors benefits that they cannot get elsewhere?

3. Literature Review Themes

3.1 History of Hedge Fund Industry Growth and Development

Cochrane (2012) describes hedge funds and their strategies. “Hedge fund” is a legal term for an unregulated investment vehicle that is not open to the general public. Another difference between hedge funds and traditional funds is the very high fees charged. Index Trackers charge 0.05% of the fund value every year and mutual funds can charge 1-1.5% of the fund value every year. Hedge funds charge 2.0% per year and an average of 20% of any profits made by the fund.

Hedge funds tend to focus on a number of different strategies:

1. Long/Short: Minimising market risk by holding both ‘long’ and ‘short’ positions in securities.
2. Global Macro Taking advantage of moves in the global economy by trading currencies, interest rates and equity indexes.
3. Risk Arbitrage or Takeover Arbitrage: Trading in the stocks of companies involved in announced takeover deals.
4. Convertible Bond Arbitrage: Taking advantage of the cheap equity options embedded in convertible bonds. (Cochrane, 2012)

While hedge funds now run a variety of strategies like equity long/short, risk arbitrage, convertible arbitrage *etc.*, in the early 1990s, most hedge funds followed global macro strategies (Ibbotson et al, 2009). In the DotCom crash of 2000, hedge funds performed as predicted and lost less money than traditional funds and this led to a huge inflow of investor money. In the book *The Hedge Fund Mirage*, Andrew Lack (2011) calculates that losses in the 2008 financial crisis wiped out all profits hedge funds had ever made. Because of the 2008 crash, equities were generally flat from 2001-2010 but bonds performed strongly. Any investor in a mixture of bonds and equities would have outperformed hedge funds over this period. (Lack, 2011)

3.2 Problems with Indexes of Hedge Fund Performance

There are many indexes that attempt to measure hedge fund performance. Some of the indexes are inaccurate because they do not weight the returns by the size of the funds. Small hedge funds tend to do better than large ones, so an unweighted index will be unrepresentative of the overall performance of the industry. hedge funds run more balanced portfolios than equities alone, so a better benchmark against which to judge them would be an index made up of 60% equities and 40% bonds. (Steinbrugge, 2014)

Another problem with indexes of hedge fund performance is that poorly performing funds have probably been closed down. This means that the funds that make up the index are just the successful ones. If one invested in a range of hedge funds when they were established, one may have been invested in the ones that lost money and are now closed down, so the index will overstate how one would have done. This is called ‘survivor bias’. In general, we can say that indexes of hedge fund performance will overstate how well hedge funds have done as a group. In his paper, Cochrane tries to calculate this difference by re-inserting the data for closed funds. (Cochrane, 2012).

Another problem with the indexes is backfill bias. This is when funds wait until they are making money before they report their performance. Only if they are consistently making money will they report all of their historical data to the index providers. If they never make consistent money, they will never appear in the index data. This will also distort the performance data in a positive direction (Ibbotson *et al*, 2011).

Figure 1 shows that larger funds perform worse than smaller funds. The graph shows fund size along the x axis and annual returns on the y axis.

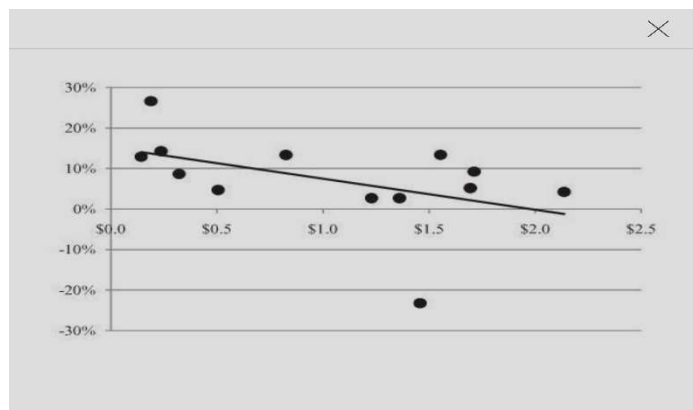


Figure 1. Hedge Fund Size vs Returns (Lack, 2012)

Looking at the differences between different databases of hedge fund performance, Joenvaara et al, (2016) found that most of the differences across the databases stem from different treatment of dead funds and incompleteness of Asset under Management data. This report compares the five most popular hedge fund databases. Barclayhedge, EurekaHedge, HFR (Hedge Fund Research), Lipper TASS and Morningstar and specifically looks at whether funds generate Alpha (market outperformance) after fees. Using the Fung and Hseih methodology (2004), they find an average Alpha of 4%. The indexes that show the highest Alpha (EurekaHedge and Morningstar) are the ones that have the smallest coverage of defunct funds and the differences between performance based on individual indexes get smaller in later periods. This suggests that methodologies across different Index providers are becoming more similar.

Results based on asset-weighted performance are explained by the fact that some indexes have many gaps in AuM for the funds measured. Approximately 70% of AuM data is missing for the early years of the Morningstar data. In the period towards the end of the study (2004-2012), the magnitude of hedge fund returns declines as the industry becomes larger. This could be caused by more money chasing fewer opportunities. The funds covered by the different databases vary widely, with 74% of funds only reported on one database.

Most investors do not buy a fund when it is established and hold it for a long period, so the report looks at how investors would fare if they were to sell low-performing funds and buy those that have recently performed better. In some indexes, performance is persistent (high-performing funds tend to sustain this performance in to the next period), but other indexes show no performance persistence, but fund size and age are negative indicators of future performance. The report is unclear on the correlation between fees charged and performance (Joenvaara *et al*, 2016).

3.3 Benchmarking Hedge Fund Performance

Despite coverage showing that hedge funds have underperformed the broad equity market indexes for a number of years, inflows have taken hedge fund assets to all-time highs. Many of these investors are sophisticated institutions, so why are they making decisions that seem irrational? The missing link is that these investors are not using broad equity market indexes like the S&P 500 as a benchmark because hedge funds have such a wide range of strategies. It may not be fair to compare them to something that is unrepresentative of their strategies. Comparing hedge funds to the broad equity market may also be wrong because hedge fund performance is less volatile than a pure equity fund, so on a 'risk adjusted' basis, they look better. (Steinbrugge, 2014)

We can break the performance of any portfolio down into two components. 'Beta' is performance that comes from the movement of the general market. 'Alpha' is the amount that a portfolio performs *above* (or *below*) the performance of the market. The big question for measuring the performance of hedge funds is: what market benchmark does one use as a proxy for market performance to work out the Beta of the hedge fund's performance and consequently, the Alpha? Merely using the performance of the broad stock market may not be appropriate because of the range of strategies hedge funds are pursuing. Since the correlation of hedge fund performance to the equity market is historically quite low, much of hedge fund performance is calculated as Alpha. If you use a range of benchmarks that better represent what hedge funds actually do, much more of their performance is actually Beta. Basically, if one chooses the right benchmarks, it looks more like they follow general market performance rather than outperforming the market. When markets are normal or rising, Betas can look quite low because markets move in different directions. When markets crash, every asset tends to go down in value at the same time. This makes Betas

much higher, so investors do not really get diversification when they *really* need it. (Cochrane, 2012)

Ibbotson et al (2010) estimate that the return of the hedge funds before they take out their fees is 11.42%. Of this, 3.78% is accounted for by fees to the hedge fund manager, 4.62% is Beta and 3.01% is Alpha. The mutual fund industry in aggregate just gives the market return with zero Alpha or outperformance. From this, the authors conclude that hedge funds do perform well and deliver value to investors.

Part of the problem is deciding which market return to use to calculate Beta. Any return above the performance of the chosen benchmark is called outperformance. This report merely uses normal equity, bond and cash returns for Betas. As we have seen, hedge funds run a variety of different strategies, which means that these benchmarks might not be representative of their strategies. Ibbotson, Chen and Zhu say this is acceptable because investors have no other way of accessing these non-traditional investment strategies other than through hedge funds, so it is right to ascribe this bit of their performance to their skill. If an investor is simply trying to capture Beta, he or she should invest in traditional equity or bond funds. Unlike other articles, Ibbotson et al say larger firms tend to outperform smaller ones. They assume that larger hedge funds can borrow more money than smaller funds, so this gives them more leverage. An article by Asness (2001) states that much hedge fund outperformance comes from 'non-traditional beta'. This is the same as the argument above in that, if one just compares hedge fund performance to traditional benchmarks like stock and bond indexes, they seem to outperform these benchmarks, but if one tries to construct benchmarks that actually reflect how they are investing, the outperformance is a lot less. What hedge fund investors are getting is not pure outperformance from smart managers, but access to non-traditional investment styles. It is not that hedge fund managers outperform these styles, it is just that there is no other way for investors to access these investment styles. This gives hedge funds something that really helps them attract money...non-correlated returns. This article gives credit for all performance over and above traditional benchmarks to the hedge fund managers even if it is just reflecting non-traditional benchmarks. This approach explains why this article shows that hedge funds deliver consistent outperformance of the market. This outperformance is not because hedge fund managers can "beat the market"; they are running investment strategies different from traditional asset managers (Ibbotson *et al*, 2010).

Dichev and Yu (2009) look at not only the published returns of hedge funds but investors' actual returns based on when they put their money into and remove it from the fund. It compares what they would receive if they invested the money in the fund and left it versus what they actually received, given the timing of their investments. The report finds that investors' actual returns are 3-7% lower than if they had just put the money in the fund and left it there. It appears that investors were chasing returns by investing in a fund after it had already risen in value, then taking the money out because the fund went down, missing any rebound in performance.

They also find that the real Alpha for hedge fund investors is close to zero. Some research (Stulz, 2007) shows annual outperformance of 3-5% per year. Other studies (Amin and Kat 2003) show no outperformance at all. This research calculates that hedge funds have returned 12.6% per year but the actual return earned by investors was only 6%. This is below both the return of the S&P 500, which was 10.9%, and only just above the risk-free rate of 5.6%. The report suggests that research by Ibbotson et al which shows consistent alpha, is contrary to that of most studies. It also appears that hedge fund returns have come down as the number and size of hedge funds have grown. This could be because more

money is attempting to capture the same number of trading opportunities, making it harder to make money (Dichev and Yu, 2009).

Lack (2012) discusses the problems of measuring hedge funds (survivor bias and backfilling) and agrees that small firms outperform larger ones. By necessity, investors are more exposed to larger funds so in addition to the problems with data on hedge fund performance, investors are more exposed to larger, worse-performing funds. The HFRX Index is a hedge fund index weighted by fund size. From 1998-2010, it shows an annual return of 7.3%. Over the same period, S&P 500 returned 5.9% and T-Bills returned 3%. Corporate bonds went up by 7.2%. As previously shown, this doesn't represent the returns investors actually made due to the timing of their investments (Lack, 2012). Boyson (2008) also finds that smaller funds outperform larger ones.

3.4 Who Invests in Hedge Funds and What Are Their Objectives?

Pension funds represent a large section of the investment universe and have been major investors in hedge funds over recent years. To arrive at an asset allocation model, they forecast expected return, volatility and correlation for each asset type. Once they have these estimates, they run an asset allocation optimisation model to determine which asset mix gives them the best return for a given amount of risk. A diversified portfolio of hedge funds adds value through its low correlation to major benchmarks with less volatility than pure equity assets. Since the financial crisis, interest rates have been lowered to help economies recover. This has meant that interest rates on bonds are also very low. For pension funds that have to meet fixed pension payments for pensioners, this is a major problem as they have historically relied on bonds to help meet their payment obligations. If hedge fund performance is higher than bond yields (which it has been), investing in hedge funds will help pension funds to meet their obligations. This explains why most of the Institutional money that has been invested in hedge funds has come from portfolios that were previously invested in bonds. (Steinbrugge, 2014)

3.5 Incentives, Regulation and Behaviour of Hedge Funds

If one buys an option on an asset, one gets the benefit of any positive price move but not the losses as your only loss is the cost of the option. A fee structure in which the hedge fund manager gets a basic fee (2% a year) *and* 20% of any profits incentivises the manager to take a lot of risk. The manager's payoff looks like an option payoff. They get a big reward if a portfolio makes money, but all the losses go to the client (unless the manager's own money is also invested in the fund). The only downside for the manager is that investors may withdraw their money if the fund suffers large losses. While hedge funds charge a 2% annual management fee and 20% of profits, 'fund of funds' take investors' money and invest it in a range of hedge funds. These 'funds of funds' charge 1% a year and 10% of profits on top of the fees charged by the underlying hedge funds. This further dilutes returns to the final investors. The main problem with monitoring hedge fund performance is that clients have no real transparency about the strategies, markets, and investment styles used by the managers, so they find it very difficult to assess them accurately (Cochrane, 2012).

Figure 2 shows how hedge fund performance is broken down between management fees and investor returns. The table shows that from 1998-2010, investors made a total of \$70bn in profits. Hedge fund fees in the same period were \$379bn. "Fund of fund" managers took out another \$61bn in fees, so investors made only \$9bn after all fees while the hedge fund managers plus the fund of fund managers took out \$440bn in fees. Managers made almost 50 times what investors made.

Year	Average HF AUM* (BNs)	Real Investor Profits (BNs)	Estimated HF Fees (BNs)**	Estimated FOF Fees (BNs)	Total Fees	Net Real Investor Profits (BNs)	Industry Share of Total Profits
1998	\$ 131	\$ 10	\$ 7	\$ 1	\$ 7	\$ 10	44%
1999	\$ 166	\$ 36	\$ 14	\$ 1	\$ 15	\$ 35	30%
2000	\$ 213	\$ 17	\$ 12	\$ 1	\$ 13	\$ 16	44%
2001	\$ 279	\$ 13	\$ 12	\$ 1	\$ 13	\$ 12	52%
2002	\$ 414	\$ 12	\$ 13	\$ 2	\$ 15	\$ 11	58%
2003	\$ 666	\$ 82	\$ 36	\$ 3	\$ 38	\$ 79	33%
2004	\$1,027	\$ 14	\$ 27	\$ 5	\$ 32	\$ 9	78%
2005	\$1,295	-\$ 6	\$ 35	\$ 7	\$ 42	-\$ 13	143%
2006	\$1,537	\$ 67	\$ 66	\$ 9	\$ 75	\$ 58	56%
2007	\$1,925	-\$ 11	\$ 59	\$11	\$ 70	-\$ 21	144%
2008	\$1,797	-\$448	\$ 36	\$10	\$ 46	-\$458	NM
2009	\$1,506	\$200	\$ 30	\$ 7	\$ 37	\$193	16%
2010	\$1,624	\$ 83	\$ 32	\$ 6	\$ 38	\$ 77	33%
Total		\$ 70	\$379	\$61	\$440	\$ 9	98%

Figure 2. Hedge Fund Fee Breakdown (Lack 2012)

Lack (2012) is highly critical of this breakdown and poses a question for investors, asking “is it right that, even if these investments give you uncorrelated returns and diversification, managers make so much more from investing your money than you do?” In summary, Lack claims that hedge funds have not served investors well. He believes managers have taken too much of the profits and too many investors invest based on past performance (Lack, 2015).

Cumming *et al* (2016) investigate whether increased regulation through measures like the Dodd-Frank Act (DFA), brought in after the financial crisis to improve stability in the financial system, has affected hedge fund performance. The DFA sought to improve transparency and regulation of the financial sector. The DFA only affects US-based hedge funds and research finds that, relative to non-US funds, US-based hedge fund performance has deteriorated (lower Alpha and return Standard Deviation) *i.e.*, the range of returns is narrower. The DFA requires US hedge funds to register with the Security Exchange Commission (SEC) and comply with minimum reporting requirements. The study finds that pre-DFA, US hedge funds produced higher Alphas than non-US hedge funds, but post-DFA, this advantage is gone. This has led to outflow from US-domiciled funds. This may be a sign that generally, increased regulation has dampened hedge fund returns (Cumming *et al*, 2016).

4. Methodology and Hypotheses

The objective of this paper is to look at the reasons behind the decision to invest in hedge funds. Are there solid financial reasons behind these decisions or are other forces at play? Hedge funds charge high fees because they promise superior performance. In the early days of the hedge fund industry, this superior performance was defined (by the funds themselves) as absolute outperformance over traditional long-only funds. For a number of years, and specifically since the financial crisis, hedge fund performance has lagged the performance of major asset markets. Unable to deny this reality, funds have focused on the fact that their performance is somewhat uncorrelated with major stock and bond markets. To

an investor looking to diversify the risk of a portfolio, holding an asset with uncorrelated returns reduces the risk of the portfolio. As noted above, low bond yields may have boosted demand for investment products with absolute returns above major bond market yields. Despite this change in emphasis, while there has been some small compression of hedge fund fees, high fees (both the annual management charge and the performance fee) have persisted.

If institutional investors have increased their holdings of hedge funds to replace low-yielding bond portfolios, have they chosen the lowest cost investment products that meet their objectives or more expensive ones? If the original rationale for high hedge fund fees, namely outperformance, is no longer credible, why have investors not insisted on lower fees? funds with high fees may have a 'luxury image' because of past performance or the prestige of the managers. Are investors putting money into these funds because it is rational to believe that higher-fee hedge funds outperform those with lower fees or are the higher-fee hedge funds 'Veblen' goods, those for which demand is driven by the high price itself rather than actual performance?

4.1 Hypothesis 1

Moves in the US Treasury 10-year bond yield are negatively correlated with hedge fund inflows. The rationale for this is that as interest rates have fallen, pension fund investors have used hedge funds to replace low-yielding bonds in an attempt to meet their fixed liabilities. To compare hedge fund net inflows with a benchmark government bond yield, I have added a six-month lag between the government bond yield and hedge fund Net Inflows. Pension funds take time to analyse movements in market variables and readjust their portfolios and six months is a reasonable period for changes in the benchmark government bond yield to lead to investment rebalancing decisions.

Independent Variable: Yield on US 10-year note

Dependent Variable: Yedge fund inflows

4.2 Hypothesis 2

There is no correlation between fee level and hedge fund performance. The rationale for this hypothesis is not only that hedge fund fees are unjustified by their performance, but also that within the hedge fund universe, high fees do not indicate higher performance.

Independent Variable: Hedge fund performance fee

Dependent Variable: Annual returns since inception

4.3 Hypothesis 3

Hedge fund inflows are positively correlated with management fee level. The rationale here is that irrespective of performance, investors are attracted to higher-fee hedge funds because they have a 'luxury' image that lower-price funds lack.

Independent Variable: Hedge fund performance fee

Dependent Variable: Hedge fund assets under management

5. Data

5.1 Dataset for Hedge Fund Performance and Assets under Management.

I will use one of the leading hedge fund databases from Eureka hedge. From this database, I will use the data for North American funds from June 2007 to July 2017 (7,263 funds). The download contains monthly performance and investment flow data. For Hypothesis 1, I have used a download of monthly yields on the US 10-year note from Bloomberg.

5.2 Choice of Dependent Variables (Hedge Fund Net Inflows and Average Annual Performance and Assets under Management)

I attempt to investigate the reasons behind hedge fund investment decisions. Money flowing in to a fund is the result of that positive decision, and money flowing out of a fund is an investor making a negative decision. For Hypothesis 1, I used net asset inflow data from the EurekaHedge North American database from June 2007. For Hypothesis 2, I used average annual performance between 2013 and 2017 YTD (July). For Hypothesis 3, I used assets under management.

5.3 Choice of Independent Variables (Hedge Fund Annual Performance Fees and US Interest Rates)

As a proxy for the yield that affects institutional investors, I have chosen the yield on the US Treasury 10-year note, the most liquid benchmark in fixed income markets. To discover whether fees are correlated with performance, the easiest performance variable to use is annual performance fee. While funds do have different management fees, the biggest differentiator between lower and high-fee hedge funds is the annual performance fee.

5.4 Data for all Variables Used

See spreadsheets below.

5.5 Statistical Methods Used

For Hypothesis 1, I looked for a correlation between the yield on the benchmark 10-year US Treasury bond and six month-lagged hedge fund net inflows. The reason for the six-month lag is that large institutions do not immediately adjust their portfolios in reaction to changes in financial markets. Six months is a reasonable period for fund managers to make such an adjustment.

For Hypothesis 2, I used a report from hedge fund consultancy, Preqin, published in 2013 and from my own data, I calculated average annual returns from 2013-17 and grouped the funds by annual performance fee.

For Hypothesis 3, I calculated total assets under management for funds in different bands of annual performance fees. While the AuM totals include performance, in aggregate, they are a good indication of where investors have decided to put their money.

6. Results

6.1 Hypothesis 1

Moves in the US Treasury 10-year bond yield are negatively correlated with hedge fund inflows.

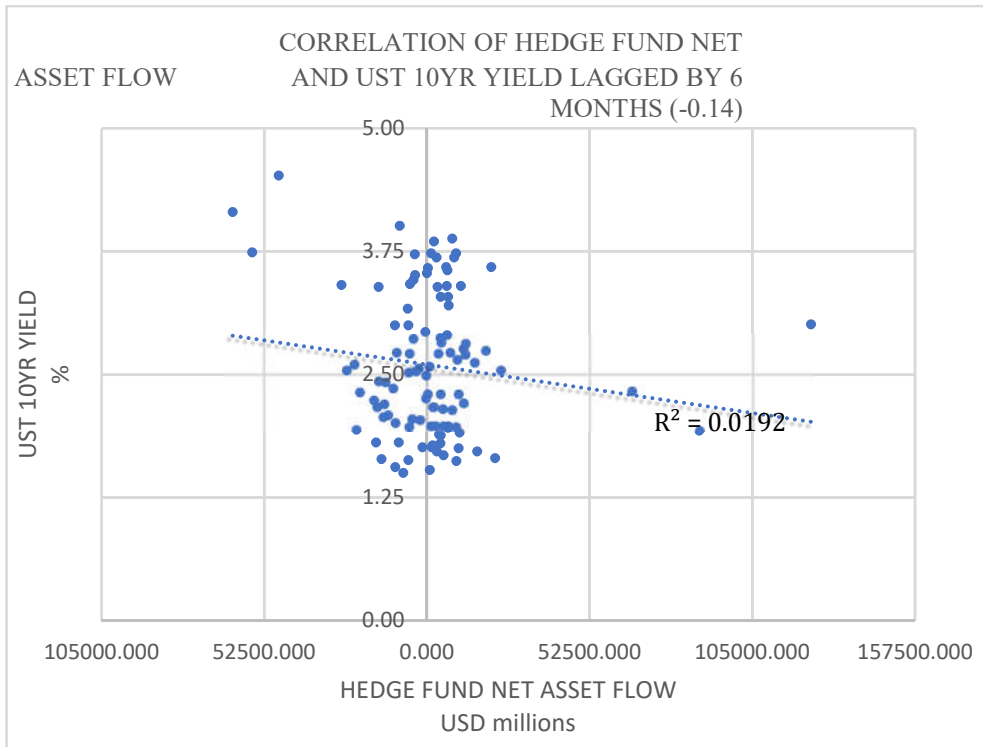


Figure 3. Hedge Fund Net Asset Flow vs Yield on UST 10-year Note

Using data from June 2007 to July 2017, Figure 3 shows that there was a correlation of -0.14 between the yield on the benchmark 10-year US Treasury bond and net asset flows to US-located hedge funds.

Over the period, the yield dropped significantly, from just under 5% to just over 1.25%, while net hedge fund inflows were strongly positive. The directional relationship looks strong, but the weak negative correlation suggests that the timing of the investment process may not be predictable or stable. Given that, according to the literature, most of the funds that have flowed in to hedge funds have come from institutional portfolios that were previously invested in fixed income assets, low yields are a plausible reason for this shift. Low bond yields are the main reason that many defined benefit pension funds are in deficit and managers are therefore looking for assets that may help them close these deficits. If hedge fund returns are higher than government bond yields, the fact that hedge fund returns are relatively uncorrelated with returns of other major asset classes is an additional benefit.

6.2 Hypothesis 2

There is no correlation between fee level and hedge fund performance. A report in 2013 by the hedge fund consultancy, Prequin, looked at the performance of hedge funds with three different levels of performance fees: less than 20%, 20% and more than 20%. For these three groups of funds, the report looked at annualised performance over three and five years. They claim that funds charging over 20% had the highest annualised returns over three and five years.

Preqin do not supply their data, but the chart in the report tells a different story than their conclusion. Figure 4 shows that in 2009 and 2012, hedge funds with performance fees below 20% performed the best and in 2013, the performance between funds charging performance fees of less than 20% is very close to the performance of funds charging performance fees of more than 20%. At the very least, the data is inconclusive and it there does not appear to be a correlation between fee levels and performance

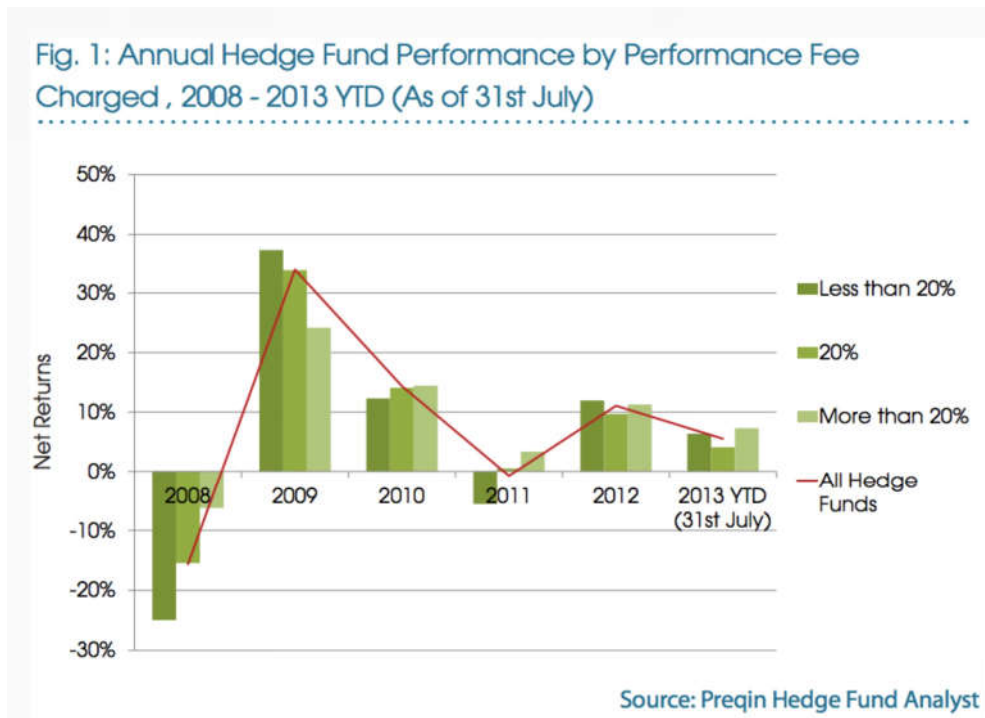


Figure 4.

Annual Hedge Fund Performance by Performance Fee Charged, 2008-2013 YTD (July 31st)

From my own data, Figure 5 shows the average annual performance of the funds in the dataset, grouped by annual performance fee, from 2013 to 2007 YTD, with the annual performance of the S&P 500 overlaid:

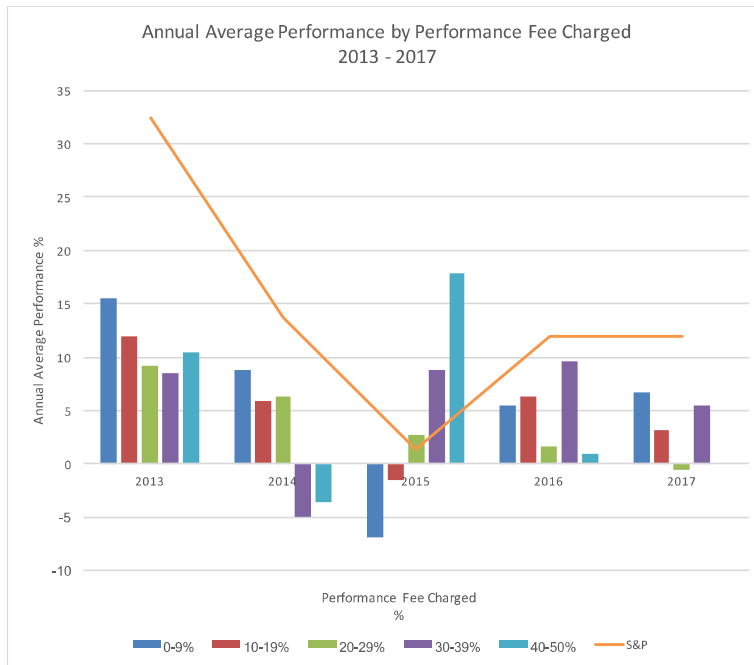


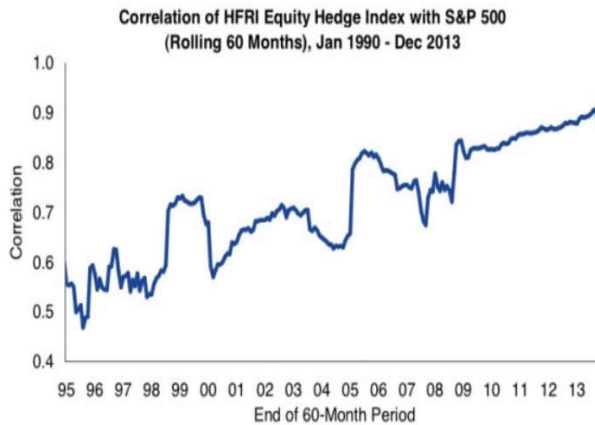
Figure 5. Annual Average Performance by Fee Charged, 2013-2017

The data show that there is no clear relationship between performance fees and after-fee returns of the funds in the dataset. Furthermore, given that the majority of strategies are equity-based, hedge fund performance consistently underperforms the main equity benchmark, the S&P 500. The fact that the higher-fee funds show similar performance after fees to lower-fee funds implies that the before-fee performance of the higher-fee funds must be superior. The managers may be more talented at being able to generate returns, but virtually the entirety of this benefit goes to the managers through their fees, rather than to the investors.

Many of the studies reviewed point out that while hedge fund returns have lagged behind those of major equity markets, hedge fund returns tend to be less volatile than conventional Equity funds so on a risk-adjusted basis, they will look better than the nominal returns. An even more important consideration for risk-adjusted performance is the maximum drawdown - the largest peak-to-trough decline. Drawdowns have been considerably worse for various equity indices relative to hedge funds during the period analysed.

Apart from risk-adjusted returns, the other reason given for investing in hedge funds is lack of correlation of hedge fund returns to major equity markets. However, a study by Morgan Stanley (Figure 6) found that correlation between one of the main hedge fund indexes and the S&P 500 has increased significantly and is now around 0.9. While hedge funds may have offered portfolio diversification in earlier years, it is hard to say that hedge fund returns offer diversification that with this high level of correlation.

Morgan Stanley

MORGAN STANLEY RESEARCH
January 13, 2014
U.S. Equity Strategy**Theme 1: Performance - Hedge Funds in Aggregate Are Essentially Long the S&P 500**

Source: Factset, Morgan Stanley Research.

Figure 6. Hedge Fund Correlation with S&P 500**6.3 Hypothesis 3**

Hedge Fund inflows are positively correlated with fee level. To test this hypothesis, Figure 7 shows current hedge fund assets under management in the EurekaHedge North American database, grouped by different performance fee levels:

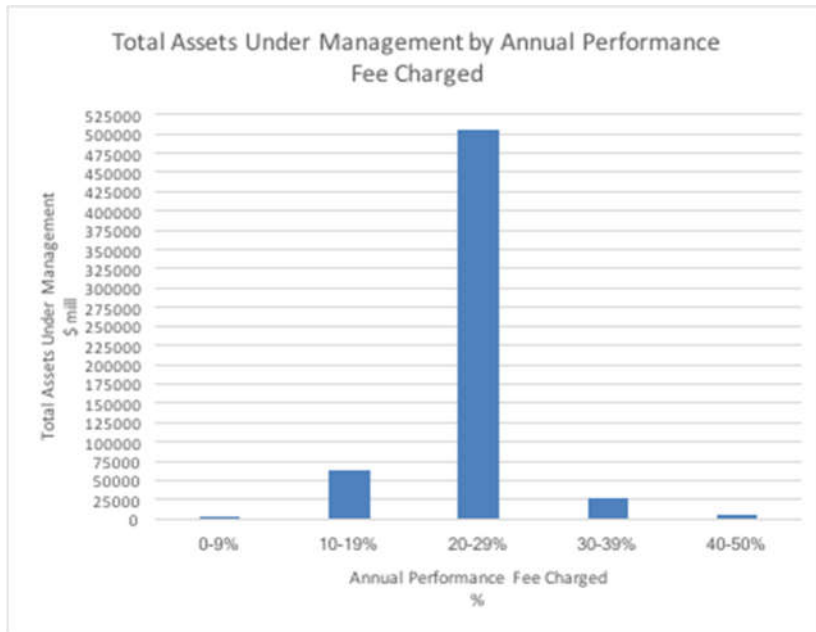


Figure 7. Total Assets under Management by Annual Performance Fee

This data shows that up to a level of 30%, demand for hedge funds seems to be positively correlated with annual Performance Fee. Only above this level does demand reduce for higher-priced funds. For many years, 20% Performance Fee was an industry standard for hedge funds but the fact that investors still have the majority of their assets sitting in funds with high performance fees suggests that this is their clear preference. They could easily have moved their money to lower-cost funds but have not done so.

Hedge funds have been the beneficiaries of low bond yields as Institutional investors look to replace parts of their Fixed Income portfolios with assets providing better returns. They have also benefitted from the fact that hedge funds are advertised as offering portfolio diversification and attractive risk-adjusted returns. If these are the macro factors driving asset flow in to hedge funds, what can we say about the type of ‘goods’ hedge funds represent?

If investors were looking for portfolio diversification and a yield pickup over fixed income assets, why would they not look for the cheapest alternative in the same way as retail investors have moved from active into passive funds as the most effective way for them to invest in equities? The data does not provide a clear financial justification for investors’ preference for high-fee hedge funds over low-fee hedge funds. There is disagreement as to whether hedge funds offer outperformance (Alpha) over traditional benchmarks. Studies that show consistent Alpha can be criticised on the grounds that using traditional benchmarks to calculate the Alpha generated by hedge funds using complex strategies is misleading. A large part of the calculated Alpha is actually accounted for by the fact that traditional benchmarks are unrepresentative. The outperformance is merely market performance of non-standard strategies. Other studies show that the actual performance captured by investors is below the quoted performance of the funds because of poor market timing as investors chase past performance and withdraw money after the fund experiences losses, potentially missing out on a rebound in performance.

These arguments about whether hedge funds provide Alpha, diversification or attractive risk-adjusted returns apply to all hedge funds. If investors are attracted despite these issues, they still seem to prefer high-fee funds. On top of this, recognising that there is a high degree of variance amongst returns, many investors either choose to invest in a range of hedge funds or a “fund of funds”. As “funds of funds” charge management and performance fees on top of the fees charged by the underlying funds, their returns are lower than the funds in their portfolio.

The managers who run the hedge funds charging higher fees are often financial superstars. Investors may be attracted to having their money managed by these famous investors despite the evidence that these managers tend to keep most of the rewards of their brilliance.

Another dynamic could be the gambling instinct. People bet money on sporting events in the hope that a win would make them an outside return. The same dynamic could be a factor here. Investors could wager some money on a financial superstar in the hope that the manager’s skill will pay off for them, even if past performance suggests this is unlikely. This is no difference from betting on a sports result that is not favoured. It might not happen but if it does, there could be substantial returns.

Rational investors seeking diversification could invest in lower-cost funds. My research shows their performance would be very similar to what they could have achieved using higher-fee funds and they would have the additional benefit of not seeing managers enrich themselves off of their investors’ money. Another alternative would be to invest in structured products. There are a number of alternative products available to investors that offer attractive, diversified returns. There are ETFs that replicate standard hedge fund strategies with much lower fees and ‘Alt Beta’ products have a similar profile at a much lower cost.

7. Conclusions

When analysing the relationship between benchmark government bond yields and hedge fund inflows, I found a weak negative correlation with a six-month time lag. I expected the correlation to be stronger but the link, while directionally obvious, is less clear than I believed it would be.

The data shows that hedge funds with higher fees do not outperform hedge funds with lower fees. The average annual returns for funds with higher fees are similar to funds with lower fees. The fact that higher-fee funds have similar performance to lower-fee funds shows that the managers of these funds are more talented, but the benefits of their talent are retained by the managers via their fees rather than passed on to investors. Despite this conclusion, investors have tended to favour high-fee hedge funds over low-fee hedge funds.

In the absence of a rational reason for current patterns of hedge fund investment, I am drawn back to my original question: are hedge funds Veblen goods?

I believe institutional investors are making an irrational investment decision as they are attracted by the past outperformance offered by hedge funds and invest in the hope that these returns may come back in the future, despite the evidence that this is unlikely. Many of the top hedge fund managers are financial celebrities and investors like the idea of these superstars managing their money, even if the fruits of the managers’ genius are retained by the managers. If what hedge funds now deliver is diversification, why do investors continue to pay ‘2+20’ or ‘1.5+15’ when there are cheaper alternatives?

The next step in my research is to generate some qualitative data by interviewing institutional investors and investment consultants about their decision-making process. This type of data will help to uncover whether there are rational financial reasons for investing in

high-fee hedge funds that are not found in my data. This approach might also uncover some of the more ‘gut feel’ reasons for these types of investments.

My view is that while hedge funds do provide some value–diversification and uncorrelated returns–the fact that they prefer high-fee funds rather than low-fee funds, and have failed to take up cheaper structured product alternatives, shows that hedge funds are Veblen goods. They have a ‘luxury’ or ‘prestige’ image that gives investors utility beyond the financial returns offered by the funds.

8. Bibliography

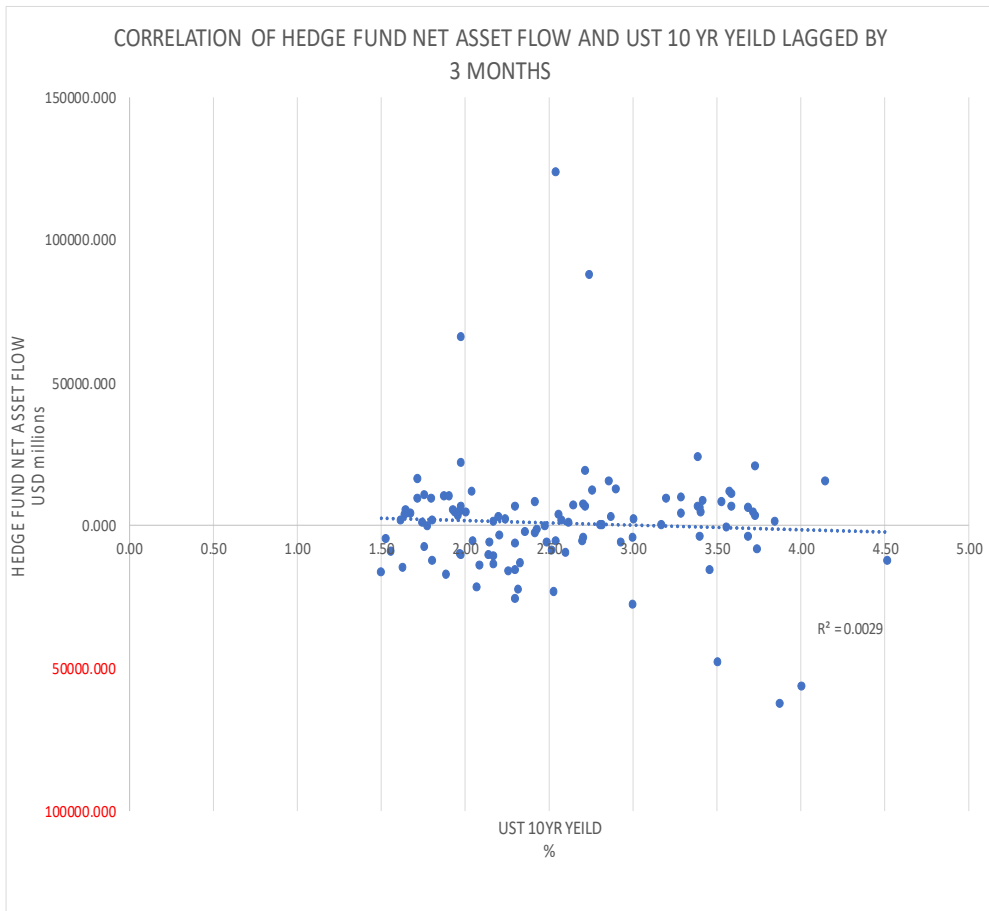
- Amin, G. and Kat H. “Hedge Fund Performance 1990-2000: Do the ‘Money Machines’ Really Add Value?”
https://www.jstor.org/stable/4126750?seq=1#page_scan_tab_contents 2003 Assess, C. “Do Hedge Funds Hedge?”
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1581559. May 2001.
- Boyson, N. “The Impact of Hedge Fund Family Membership on Performance and Market Share.”
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1286434. 2008. Cochrane, J., Chicago Booth School of Business–Advanced Hedge Fund Notes ~
https://faculty.chicagobooth.edu/john.cochrane/teaching/35150_advanced_investments/hedge_notes_and_questions.pdf. 2012
- Cumming, Dai and Johan. “Dodd-Frinking the Hedge Funds”,
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2527679. May 2016
- Dichev and Yu, “Higher Risk, Lower Returns: What Hedge Fund Investors Really Earn.”
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1354070. July 2009 Ibbotson,
- Chen and Zhu. “The ABCs of Hedge Funds–Alphas, Betas and Costs.”
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1581559 March 2010
- Joenvaara, Kosowski and Tolonen. “Hedge Fund Performancer: What Do We Know?”
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1989410. March 2016
- Lack, A. “The Hedge Fund Mirage–The Illusion of Big Money and Why It’s Too Good to be True”. November 2011.
- Steinbrugge, D. CAIA, “Why Are Hedge Fund Assets Reaching All-Time Highs?”,
<http://www.allaboutalpha.com/blog/2014/05/12/why-are-hedge-fund-assets-reaching-all-time-highs-while-they-underperform-the-sp-500/>. May 2014
- Stultz, R. “Hedge Funds: Past, Present and Future.” *Journal of Economic Perspectives*–
 Volume 21, Number 2, Spring 2007.

9. Appendices

9.1 Hedge Fund Net Asset Flow vs UST 10-Year Note (Source: EurekaHedge)

3 MONTH LAG					
Month	Net asset flow	UST 10-Yr. Yield			
Jul 17		2.30	Dec 14	25820.241	2.53
Jun 17	15518.656	2.48	Nov 14	23360.377	2.42
May 17	173.984	2.42	Oct 14	2581.801	2.54
Apr 17	8161.576	2.43	Sept 14	5607.339	2.60
Mar 17	1448.703	2.49	Aug 14	9695.470	2.56
Feb 17	6001.902	2.14	Jul 14	3738.183	2.71
Jan 17	10217.735	1.76	Jun 14	4372.317	2.72
Dec 16	7693.825	1.63	May 14	6490.137	2.71
Nov 16	14668.754	1.56	Apr 14	7528.502	2.86
Oct 16	9099.576	1.50	Mar 14	15379.699	2.90
Sep 16	16479.460	1.64	Feb 14	12495.835	2.72
Aug 16	3952.089	1.81	Jan 14	19008.949	2.62
Jul 16	1805.460	1.81	Dec 13	824.029	2.81
Jun 16	12494.167	1.89	Nov 13	275.790	2.74
May 16	17051.339	1.78	Oct 13	87849.243	2.58
Apr 16	302.220	2.09	Sept 13	1546.855	2.30
Mar 16	14086.420	2.24	Aug 13	6753.202	1.93
Feb 16	2299.151	2.26	Jul 13	5343.884	1.76
Jan 16	15991.471	2.07	Jun 13	10521.423	1.96
Dec 15	21647.648	2.17	May 13	3203.737	1.98
Nov 15	10883.922	2.17	Apr 13	21952.599	1.91
Oct 15	13780.896	2.32	Mar 13	10248.439	1.72
Sept 15	22698.909	2.36	Feb 13	16225.877	1.65
Aug 15	2137.246	2.20	Jan 13	5354.847	1.75
Jul 15	2821.998	1.94	Dec 12	872.701	1.72
Jun 15	4378.763	2.04	Nov 12	9513.114	1.68
May 15	11813.929	1.98	Oct 12	4319.800	1.53
Apr 15	66069.528	1.88	Sept 12	4762.764	1.62
Mar 15	10230.983	2.21	Aug 12	1931.968	1.80
Feb 15	3574.833	2.33	Jul 12	9366.422	2.05
Jan 15	13324.560	2.30	Jun 12	5662.356	2.17

May 12	1308.571	1.97	Dec 09	6617.623	3.40
Apr 12	10113.959	1.97	Nov 09	3914.953	3.59
Mar 12	5163.188	1.98	Oct 09	6772.539	3.56
Feb 12	6762.201	2.01	Sept 09	495.403	3.72
Jan 12	4460.410	2.15	Aug 09	4593.715	3.29
Dec 11	5996.483	1.98	Jul 09	4293.829	2.93
Nov 11	10358.033	2.30	Jun 09	5937.980	2.82
Oct 11	6218.532	3.00	May 09	21.342	2.87
Sept 11	4375.438	3.00	Apr 09	3095.714	2.52
Aug 11	27562.235	3.17	Mar 09	8768.529	3.53
Jul 11	285.670	3.46	Feb 09	8105.524	3.69
Jun 11	15638.311	3.41	Jan 09	3882.647	4.01
May 11	4455.224	3.58	Dec 08	56375.054	3.88
Apr 11	11783.158	3.39	Oct 08	62690.583	3.51
Mar 11	23836.567	3.29	Aug 08	47792.181	3.74
Feb 11	9830.857	2.76	Jun 08	8382.244	4.15
Jan 11	12356.884	2.54	Apr 08	15670.940	4.52
Dec 10	123895.242	2.65	Feb 08	12248.981	
Nov 10	6971.122	2.70	Dec 07	16328.662	
Oct 10	5493.653	3.01	Oct 07	8972.132	
Sept 10	2238.577	3.20		30159.382	
Aug 10	9399.928	3.42			
Jul 10	8740.036	3.85			
Jun 10	1287.613	3.73			
May 10	20733.003	3.69			
Apr 10	6387.988	3.73			
Mar 10	3352.005	3.59			
Feb 10	10954.916	3.40			
Jan 10	6134.523	3.39			

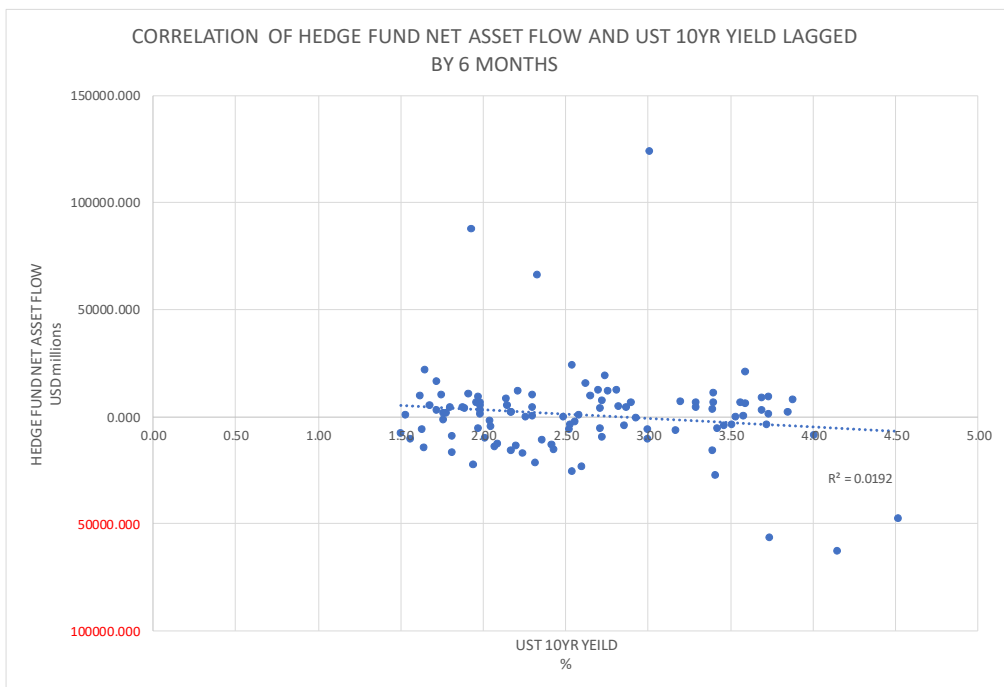


Correlation: -0.053741502

<u>6 MONTH LAG</u>			Nov 16	14668.754	1.81
Month	Net Asset Flow	UST 10-Yr. Yield	Oct 16	9099.576	1.81
Jul 17		2.43	Sep 16	16479.460	1.89
Jun 17	15518.656	2.49	Aug 16	3952.089	1.78
May 17	173.984	2.14	Jul 16	1805.460	2.09
Apr 17	8161.576	1.76	Jun 16	12494.167	2.24
Mar 17	1448.703	1.63	May 16	17051.339	2.26
Feb 17	6001.902	1.56	Apr 16	302.220	2.07
Jan 17	10217.735	1.50	Mar 16	14086.420	2.17
Dec 16	7693.825	1.64			

Feb 16	2299.151	2.17	Mar 13	10248.439	1.72
Jan 16	15991.471	2.32	Feb 13	16225.877	1.68
Dec 15	21647.648	2.36	Jan 13	5354.847	1.53
Nov 15	10883.922	2.20	Dec 12	872.701	1.62
Oct 15	13780.896	1.94	Nov 12	9513.114	1.80
Sep 15	22698.909	2.04	Oct 12	4319.800	2.05
Aug 15	2137.246	1.98	Sep 12	4762.764	2.17
Jul 15	2821.998	1.88	Aug 12	1931.968	1.97
Jun 15	4378.763	2.21	Jul 12	9366.422	1.97
May 15	11813.929	2.33	Jun 12	5662.356	1.98
Apr 15	66069.528	2.30	May 12	1308.571	2.01
Mar 15	10230.983	2.53	Apr 12	10113.959	2.15
Feb 15	3574.833	2.42	Mar 12	5163.188	1.98
Jan 15	13324.560	2.54	Feb 12	6762.201	2.30
Dec 14	25820.241	2.60	Jan 12	4460.410	3.00
Nov 14	23360.377	2.56	Dec 11	5996.483	3.00
Oct 14	2581.801	2.71	Nov 11	10358.033	3.17
Sep 14	5607.339	2.72	Oct 11	6218.532	3.46
Aug 14	9695.470	2.71	Sep 11	4375.438	3.41
Jul 14	3738.183	2.86	Aug 11	27562.235	3.58
Jun 14	4372.317	2.90	Jul 11	285.670	3.39
May 14	6490.137	2.72	Jun 11	15638.311	3.29
Apr 14	7528.502	2.62	May 11	4455.224	2.76
Mar 14	15379.699	2.81	Apr 11	11783.158	2.54
Feb 14	12495.835	2.74	Mar 11	23836.567	2.65
Jan 14	19008.949	2.58	Feb 11	9830.857	2.70
Dec 13	842.029	2.30	Jan 11	12356.884	3.01
Nov 13	275.790	1.93	Dec 10	123895.242	3.20
Oct 13	87849.243	1.76	Nov 10	6971.122	3.42
Sep 13	1546.855	1.96	Oct 10	5493.653	3.85
Aug 13	6753.202	1.98	Sep 10	2238.557	3.73
Jul 13	5343.884	1.91	Aug 10	9399.928	3.69
Jun 13	10521.423	1.72	Jul 10	8740.036	3.73
May 13	3203.737	1.65	Jun 10	1287.613	3.59
Apr 13	21952.599	1.75	May 10	20733.003	3.40

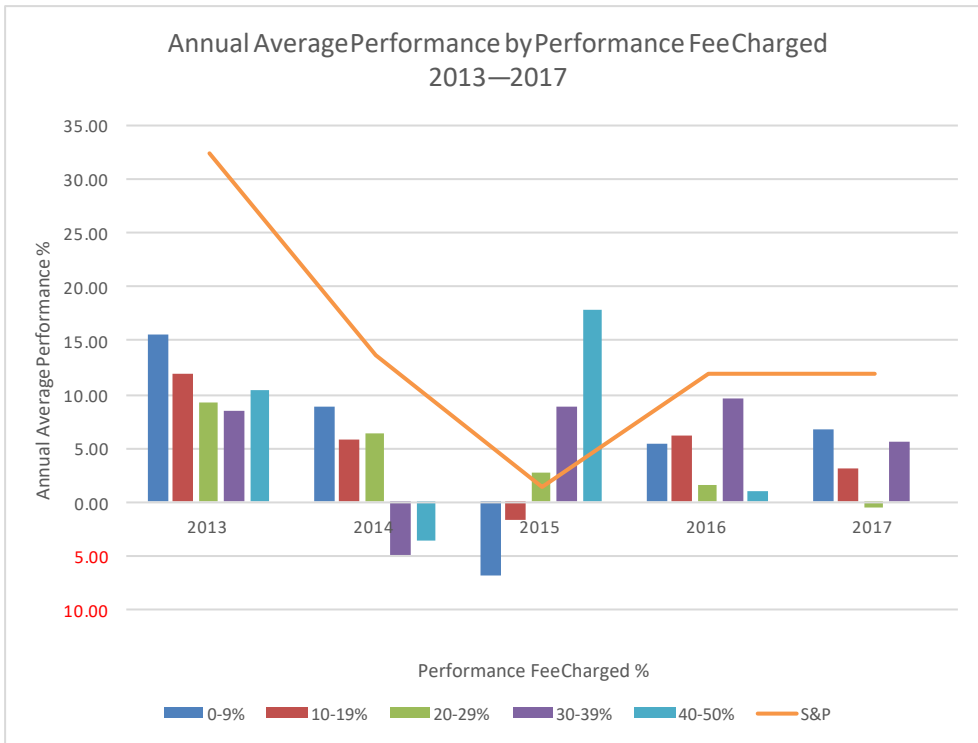
Apr 10	6387.988	3.39	Mar 09	8768.529	3.88
Mar 10	3352.005	3.40	Feb 09	8105.524	3.51
Feb 10	10954.916	3.59	Jan 09	3882.647	3.74
Jan 10	6134.523	3.56	Dec 08	56375.054	4.15
Dec 09	6617.623	3.72	Oct 08	62690.583	4.52
Nov 09	3914.953	3.29	Aug 08	47792.181	
Oct 09	6772.539	2.93	Jun 08	8382.244	
Sep 09	495.403	2.82	Apr 08	15670.940	
Aug 09	4593.715	2.87	Feb 08	12248.981	
Jul 09	4293.829	2.52	Dec 07	16328.662	
Jun 09	5937.980	3.53	Oct 07	8972.132	
May 09	21.342	3.69		30159.382	
Apr 09	3095.714	4.01			



Correlation: -0.138529198

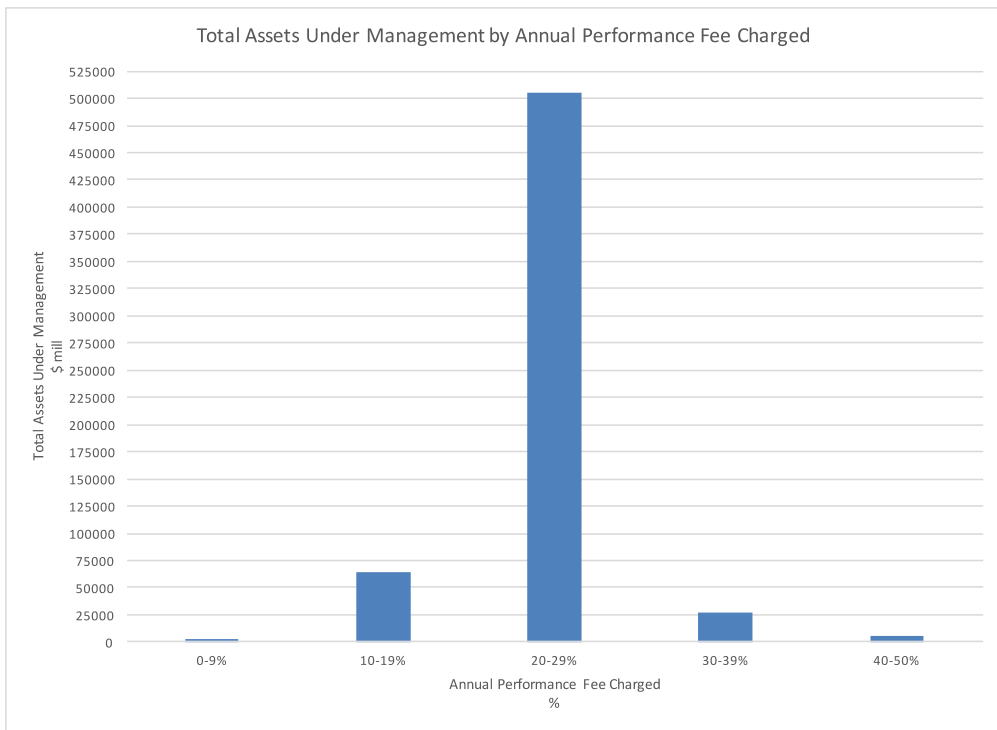
9.2 Hedge Fund Performance Fee vs Average Annual Hedge Fund Returns
 (Source: EurekaHedge)

	2013	2014	2015	2016	2017
0-9%	15.56	8.84	6.85	5.50	6.69
10-19%	11.89	5.84	1.57	6.27	3.13
20-29%	9.19	6.35	2.69	1.62	0.55
30-39%	8.57	4.94	8.82	9.66	5.54
40-50%	10.46	3.54	17.88	0.97	0.00
S&P	32.39	13.69	1.38	11.96	11.93



9.3 Hedge Fund Performance Fee vs Assets under Management (Source: EurekaHedge)

<u>PERFORMANCE FEE</u> %	<u>TOTAL AUM</u> \$mill
0-9	2614
10-19	64324
20-29	505280
30-39	26799
40-50	5625



Brain Plasticity: The Effects of BrainPort Sensory Substitution

Elexis Hernandez Sanchez

Author Background: Elexis Hernandez Sanchez grew up in the United States and currently attends Dr. Kirk Lewis Career and Technical High School in Houston, Texas. Her Pioneer seminar topic was in the field of neuroscience and titled "Understanding the Sense of Touch through Neuroscience."

Abstract

The BrainPort V100 is a modern sensory substitution device which translates visual stimuli into pulses applied to the tongue. The goal of this experiment will be to analyze the effectiveness of sensory substitution training on different age groups and conditions. The subjects tested will include young adults and children, some of which will be congenitally blind, acquired blind, and sighted. These will be placed into 12 groups with respect to their age and conditions. Half of the 12 groups of subjects tested will undergo 432 hours of training with the BrainPort, while the other half will not receive such training. Each subject will undergo a PET scan four times, every two months for the duration of the experiment. Three tactile tests will be performed on each subject to analyze the nature of the effects that sensory substitution via the use of the BrainPort has on brain function and tactile sensitivity.

Introduction

Sensory substitution devices are those which allow for the compensation of a lost sense by allowing another sense to convey to the brain the same information that would normally be garnered by the lost sense. Sensory substitution devices have long been under study, with some early inventions being the walking cane and braille (Ojala, 2016). Paul Bach-y-Rita was one of the first to seriously believe that neuroplasticity was still present in adult brains, and he developed one of the first sensory substitution systems. His work in the 1960s, such as introducing a chair that worked to allow the blind to view their surroundings, did much to advance the field. This device was named TTVS (Tongue Tactile Vision Substitution) and functioned by using a camera to record what was within the user's range of vision. See Figure 1. These visual stimuli were then translated into an assortment of vibrating tactile stimuli located on the back of the chair where the subject was sitting. In this way, the chair could convey information about shadows, items, and faces present in the room (Bach-y-Rita, Collins, Saunders, White, & Scadden, 1969). Bach-y-Rita's work would help open a new field of study on the effects of neuroplasticity and the extent to which it was present in the adult brain.

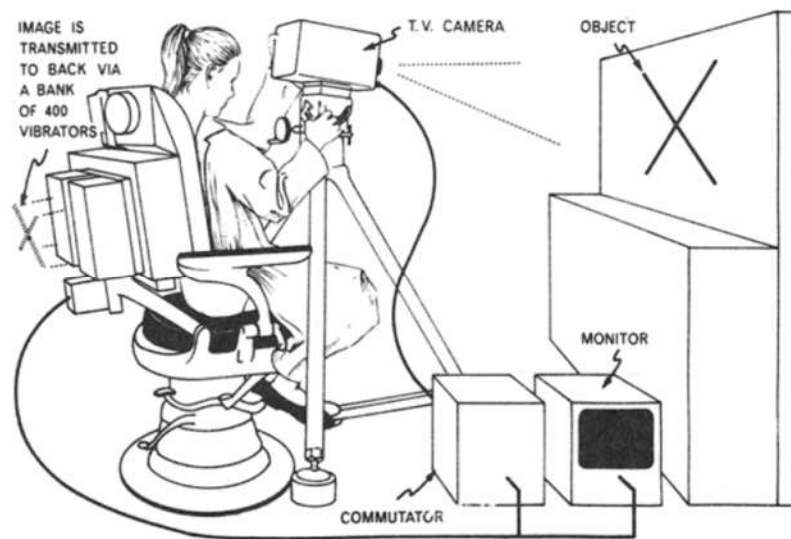


Figure 1. Illustration of Bach-y-Rita's TTVS. (Source: White, Saunders, Scadden, Bach-y-Rita, & Collins, 1970).

Other sensory substitution devices that have been made throughout the years have advanced the translation of visual information to tactile stimuli. In general, visual sensory substitution devices act by allowing visual information that would normally be detected by the eyes to be processed by another sensory modality. One way in which this may work is to allow the blind person to not directly sense the object, as a sighted person would, but rather to associate the object directly with the touch that signifies its presence—meaning that they perceive the object as being before them, but don't create a visual image of it in their head. Further, as the perception of seeing the object is created through the sense of touch, it may be possible that the visual cortex, as well as the somatosensory cortex, is activated when using such sensory substitution devices. The two areas form a system for the registration of the visual information being conveyed through tactile stimulation (Ojala, 2016). Thus, brain plasticity plays a vital role in the processes involved in sensory substitution devices, as it allows for the brain to be rewired in a way that allows for these two areas to become intertwined and more involved with each other than would have before been possible. It is this prolonged cooperation between these two areas that makes continuous perception of vision possible for blind individuals using a sensory substitution device.

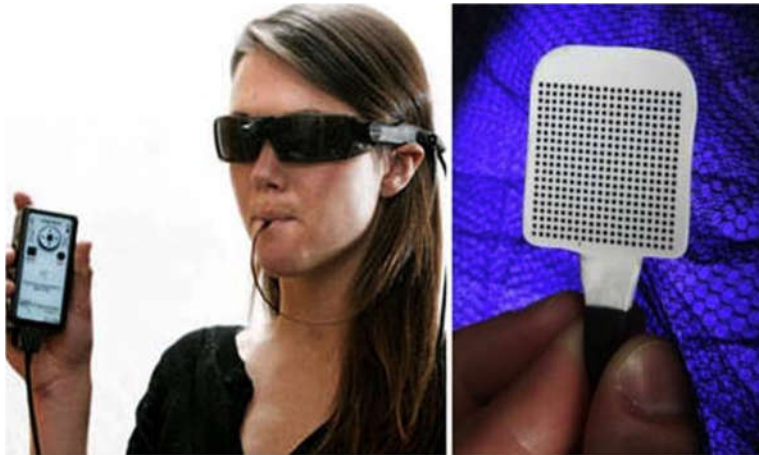


Figure 2. *The BrainPort allows the blind to perceive objects via tactile stimulation. (Source: Salton, 2009)*

The BrainPort V100 was approved by the Food and Drug Administration in 2015, and substitutes vision with tactile stimuli. See Figure 2. The sensory substitution device is composed of a tiny, 1.5 cm camera that is affixed onto the center of a pair of sunglasses and is used to record the amount of light in a given area. The device also includes a base unit, which is held by the user and is about the size of a phone. This base unit holds controls for shock intensity, zoom, and light settings. It also houses a central processing unit (CPU) that works to translate the digital image taken by the camera into an assortment of pulses of electricity. See Figure 3. These pulses of electricity have been reported to feel like bubbles on the tongue's surface. The CPU then transmits its data to an electrode array, which measures approximately 9 cm², and is placed on the user's tongue. Each of the 400 electrodes on the electrode array function to convey information from a set of pixels. A strong pulse is delivered for the presence of white pixels, while no signal is delivered for the presence of black pixels. With the use of this device, blind users can distinguish what is within their eye-line (Salton, 2009; Arnoldussen & Fletcher, 2012). In effect, this allows them to see, as the presence or absence of stimuli acts to signify the presence or absence of an object, similar to the information that would be conveyed through visual stimuli.

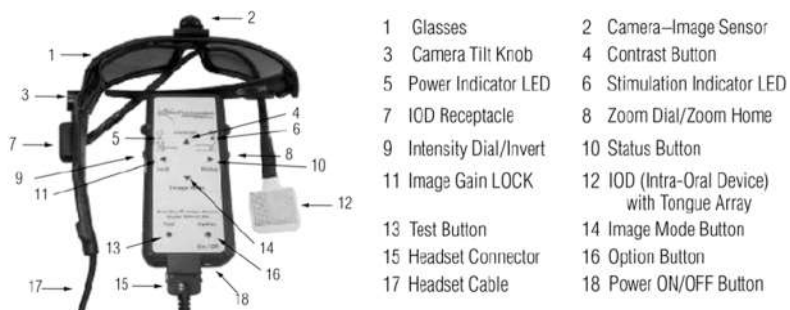


Figure 3. *Depiction of the BrainPort device hardware. (Source: Nau, Pintar, Arnoldussen, & Fisher, 2015).*

It has been shown in previous studies that portions of the brain traditionally involved in the interpretation of visual stimuli are recruited for the interpretation of stimuli from other senses when blind subjects undergo training with sensory substitution devices. Such observations have been made in the activation of the occipital cortex (Ortiz, et al., 2011), as well as the extrastriate body area of the visual cortex (Striem-Amit & Amedi, 2014), and the lateral-occipital tactile-visual area (Amedi, et al., 2007). It can therefore be seen that the brain interprets sensory information in the correct context no matter what sense is receiving the stimulation—so long as it obtains the necessary information in the correct format. Thus, although the subject is receiving tactile stimulation, s/he does not attribute the visual information that results to the tactile stimulation; rather s/he attributes it as being external to herself/himself (Bach-y-Rita & Kercel, 2003). Thus, in the use of the BrainPort, although the stimulation is received by mechanoreceptors on the tongue, it may be interpreted similarly to the interpretation of visual stimuli, as the subject becomes accustomed to the device and begins to perceive the images s/he sees as existing in space, rather than on her/his skin (Bach-y-Rita & Kercel, 2003).

The experiment outlined in the remainder of this paper will set out to test five hypotheses:

(1) *If the brains of child subjects receive an overall higher percentage of change after training with the BrainPort device, then children, whether blind or sighted, harbor more brain plasticity than adults.*

This is evidenced by previous studies which have shown that plasticity of the brain is greater at younger ages because of an excess of neurons and synaptic connections (Johnson, Nishimura, Harum, Pekar, & Blue, 2001). This would lead to the conclusion that blind children should be taught how to use sensory substitution devices at an earlier age than currently practiced.

(2) *If there is a larger percentage of change in brain activity in the congenitally blind subjects versus those subjects who lost their vision later in life, when both groups were trained, then plasticity in the brain would be greater when no prior activation of the visual cortex by visual stimuli throughout life was present.*

This would illustrate that the congenitally blind, who never used their visual cortex for the processing of actual visual stimuli presented more plasticity because their brain was freer to recruit the visual cortex for the processing of tactile stimuli, as received from the BrainPort. This is reinforced by previous work that illustrates that hypercompensation has been witnessed in blind subjects who have had tactile abilities superior to that of their sighted counterparts (Rieser, 2008). Thus, this would suggest that there are differences in the ability that different populations of blind individuals have for neuroplasticity, a topic that will be tested in this paper.

(3) *This experiment will also test the hypothesis that if the visual cortex becomes active during the PET scans taken while all subjects are participating in the shape differentiation task, then this will either be a direct result of the processing of the visual information delivered by the BrainPort or it will simply be an aftereffect of this processing.*

I will investigate what is the cause of the activation of the visual cortex by examining the PET scans.

(4) *I will also test the hypothesis that if subjects are trained with the BrainPort, then the brain plasticity that this evokes will cause a rise in sensitivity in the tactile channel(s) that are directly involved in the perception of the tactile stimuli delivered by the BrainPort.*

This hypothesis will be tested with the use of touch sensitivity measurements, including vibrotactile detection and grating orientation detection, the thresholds of which

will be expected to lower in only the tactile channels(s) that are stimulated by the BrainPort device.

(5) *The final hypothesis that will be tested will be that if tactile channels gain sensitivity after training with the BrainPort, then this activity will either be very specific to the mechanoreceptors that are stimulated throughout the training or generalized throughout the areas of the body in which these mechanoreceptors are present, such as the thenar eminence of the hand.*

Thus, I have set out with five propositions to test through experimentation with the use of the BrainPort device. To test my hypotheses. I will have twelve groups total. I will have six groups made up of congenitally blind children, late-blind children, congenitally blind adults, late-blind adults, sighted children, and sighted adults. I will break this down further into sub-groups with half of the subjects in each of the six groups receiving training with the BrainPort, and the other half receiving no training. This will add up to twelve groups in total. This will allow me to isolate all variables among subjects and come to a conclusion on each of my hypotheses.

The reasoning behind having children among the subjects in my research will be to test my hypothesis that the brain's ability to train and become accustomed to the BrainPort will be greater in youth. This is reasoned from evidence that shows that children go through a critical period in the establishment of binocular vision from the ages of 1 to 3 (Banks, Aslin, & Letson, 1975). In addition, it has been seen that there are several other critical periods involved in the visual system, with more specialized layers developing these cemented properties later than others. Overall, the critical periods for vision occur from the opening of the eyes until puberty (Daw, 1998). These findings illustrate the development of cortical attributes early in life, creating a strong justification for plasticity in young brains. Further, this gives added support for the thought that the children involved in this experiment will display a greater mastery of the BrainPort device than their adult counterparts, as their cortical properties will be more malleable.

It may be seen that aspects of vision are established in the visual cortex early in life and continue to govern the interpretation of visual stimuli as an individual matures. Lack of stimuli during these sensitive periods may lead to a deficiency in the eyesight of the child which lasts a life time. This is something that has been seen in the study of kittens, in which it has been shown that after a specific period, the degrading effects caused by a lack of visual stimuli stop being reversible due to the passing of critical periods (Hubel & Wiesel, 1970). This has also been seen in the study of strabismus, when developed early in a child's life, for which treatment before the passing of the critical period involved is key to restoring eyesight in both eyes and allowing for the development of binocular vision (Banks, et al., 1975). Thus, it can be observed that early intervention is often vital in the development of the mechanisms involved in the interpretation of visual stimuli. It may be that in blind children the same type of early intervention is necessary to allow for the sensory substitution device to become fully integrated with the same mechanisms involved in visual interpretation.

The young subjects were selected between 4 and 5 years of age to ensure that none of the critical periods related to the visual system had passed. Thus, I have decided to have the child subjects be as young as possible, while still being able to actively participate in the study.

I will be using groups made up of individuals who have congenital and later on-set blindness. This will be to allow the comparison of the percentages of difference in brain plasticity between these two groups after training with the BrainPort device. In blind individuals, the visual cortex is frequently recruited for use by the other senses. This,

however, is often of greater advantage when vision is lost at birth, whereas, if people become blind later in life, their performance in tactile tasks is much closer in sensitivity to those of their sighted counterparts (Striem-Amit, Bubic, & Amedi, 2012; Rieser, 2008). This may be due to the brain in individuals who lose their sight later in life having already become adjusted to interpreting visual stimuli received from the retinas in the visual cortex—leaving no room for said cortex to be recruited by the other senses. Thus, after the loss of vision, these patterns of having visual stimuli interpreted in the visual cortex must first be repressed for the visual cortex to begin to translate tactile stimuli as visual perception—potentially making the process of getting accustomed to sensory substitution devices more challenging and time-consuming. Thus, the other senses may not be able to become enhanced with the absence of vision during the same span of time among all blind subjects.

Moreover, I will not only be comparing brain plasticity between conditions, as mentioned before, but will also be comparing brain plasticity across age groups. This will be possible as there will be children who are congenitally blind as well as those who have lost their vision later in life. I will therefore be able to observe if having visual input early in life will have an inhibitory effect on the plasticity of the brain, even when the critical periods associated with the visual cortex have not yet passed, as compared to those subjects who were born blind and had no such visual input.

In addition, the changes that must be made in the brain to allow for the interpretation of tactile stimuli to visual sensation are made harder when the critical periods in the visual cortex, as discussed earlier, have already passed in adult subjects (Daw, 1998). Although some plasticity remains in the brain, it is not present to the extent that will allow marked changes in the hard-wiring of the mechanisms in the visual cortex, as would be possible before the passing of these critical periods.

The presence of sighted individuals in the study will act as control groups with whom to compare the changes that take place in all eight groups of blind individuals. These control groups will allow me to compare the brain plasticity that is present in both sighted children and young adults, and to see how this compares to their blind counterparts. This comparison will further enhance my view by allowing me to see if having long-term, continual visual stimuli input in the visual cortex will have a detrimental effect on brain plasticity regarding the allowed changes in the structure of the brain after training with the BrainPort.

Moreover, the thought that the regions of the brain that are traditionally used for the interpretation of visual stimuli will become active while the BrainPort is in use, as seen in my third hypothesis, is supported by previous studies that show such activation with the use of sensory substitution devices (Ortiz, et al., 2011; Striem-Amit & Amedi, 2014; Amedi, et al., 2007).

I will set out to replicate previous findings that have been seen in sighted subjects, in which the presence of mental imagery lead to the activation of the primary visual cortex (Kosslyn, Thompson, Kim, & Alpert, 1995; Chen, Kato, Zhu, Ogawa, & Ugurbil 1996). The belief that imagery will be present in blind subjects while the BrainPort is in use is supported by previous studies that have shown that visual imagery in the blind can be stimulated by haptic feedback (Renzi, Cattaneo, Vecchi, & Cornoldi, 2013). It is therefore possible that subjects who are blind, do have the ability to create mental images of what may be before them (Renzi, et al., 2013; Cattaneo & Vecchi, 2011). Thus, it may be seen that with the haptic feedback given by the BrainPort, visual imagery may be created for the subjects. This is further supported by findings from a previous study in which a group of

seven subjects that lost their vision early in life came to experience the subjective perception of light after sensory substitution training (Ortiz, et al., 2011).

There are two possible reasons for which the visual cortex becomes active during the shape differentiation task. The first of these possibilities is that this activation is an artifact of visual processing. This is supported by previous findings in which it was revealed that in the absence of the primary visual cortex, vivid mental imagery could still be experienced by a subject (Bridge, Harrold, Holmes, Stokes, & Kennard, 2012). The presence of such mental imagery in the absence of the primary visual cortex supports the idea that activation of this region is not an integral part of visual perception and that, indeed, its activation is just an after-effect of visual perception.

The second possible reason for which the occipital lobe is activated during the shape differentiation task could be that this area of the brain is fundamental to the processing of the images that are delivered through the tactile stimulus of the BrainPort. This result would be supported if the PET scans revealed greater activation in the occipital lobes of the congenitally blind subjects. This increased activation would be because these subjects have no prior experience with the processing of visual stimuli in the traditional sense of vision, as do the late-blind and sighted subjects. Such results may also lead to the increased perception of tactile stimuli as visual perception, as seen in previous studies with the use of the BrainPort (Nau, et al., 2015). This finding may also allow for greater brain plasticity to present itself in congenitally blind subjects, as they would be able to make more connections distinguishing one shape from another. This would be due to the regions of their brain, which would usually be dedicated to the processing of visual stimuli, being void of such use and instead being available for the interpretation of the BrainPort's tactile stimuli.

The percentage of change in brain activity as well as the percentage of change in the location of this activity among subjects will be measured with the use of PET scans. When undergoing a PET scan, a person is injected with a radiopharmaceutical. Radiopharmaceuticals are radioactive medicinal compounds that are injected into a patient to diagnose and provide therapy (Radiopharmaceuticals: General monograph, 2014). The individual is then placed on an examination table that slides into the PET scanner, which is a large machine that records the radioactivity emitted from the radiopharmaceutical within the body. See Figure 4. The PET machine records this activity with the use of specialized cameras, which allow for the creation of pictures that show both the structure and functioning of the organ in which the radiopharmaceutical has accumulated—which in this case will be the brain (Nordqvist, 2017). A BrainPort device compatible with the use of a PET machine will be used by the subject during the scanning process while they complete an assortment of shape differentiation tasks. Said shape differentiation tasks will be described in detail within the Methods section of this paper. These measurements will be taken throughout training with the BrainPort device and will be converted to percentages of change for each group. Thus, these percentages of change in brain activity and in the location of said activity will be compared among groups as a measurement of relative brain plasticity present among each demographic in each group tested.



Figure 4. A PET scanner. (Source: Nordqvist, 2017)

In addition, to isolate the brain activity that is directly caused by the use of the BrainPort, I will be using a subtraction technique when carrying out the PET scans. See Figure 5. This subtraction technique will consist of taking two PET scans: one while the subject is using the BrainPort device as well as one while the subject is conscious but not receiving any sensory stimulation. These two scans will then be subtracted from one another to achieve a final scan that isolates the brain activity caused by the use of the BrainPort device (Bear, Connors, & Paradiso, 2007).

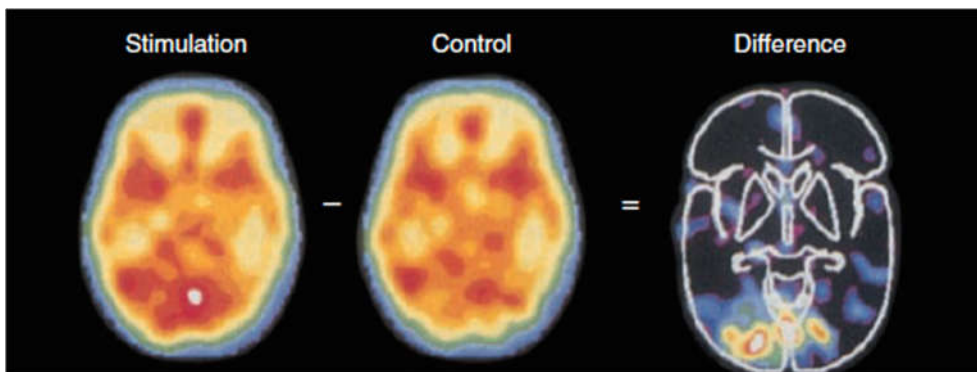


Figure 5. A PET image with the implementation of a subtraction technique. (Source: Bear et al., 2007).

In addition, two other tactile detection tasks will be used to gauge the level of sensitivity acquired after training with the BrainPort. To accomplish this, the thresholds that result from these two touch sensitivity tasks will be compared and evaluated, as explored in the remainder of this proposal. To make such measurements, each subject will participate in each of the tactile sensitivity tasks both before and after training with the BrainPort device. Thus, the sensitivity in which the subject has post-training will be measured relative to the sensitivity that the subject started out with prior to the training to measure the relative amount of improvement.

The two tactile sensitivity measurements for touch will be used to test whether generalized learning from brain plasticity occurs throughout the sense used for sensory substitution—or if such learning is isolated to specific channels. These two tactile perception measurements will be conducted on both the thenar eminence of the hand and the dorsal surface of the tongue to evaluate the extent to which such learning, if present, spreads throughout the tactile channels. This will be very easy to do as the tactile channels that are present on the thenar eminence of the hand are also present on the tongue, with the exception of the PC channel (Haggard & de Boer, 2014; Verrillo, 1966).

Vibrotactile detection thresholds will be the first tactile sensitivity measurements taken. This test will be used to determine which receptors, if any, improve in their performance after training with the BrainPort. By evaluating the extent to which the threshold of each receptor channel decreases, I will be able to determine which receptors are mainly involved in using the BrainPort. This study will also show what receptors are mainly involved in plasticity, as the results of the PET scanning will be compared with the results garnered from this test. Different contactor sizes and frequencies will be used to test each receptor channel. See Figure 6. A large 2.9 cm² contactor will be used with 250Hz vibration to stimulate the PC channel. The PC channel is optimally responsive to high-frequency vibrations, due to its ability for spatial and temporal summation as well as the large receptive fields of its Pacinian-corpusele fibers. Further, to stimulate the SA II channel, a smaller .008 cm² contactor with a 250Hz frequency will be used. The SA II channel is most responsive to the skin being stretched, making it vital in an individual's awareness to hand position. This is attributable to the channel having slowly adapting receptors that are sparingly interspersed on the skin in combination with nerve fibers with sizeable receptive fields. In addition, because the size of a contactor doesn't play a role in the vibrotactile detection thresholds of smaller frequencies, a large 2.9 cm² contactor will be used to stimulate RA receptors with a 10Hz frequency and SA I receptors with a 1Hz frequency. The RA channel is optimally responsive to tactile stimuli that are mobile over the skin, in contrast to those which are static. This is a result of the RA channel having fibers that rapidly adapt with minute receptive fields. The SA I channel is optimal in the perception of changes in the three-dimensional characteristics of stimuli, a property that is especially well suited for the reading of Braille. This is attributable to the channel having fibers that slowly adapt and have minute receptive fields (Gescheider, Wright & Verrillo, 2010).

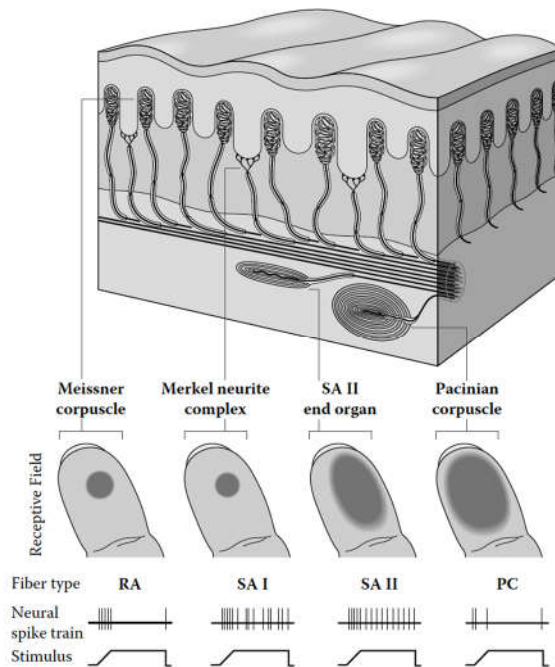


Figure 6. Illustration of nerve fibers and tactile receptors, as found in glabrous skin.
Source: (Gescheider et al., 2010).

Moreover, a grating orientation detection task will also be used to measure the relative increase in tactile function after training with the BrainPort. Through this test, I will be able to determine whether detection thresholds see a marked decrease after training with the BrainPort and how this shows variance among age groups and individuals of different conditions. These results will also be compared to those gathered from the PET scans to measure the amount of brain plasticity that allows for learning throughout the channels involved in the sense of touch.

Methods

In this research proposal, there will be five hypotheses that will be tested. The first hypothesis is that the brains of children should exhibit more changes due to training with the BrainPort device than the brains of adults who have also been trained with the BrainPort. This hypothesis is based on the notion that the brains of young children exhibit more potential for brain plasticity than would occur in adults. This has been seen in the critical periods of both kittens and adults (Hubel & Wiesel, 1970; Banks et al., 1975). I will also be testing the hypothesis that those who are congenitally blind should display larger changes in the structure of the brain because of greater plasticity, as compared to those who are sighted or have experienced late-blindness. This is because the congenitally blind have never used their visual cortex for vision. Further, I will be testing to see the cause of the activity seen in the visual cortex because of the use of the BrainPort, if such activity as has been seen with the use of other sensory substitution devices is replicated in the PET scans (Striem-Amit & Amedi, 2014; Amedi, et al., 2007; Ortiz, et al., 2011). I will also test the hypothesis that because of brain plasticity, training with the BrainPort device will lead to a rise in other

measurements of tactile sensitivity, particularly in the tactile channels that are stimulated by the pulses delivered by the BrainPort. Lastly, I will be examining whether the sensitivity gained after training with the BrainPort is specific to those mechanoreceptors that are stimulated throughout the training or is able to spread throughout the body's tactile channels. This will be examined by conducting two tactile sensitivity measurements on both the dorsal surface of the tongue and the thenar eminence, which will allow me to observe the change in thresholds in the receptors of the area being directly stimulated during BrainPort training and an area similar to that which is stimulated (Haggard & de Boer, 2014; Verrillo, 1966; Fitzpatrick, 2004). In effect, this will allow me to conclude how specific the sensitivity gained from training with the BrainPort is.

Subjects

Subjects will include 12 congenitally blind children (6 receiving training, while 6 will not). Also, included will be 12 children who have lost their sight later in life and within the time span of 1 to 2 years of age (6 receiving training, while 6 will not). Child subjects will be from 4 to 5 years of age. Subjects will also include 12 congenitally blind young adults (6 receiving training, while 6 will not) and 12 adults who have lost their sight at 20-25 years of age (6 receiving training, while 6 will not). Control groups will also be included in this study to ensure that all observations can be attributed to the etiology of the subject's loss of sight. Control groups in this study will include 12 sighted children (6 receiving training, while 6 will not) and 12 sighted young adults (6 receiving training, while 6 will not). In addition, subjects will be in equal proportions of male to female within each group. These subjects will be placed into 12 groups. See Table 1. All blind subjects will be recruited from the National Federation of the Blind for participation in the study. All subjects will give informed written consent and will be made aware of all aspects of the study. Subjects will be disqualified from the study if they are subject to any impairment that would hamper their ability to utilize the BrainPort device, such as brain trauma or malnutrition.

Table 1. Groups and demographic information.

Group	Con- genital Blindness	Later On-set Blind- ness	Sighted	Children	Young Adult	Trained	Not Trained
1	X			X		X	
2	X			X			X
3		X		X		X	
4		X		X			X
5	X				X	X	
6	X				X		X
7		X			X	X	
8		X			X		X
9			X	X		X	
10			X	X			X
11			X		X	X	
12			X		X		X

Intervention

All subjects in groups 1, 3, 5, 7, 9 and 11 will receive training with the BrainPort device. This can be seen in Table 1. These subjects will receive multi-faceted training. Training will include an introduction to the BrainPort and its components, such as the development of familiarization with the device, how to care for it, and how it works to translate stimuli from its visual to tactile form. The lack of color and depth perception when the device is used will also be discussed with each subject as an indication of the realistic limitations of the BrainPort. I will then follow a training program as outlined by established studies to maximize the ability of those subjects who will be trained with the BrainPort to build their comfort with the device (Nau et al., 2015). Thus, subjects who will be trained to use the BrainPort device will develop knowledge to recognize symbols, such as letters and numbers. They will also develop tongue-to-hand coordination, the equivalent to hand-eye coordination in this scenario, and skills in navigation while wearing the BrainPort device. In total, the training for subjects will take 6 months, during which the subjects will participate in at least two 3-hour supervised training sessions twice per day for 3 days each week. This will amount to 72 days during the 6-month period in which the subjects will complete 432 hours of training. Those in groups 2, 4, 6, 8, 10, and 12 will not be trained and will continue with their usual lives during the 6-month period.

Each subject will undergo a PET scan four times. This will occur once at the onset of the research, before any subject receives training, once two months after the onset of training, once four months after the onset of training, and again at the end of the six-month period of training. Subjects in groups 2, 4, 6, 8, 10, and 12, who will not receive any training, will also undergo the PET scans to ensure that all measurements taken are constant among subjects. During each PET scan a picture of the activity throughout the entire brain will be gathered to see what parts of the brain are activated due to the use of the BrainPort device.

Instruments

To effectively test my hypotheses, various tactile detection tasks will be used. The first involves the identification of shapes with the use of the BrainPort device in all subjects. Sighted subjects will be blindfolded during the tasks to ensure that all subjects receive equal amounts of external stimuli. Each object will be presented on the screen of a 17-inch computer and will be placed in front of a black wall to provide an environment with high contrast. Each shape will be displayed directly in front of the subject at a distance of 50 cm. The shapes identified will include the following: an ellipse, circle, square, parallelogram, triangle, rectangle, star, and the letters E, A, R, I, O, T, N, S, L, C, U, D and P. The letters will be presented with the use of 95-point Times New Roman font, resulting in letters that are 3.5 cm tall. These 20 shapes will be presented in randomized order for each of 10 trials for each subject. While the subjects complete the task, they will undergo a PET scan and will be wearing a BrainPort device that is compatible with the machine. Each subject will be given 3 minutes to identify the shape, if no correct response is spoken in this time frame, then the trial will be labeled as incorrect. Feedback will be provided for each trial, letting subjects know if they have given the correct or incorrect answer. If the answer is incorrect, they will be told what the correct answer should have been. Additionally, no ambiguous answers will be allowed. Once the answers are gathered from the shape differentiation task, the amount that each subject has named correctly and the time that they have spent in giving each correct answer will be noted.

Procedure

Position Emission Tomography Imaging

Each of the 72 subjects, whether trained or not trained, will receive position emission tomography, or PET, imaging every two-month period, as mentioned above. Position emission tomography imaging will allow for the evaluation of the flow of blood and metabolic activity within the brain.

Each subject will be injected with a radiopharmaceutical called fluorodeoxyglucose (18F) (Saha, MacIntyre, & Go, 1994). Fluorodeoxyglucose is treated similarly to glucose when it is in the body, meaning that it is easy to observe where it becomes concentrated to evaluate which regions of the brain are being used to a greater extent (Healthwise, 2015). After the injection of this radiopharmaceutical, each subject will participate in the shape differentiation task outlined above, while placed in a PET scanner that is compatible with the BrainPort. Each subject will undergo a scanning session. The manufacturer's software will be used to correct scatter, dilution, dead time, radioactive decay, and accidental concurrences, as seen in similar studies (Lee et al., 2014).

The results from the PET scan and the shape differentiation task will be considered for each subject and compared with the subject's performance in the same task in each two-month increment. The relationship between these two sets of results will be used to translate

responses into percentages of change in brain activity and percentage of change of the location of this brain activity to easily evaluate the amount of change in the structure of the brain due to training with the BrainPort. In effect, this will allow me to observe how much brain plasticity there is present in each group tested because of acquired tactile sensitivity with the sensory substitution device (Kujala, Alho, & Näätänen, 2000; Kupers & Ptito, 2004).

Additional Measurements

Additionally, each subject will also participate in two touch sensitivity tasks. These will be used to evaluate if the rewiring of the brain involved in training with the BrainPort device is specialized to one tactile channel or is generalized throughout all tactile channels. The two sensitivity tasks will be the measurement of thresholds in vibrotactile detection and grating orientation detection. These measurements will be taken on both the dorsal surface of the tongue, where the BrainPort's electrode array is placed when the device is in use, and the thenar eminence of the palm of the hand, a location similar to the tongue in regard to tactile sensitivity (Salton, 2009; Haggard & de Boer, 2014; Verrillo, 1966).

Vibrotactile Detection Threshold

The vibrotactile detection threshold measurement will be used to determine which receptors, if any, improve in performance after training with the BrainPort device. It is known that the sense of touch is composed of four distinct channels, which remain independent until they converge in the somatosensory cortex (Gescheider & Wright, 2012). A single type of mechanoreceptor and afferent nerve fiber are linked to each channel in the tactile sensory system. These channels include the RA channel, composed of Meissner capsules and rapidly adapting (RA) nerve fibers, the PC channel, composed of Pacinian corpuscles and PC nerve fibers, the SA II channel, composed of SA II end organs and slowly adapting type II, or SA II, nerve fibers, and the SA I channel with its Merkel-neurite complexes and slowly adapting type I, or SA I, nerve fibers. This specificity in the composition of each tactile channel allows for each to have a high frequency-selectivity. This in turn leads to the possibility that each tactile channel may be innervated and isolated with the careful choosing of intensity and frequency used to test subjects in specific test sites (Gescheider & Wright, 2012). Thus, I will identify which receptors are mainly involved in brain plasticity, as the results yielded from this experiment will be compared to those from the PET scans. These measurements will also be used to determine which tactile channels are improved by training with the BrainPort.

This will be accomplished by utilizing specific frequencies and contactor sizes known to activate each channel. As noted earlier, a large 2.9 cm² contactor will be used with 250Hz vibration to stimulate the PC channel. A smaller .008 cm² contactor will be used with a 250Hz frequency to stimulate the SA II channel. In addition, a large 2.9 cm² contactor will be used to stimulate RA receptors with a 10Hz frequency and SA I receptors with a 1Hz frequency. The same frequencies and contactor sizes will be used on both the dorsal surface of the tongue and the thenar eminence, with adjustments being made to make the apparatus used compatible for usage on the tongue.

Apparatus

The vibrotactile detection threshold for each subject will be found with the use of an apparatus consisting of a circular contactor shaped to fit the curves of the skin's surface. See Figure 7. Separate contactors will be created for use on either the dorsal surface of the tongue or the thenar eminence of the hand. This contactor will be bordered by a rigid

surround with a 1mm gap present between the two and will be attached to a vibrating component of the apparatus. The function of this rigid surround will be to limit the extent to which the stimulus can spread throughout the surface of the skin, thereby restraining it to the area directly around the contactor (Gescheider et al., 2010; Gescheider & Wright, 2012). The skin temperature will be kept constant by a water bath attached to the contactor, which will act to circulate water throughout the apparatus' hollow chambers. The temperature of the apparatus will be kept at 30°C. This will be implemented to reduce the role that skin temperature plays in tactile sensitivity (Gescheider & Wright, 2012). In addition, the apparatus will be altered to facilitate its usage on the surface of the tongue. Each subject will be isolated from all superfluous stimuli with the use of earphones, which will deliver narrow-band noise to conceal the sound made by the vibrator, and will be placed in a booth that blocks out nonessential external vibration and noise.

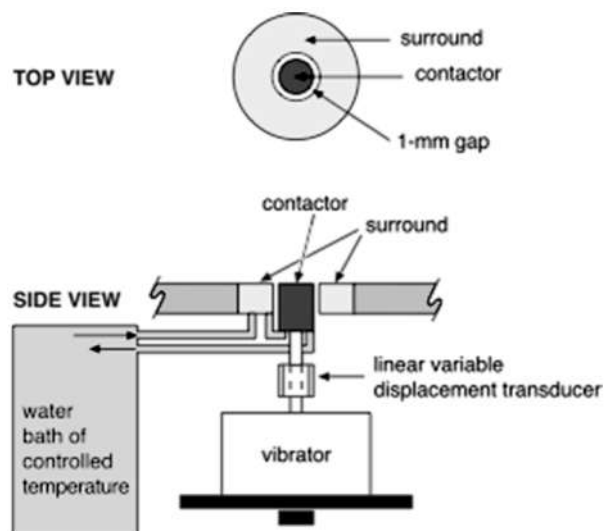


Figure 7. The apparatus used in the measurement of vibrotactile detection thresholds. (Source: Gescheider & Wright, 2012).

Measurement of Vibrotactile Detection Thresholds

The vibrotactile detection thresholds of each subject will be measured for a variety of frequencies to stimulate each tactile channel involved in the processing of tactile stimuli. These will be issued on the dorsal surface of the tongue and the thenar eminence of the hand with contactors in an assortment of sizes, as stated above. The vibrotactile detection thresholds will be measured with the use of a single vibratory stimulus, present in only one of two observation intervals presented to each subject. Each observation interval will be differentiated by lights. Each subject will indicate which observation interval contains the stimulus by pressing one of two buttons and feedback will be provided with the usage of a light, indicating to the subject if their response was correct or incorrect. The subjects will respond with the use of a two-alternative forced-choice tracking method. In addition, the stimuli will be increased by 1.0 decibel each time there is an incorrect answer and decreased by 1.0 decibel each time there are three correct answers. The stimulus intensity that leads to the subject being correct 75% of the time when detecting the presence of the stimulus will

be recorded as their threshold. These responses will include nonconsecutive correct answers (Zwislocki & Relkin, 2001; Gescheider & Wright, 2012).

The vibrotactile detection thresholds gathered from this testing will then be plotted as a function of stimulus frequency to create a threshold curve for each subject. The graphs that result prior to and after training will then be compared to one another, as explored later in this proposal, to evaluate which thresholds, if any, are lowered, and the tactile channels involved with each of these thresholds. It is with these measurements that I will be enabled to conclude whether the rise in tactile sensitivity due to training with the BrainPort is isolated to specific tactile channels or whether it spreads throughout all tactile channels. Thus, I will be able to determine which channel or channels were responsible for the learned performance on the device. This will also allow identification of those channels involved in the restructuring of the brain, as the results gathered from the vibrotactile detection thresholds are evaluated in conjunction with the changes seen in the PET scans and the grating orientation detection thresholds.

Grating Orientation Detection Thresholds

The second tactile sensitivity measurement will be taken via a grating orientation task. Thresholds for the tactile grating orientation detection will be measured by presenting grating stimuli to the subject. This grating stimulus will consist of alternating ridges and grooves that will be varied in width to measure the threshold for the discrimination of the orientation of the grating. See Figure 8. The measurements of thresholds that result will be consistent with the slowly adapting type I, or SA I, mechanoreceptors and their density on the fingertip (Johnson & Hsiao, 1992). The SA I fibers are responsible for perception of spatial stimuli on the skin, a characteristic necessary in the reading of Braille (Gescheider et al., 2010).

The tactile grating that will be presented to the subjects will consist of 25-mm square plastic blocks in which square gratings are cut. The gratings will be cut in an assortment of distances apart from one another as follows: 1.0, 1.5, 2.0, 2.5, 3.0, 4.0, and 5.0 mm. The slots cut into the blocks will be 1.5 mm in depth and their width will be half that of the distance between each grating (Johnson & Phillips, 1981).

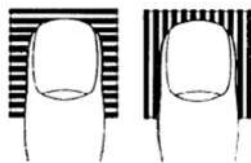


Figure 8. Diagram of the two configurations of the grating stimuli that will be presented to subjects. (Source: Johnson & Phillips, 1981).

The stimuli will be presented in one of two configurations: the grating bars parallel along the finger or the grating bars aligned perpendicular to the finger. In addition, the same grating stimuli will be presented twice in two possible sequences: with the grating bars parallel along the finger both times, or with the grating bars arranged parallel to the finger the first time but perpendicular to the finger the second time. In addition, this orientation detection task will be altered for use on the tongue. The same procedure will be followed for the use of the grating orientation task on the tongue. This will be seen, as the two configurations will also be presented in two possible sequences on the tongue: with the grating bars parallel along the tongue both times, or with the grating bars parallel along the

tongue the first time but perpendicular to the tongue the second time. On both the thenar eminence and the tongue, the stimuli will be presented passively for a set period of 2.5 s. This is in accordance to past studies that have measured grating orientation thresholds (Johnson & Phillips, 1981). The subjects will answer with the use of a two-alternative forced-choice tracking method in both variations of the orientation detection task by indicating the sequence of stimuli presented. The grating spacing that leads to the subject being correct 75% of the time when detecting the sequence of stimulus presented will be recorded as their threshold.

Analysis

For this study, the parts of the brain that are activated while the BrainPort is in use during the PET scan will be noted, specifying the location of activation in the brain and the percentage of the brain that is activated. This process will be repeated for each PET scan and the changes within each scan will be noted, as the areas of the brain activated while the BrainPort is in use during the PET scan will be expected to change throughout the training period. These measurements will be considered a measure of brain plasticity, as the structure of the brain involved in the interpretation of tactile stimuli into visual information is altered, and as each subject trained with the BrainPort will become more accustomed to the device and how it functions to convey information.

The vibrotactile detection thresholds will be gathered with stimulation to the tactile channels located on the dorsal surface of the tongue and the thenar eminence on the palm of the hand. The detection thresholds measured from each subject will then be plotted as a function of the frequency of the stimulus applied; the resulting graphs for each subject, both before and after the 6-month period, will then be evaluated with one another in conjunction with the similarly plotted function found in previous studies, which is showcased below (Gescheider et al., 2010). See Figure 9. The same comparison will be carried out on all subjects, regardless of whether they did or did not receive training, to ensure that all changes in thresholds in trained subjects are directly caused by BrainPort training.

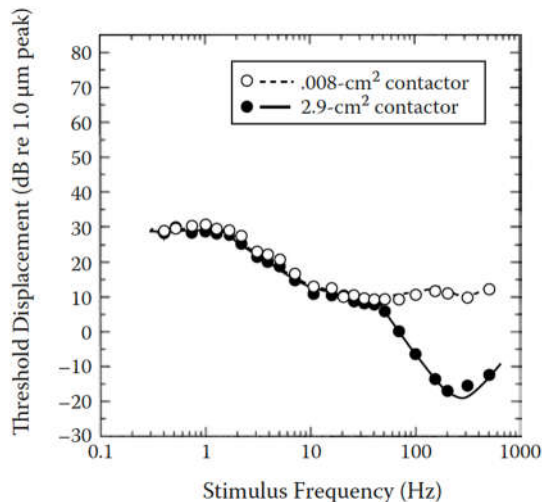


Figure 9. Detection thresholds plotted as function of the applied stimuli, as done via .008 cm² and 2.9 cm² contactors. (Source: Gescheider et al., 2010).

The four neural systems involved in touch facilitate the detection of vibratory stimuli on glabrous skin. Thus, each part of the above function is directly related to a specific tactile channel. It is because of this that I will be constructing a threshold curve for each subject, whether trained or not, both at the beginning of the training period and the end of training period. This will allow me to compare the changes seen in the two resulting threshold curves to determine the effects that BrainPort training has on the vibrotactile detection thresholds of each tactile channel.

The lowering thresholds and the effect that learning has on the channels will thereby be directly evaluated. This said, the PC fibers determine the thresholds for stimuli with a frequency above 40 Hz with the use of a large contactor. The SA II fibers dictate the threshold when the stimuli's frequency is above 100Hz with the use of a small contactor. The SA I fibers dictate the threshold when the frequency of the stimuli is between 0.4 Hz and 1.5 Hz, regardless of the size of the contactor used. The RA fibers control the thresholds when the frequency of the stimuli is between 1.5 Hz and 40 Hz with the use of a large contactor and those frequencies that are between 1.5 Hz and 100 Hz with the use of a small contactor (Gescheider et al., 2010). These specifics can be seen in more details in the graphs shown below (Gescheider et al., 2010). See Figure 10.

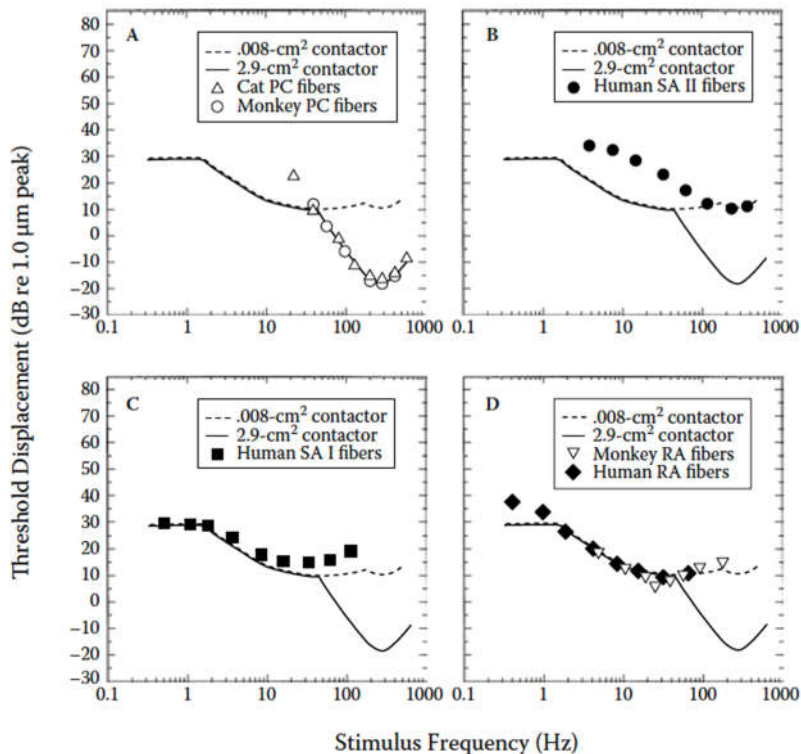


Figure 10. Graphs illustrating the optimal frequency and thresholds to which each of the four neural systems respond. (Source: Gescheider et al., 2010).

Further, the results yielded from the three tactile acuity tests will be compiled. The results taken from the PET scans will be converted into percentages of change in brain activity and percentages of change in the location of this activity. This will be done for each

scan with two month increments between each scanning session throughout the training process. The thresholds for the vibrotactile detection task and the grating orientation detection task will be noted for each subject in the study. The results from the vibrotactile detection threshold task will be used to determine which receptors, if any, improve in their performance after training with the BrainPort. This will be observed as the threshold curve noted above will be measured for both the dorsal surface of the tongue and the thenar eminence of each subject in all 12 groups before and after the six-month interval. The change in this curve will be noted, as the vibrotactile detection thresholds lower or remain unchanged among subjects. This will indicate which receptors are mainly involved with plasticity in the brain, as these receptors are trained and improved with the BrainPort device. Any discrepancy that arises between the changes in thresholds seen on the tongue and on the hand, will also display the extent to which the tactile sensitivity that comes after training with the BrainPort spreads throughout the tactile channels. Finally, the results from the grating orientation detection task will be used to determine whether the space necessary between gaps in the grating decrease in size after training with the BrainPort. How this varies among groups of subjects with different ages and visual acuity, as well as among the two different locations at which this measurement will be taken, will also be evaluated.

All these tests will be taken in unison to answer my original five hypotheses. The PET scans among groups will be used to determine whether there is heightened brain plasticity among the blind, as compared to the sighted, or among children, as compared to young adults. The PET scans will also be used to determine why the visual cortex becomes active in the interpretation of the images presented through the tactile stimuli delivered by the BrainPort. The vibrotactile detection thresholds will be used to identify which touch receptors, if any, are mainly involved in brain plasticity. This will occur via the observation of lowering thresholds among tactile mechanoreceptors. The grating orientation task, will be used in direct partnership with the vibrotactile detection task to evaluate which mechanoreceptors are involved in the changes in the brain that come with training with the BrainPort and whether lower thresholds are necessary to detect certain tactile differences. These two tactile sensitivity measurements will also be used together to evaluate the extent to which the training that comes with prolonged use of the BrainPort device spreads throughout the body. This will be done with an examination of the changes in the thresholds of the mechanoreceptors on the dorsal surface of the tongue and the thenar eminence of the hand, and a comparison of the two.

The flowchart below summarizes the experimental design of this research proposal. It also acts as a timeline for completing this study. Accordingly, 3-months will be taken at the onset of the study to ensure the recruitment of suitable subjects. A 6-month period will then be utilized for the observation of the subjects via the variety of tests outlined above. These tests will include a PET scan every two months, and the testing of the vibrotactile detection and grating orientation detection thresholds for each subject, which will occur at the beginning and end of the study. Lastly, an additional 2 months will be used for the analysis of the collected data. In total, an 11-month period will be used for the completion of this experiment. See Figure 11.

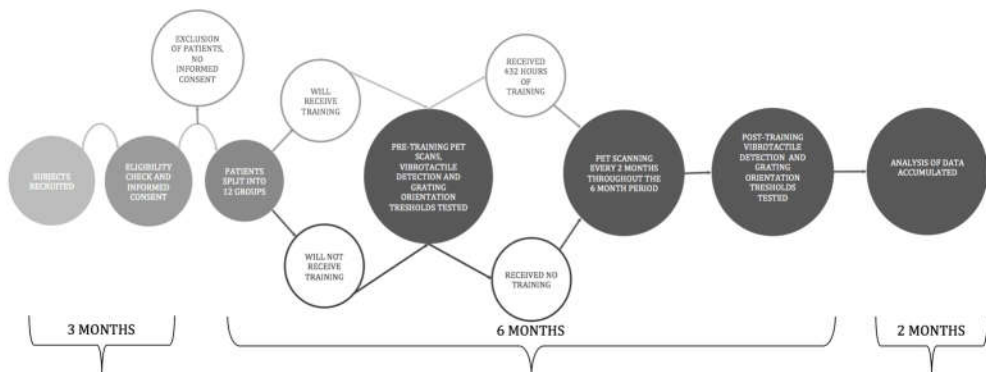


Figure 11. Outline of experimental design and period in which each stage will be completed.

Possible Results

There are multiple conclusions that may come as the result of the experiment outlined above. I will therefore assess all potential scenarios. If PET scans show no differences in the percentage of changes which occur between the brains of young adults and the brains of children, then this will illustrate that the critical periods involved in human growth and development have no effect on the plasticity evoked by the BrainPort. This would further demonstrate that the plasticity that results from training with the BrainPort is the same, regardless of the age of the individual who goes through the training. However, if the brains of the children show a heightened ability to change their wiring, as I predict in this experiment, then it will illustrate that the critical periods that children pass through as they develop do play a role. This would also show evidence that the plasticity present in younger brains is greater than that found in the brains of young adults.

A sample of the data that could be collected during this experiment is shown below. See Table 2. This table displays the frequency of brain activity observed with the use of the BrainPort. Baseline was used in the third column to represent the expected brain activity when the BrainPort was in use by a subject who received no training for how the device functioned. This baseline value was established according to past experiments in which light perception was tested in untrained subjects using the BrainPort device (Lee et al., 2014). Based on these previous findings it may be predicted that chance level for the light perception task will be found at 53.6% ($p > 0.05$). The sample data for subjects after the 6-month period of experimentation is predicted below, with adults showcasing 93.6% correct and children showing an elevated 96.6% correct in the perception of light. The chi-square value illustrates that the sample data is significant. This is seen as the chi-square value found is larger than the critical value of 7.82 for the three degrees of freedom used. See Table 3. Thus, the null hypothesis that there is no relationship between the age of subjects and their ability to perceive light with the use of the BrainPort would be able to be rejected. Further, not only are these results significant, as shown in the chi-square value below, it may also be seen that children displayed a higher improvement in the perception of light after training with the BrainPort. This is especially true when compared to the controls in the group, the untrained blind young adults and children, whom displayed very little change in their performance. If these proposed findings were to be seen in the research outlined above, it may be concluded that the brains of children do harbor more plasticity than those of adults.

Table 2. Observed and expected values for brain activation with the use of the BrainPort, as measured with PET scans.

Subjects:	Observed (O)	Expected (E)	(O-E)	(O-E) ²	(O-E) ² /E	Σ(O-E) ² /E
Trained Blind Children	96.6	53.6	43	1849	34.49626866	64.49104
Trained Blind Young Adults	93.6	53.6	40	1600	29.85074627	
Untrained Blind Children	56	53.6	2.4	5.76	0.107462687	
Untrained Blind Young Adults	55	53.6	1.4	1.96	0.036567164	

Table 3. Chi-square value derived from sample data shown above.

$\chi^2=$	64.49104
$\chi=$	8.03063

If the PET scans show no marked difference in the brain activity of the congenitally blind, acquired-blind, and sighted subjects, this will show that plasticity is equal among all individuals within a given age groups—either that of children or that of young adults in this case. This would give evidence that training with the BrainPort should not be limited to the blind, and could be used to enhance the ability of the sighted in everyday situations. If my prediction is correct, however, the PET scans taken while each group of subjects undergoes training with the BrainPort will show that congenitally blind subjects will have more changes in brain activity after training with the BrainPort. This will be because the congenitally blind have never previously used their visual cortex for the processing of visual stimuli. The implication is that their brain has no preconceived notions to overcome about what the visual cortex should be used for before allowing the area to be overtaken by the tactile stimuli delivered by the BrainPort vision substitution device. Thus, the congenitally blind will have more room for hypercompensation to develop. In addition, I also foresee that subjects who have lost their sight later in life will display levels of brain plasticity more similar to that of their sighted counterparts. This prediction is based upon the concept that both people who have acquired blindness and people who are still able to see have visual cortexes that have become accustomed to some extent to the processing of visual stimuli from the traditional sense of sight. Thus, they have less ability to develop the same type of hypercompensation as congenitally blind subjects because they have visual cortexes that have established habits which must first be broken down before allowing the tactile stimulation delivered by the BrainPort to be processed in this same area. Although training with the BrainPort may still be possible, this habit that has already been established by the brain, will elongate the time needed for training and will lessen the resulting brain plasticity in these subjects.

The examination of the PET scans will also allow for a conclusion to be drawn for why the visual cortex becomes activated while the BrainPort is in use. If the primary visual cortex becomes more activated in sighted and late-blind subjects, then it may be concluded

that this activity is simply an aftereffect of the processing of the images presented and can be attributed to the presence of mental imagery. This is supported by the results of past studies that have shown that the presence of visual imagery in the minds of subjects leads to the activation of the primary visual cortex (Kosslyn et al., 1995; Chen et al., 1996) and that in the absence of the primary visual cortex mental imagery can still be perceived (Bridge et al., 2012). It is also important to note, that these types of mental images would only be possible for individuals who have had previous experience with sight, supporting that sighted and late-blind subjects would show a more heightened activation of their primary visual cortex than the congenitally blind in this situation. However, if the visual cortex shows higher levels of activity in congenitally blind subjects, then the activation of the visual cortex is due to this region of the brain being highly involved in the processing of the tactile stimuli. This would show a great amount of brain plasticity, as the visual cortex would be used for the interpretation of tactile stimuli.

The results taken from the remaining two tactile sensitivity measurements will be used to determine the effect that training with the BrainPort has on the four tactile channels. If the results from the vibrotactile detection thresholds are lowered by the same amount for each tactile channel, this would show that the BrainPort has caused an overall increase in sensitivity throughout all tactile channels. Such a result would imply that all four of the tactile channels are involved in the processing of the pulses delivered by the BrainPort. The most likely result, however, will be that training with the BrainPort will cause a lowering of the threshold for the tactile channel that is mainly stimulated by the device, which I hypothesize to be the RA channel. Thus, the threshold found at 10Hz should be reduced to a greater extent than the thresholds measured at other frequencies. This prediction is due to the presence of RA fibers on the tongue and the sensitivity of such fibers. In addition, rapidly adapting afferents make up 2/3 of the mechanoreceptors that are present on the tongue's surface (Lozano, Kaczmarek, & Santello, 2009). The PC tactile channel can also rapidly adapt, yet this channel is not present on the dorsal surface of the tongue, where the BrainPort is placed (Haggard & de Boer, 2014; Verrillo, 1966). Both the SA I and SA II nerve fibers are present in all the soft tissues of the mouth (Haggard & de Boer, 2014). The slowly adapting characteristics of the SA I and SA II channels, however, lead me to the conclusion that they would not be optimal for the processing of the quick pulses delivered by the BrainPort. Thus, because of the relatively small area of the tongue that is covered by the BrainPort's electrode array, I hypothesize that the tactile channel that is most stimulated by the pulses delivered from the BrainPort device is the RA channel. The rapidly adapting nature of the RA channel is especially important because of the fast pulses delivered to the tongue by the BrainPort device. The small receptive fields of the RA fibers are also vitally important for the discrimination of the stimuli applied by the BrainPort. These characteristics would allow the RA mechanoreceptors to quickly sense the changes in pulses and convey this information to the brain, where it could be interpreted. Thus, the changes seen in the PET scans in the subjects who received training would be attributed to the RA channel specifically, as this would be the only one whose threshold would lower after the training.

The grating orientation detection thresholds measured will also be used to establish which tactile channels, if any, are trained with continual use of the BrainPort. The grating orientation detection task is known to stimulate the SA I channel (Johnson & Hsiao, 1992). This will therefore allow me the opportunity to evaluate whether continual training with the BrainPort leads to an increase in sensitivity within the SA I tactile channel. It may be found that the grating orientation detection thresholds become lower after the six-month training with the BrainPort. Such a finding would suggest that brain plasticity is mediated by more

than the RA channel alone. Thus, if the sensitivity of the SA I channel became heightened during this experiment, my hypothesis that the BrainPort stimulates the RA channel exclusively would be found to be incorrect. Indeed, further investigation into the role of SA I fibers during the use of the BrainPort device would be necessary. On the other hand, if the grating orientation thresholds remain uniform among all subjects, this would signify that training with the BrainPort does not cause increased sensitivity in the SA I channel. This result would also signify that the BrainPort has no effect on the tactile sensitivity of this channel within any of the groups tested, without regard to whether they are blind or sighted, or whether they received training or did not. This would give further support to show that the SA I type mechanoreceptors are not responsible for the changes that will be observed in the PET scans among subjects.

Additionally, the two tactile sensitivity measurements will be carried out on both the dorsal surface of the tongue and the thenar eminence of the hand to examine the specificity of the sensitivity acquired by the mechanoreceptors on the area of the body trained. There are two possibilities that could result from these measurements. The first possibility is that the thresholds for both the vibrotactile detection and the grating orientation detection tasks lower on both the tongue and the thenar eminence. It may then be concluded that the tactile sensitivity acquired after training with the BrainPort has translated to the tactile channels involved throughout the body. My prediction, however, is that the results will show that the thresholds for both sensitivity measurements will become lower on the tongue, but remain the same on the thenar eminence. This will demonstrate that the tactile learning that results from training with the BrainPort is extremely specific to the mechanoreceptors that are directly stimulated throughout the training process and will not spread throughout the body.

Overall, we are on the verge of bountiful technological advancements with the creation of sensory substitution devices. Not only do these new inventions allow us to restore the loss of a sense to others but they may also lead to the advancement and supplementation of senses that we already have by taking full advantage of all our faculties. It is therefore important to conduct experiments that will allow us to test different aspects of the device to examine how its use can influence the brain of the user as well as her/his sense of touch. The research outlined above would be an important step forward in our journey as it would allow us to see the optimal age and etiology of individuals that could be trained with the BrainPort. Only through continual study may we be able to conclude what the most effective use of sensory substitution devices, such as the BrainPort, will be.

References

- Amedi, A., Stern, W. M., Camprodon, J. A., Bermpohl, F., Merabet, L., Rotman, S., Meijer, P., Hemond, C. & Pascual-Leone, A. (2007). Shape conveyed by visual-to-auditory sensory substitution activates the lateral occipital complex. *Nature Neuroscience*, 10(6), 687- 689.
- Arnoldussen, A. & Fletcher, D. (2012, January 1). *Visual perception for the blind: The BrainPort vision device* . Retrieved September 4, 2017, from Retinal Physician: <http://www.retinalphysician.com/issues/2012/jan-feb/visual-perception-for-the-blind-the-brainport-vis>
- Bach-y-Rita, P., Collins, C. C., Saunders, F. A., White, B., & Scadden, L. (1969). Vision substitution by tactile image projection. *Nature*, 221(5184), 963-964.
- Bach-y-Rita, P., & Kercel, S. W. (2003). Sensory substitution and the human-machine interface. *Trends in Cognitive Sciences*, 7(12), 541-546.

- Banks, M. S., Aslin, R. N., & Letson, R. D. (1975). Sensitive period for the development of human binocular vision. *Science*, *190*(4215), 675-677.
- Bridge, H., Harrold, S., Holmes, E. A., Stokes, M., & Kennard, C. (2012). Vivid visual mental imagery in the absence of the primary visual cortex. *Journal of Neurology*, *259*(6), 1062-1070.
- Cattaneo, Z., & Vecchi, T. (2011). The importance of blindness-onset. In Cattaneo, Z., & Vecchi, T., *Blind Vision: The Neuroscience of Visual Impairment* (pp. 155-172). Cambridge, Massachusetts: MIT Press.
- Chen, W., Kato, T., Zhu, X. H., Ogawa, S., & Ugurbil, K. (1996). Primary visual cortex activation during visual imagery in human brain: A fMRI mapping study. *Neuroimage*, *3*(3), S204.
- Daw, N. W. (1998). Critical periods and amblyopia. *Archives of Ophthalmology*, *116*(4), 502-505.
- Fitzpatrick, D. (2004). Mechanoreceptors specialized to receive tactile information. In Purves, D., Augustine, G.J., Fitzpatrick, D., Hall, W.C., LaMantia, A., McNamara, J.O., & Williams, S.M. (Eds.), *Neuroscience (3rd Edition)* (pp. 192-193). Sunderland, Massachusetts, U.S.A.: Sinauer Associates.
- Gescheider, G. A., Wright, J. H., & Verrillo, R. T. (2010). *Information-processing channels in the tactile sensory system: A psychophysical and physiological analysis*. New York, NY: Taylor & Francis.
- Gescheider, G. A., & Wright, J. H. (2012). Learning in tactile channels. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(2), 302-313.
- Haggard, P., & de Boer, L. (2014). Oral somatosensory awareness. *Neuroscience & Biobehavioral Reviews*, *47*, 469-484.
- Healthwise. (2015, August 21). *Positron emission tomography (PET)*. Retrieved August 26, 2017, from WebMD: <http://www.webmd.com/cancer/lymphoma/positron-emission-tomography#1>
- Hubel, D. H., & Wiesel, T. N. (1970). The period of susceptibility to the physiological effects of unilateral eye closure in kittens. *The Journal of Physiology*, *206*(2), 419-436.
- Johnson, K. O., & Hsiao, S. S. (1992). Neural mechanisms of tactual form and texture perception. *Annual Review of Neuroscience*, *15*(1), 227-250.
- Johnson, K. O., & Phillips, J. R. (1981). Tactile spatial resolution. I. Two-point discrimination, gap detection, grating resolution, and letter recognition. *Journal of Neurophysiology*, *46*(6), 1177-1192.
- Johnston, M. V., Nishimura, A., Harum, K., Pekar, J., & Blue, M. E. (2001). Sculpting the developing brain. *Advances in Pediatrics*, *48*, 1-38.
- Kosslyn, S. M., Thompson, W. L., Kim, I. J., & Alpert, N. M. (1995). Topographical representations of mental images in primary visual cortex. *Nature*, *378*(6556), 496.
- Kujala, T., Alho, K., & Näätänen, R. (2000). Cross-modal reorganization of human cortical functions. *Trends in Neurosciences*, *23*(3), 115-120.
- Kupers, R., & Ptito, M. (2004, August). "Seeing" through the tongue: cross-modal plasticity in the congenitally blind. In *International Congress Series* (Vol. 1270, pp. 79-84). Elsevier.
- Lee, V. K., Nau, A. C., Laymon, C., Chan, K. C., Rosario, B. L., & Fisher, C. (2014). Successful tactile based visual sensory substitution use functions independently of visual pathway integrity. *Frontiers in Human Neuroscience*, *8*, 1-12.

- Lozano, C. A., Kaczmarek, K. A., & Santello, M. (2009). Electrotactile stimulation on the tongue: Intensity perception, discrimination, and cross-modality estimation. *Somatosensory & Motor Research*, 26(2-3), 50-63.
- Nau, A. C., Pintar, C., Arnoldussen, A., & Fisher, C. (2015). Acquisition of visual perception in blind adults using the BrainPort artificial vision device. *American Journal of Occupational Therapy*, 69(1), 6901290010p1-6901290010p8.
- Nordqvist, C. (2017, June 23). *PET scans: Uses, risks, and procedure*. Retrieved August 28, 2017, from Medical News Today: <http://www.medicalnewstoday.com/articles/154877.php>
- Ojala, J. (2016, November 30). *Sensory substitution: An augmented human technologies seminar paper*. Retrieved August 9, 2017, from LinkedIn: <https://www.linkedin.com/pulse/sensory-substitution-augmented-human-technologies-seminar-jouni-ojala>
- Ortiz, T., Poch, J., Santos, J. M., Requena, C., Martínez, A. M., Ortiz-Terán, L., Turrero, A., Barcia, J., Nogales, R., Calvo, A., Martinez, J.M., Cordoba, J.L., Pascual-Leone, A. (2011). Recruitment of occipital cortex during sensory substitution training linked to subjective experience of seeing in people with blindness. *PLoS ONE*, 6(8), e23264.
- Radiopharmaceuticals: General monograph.(consultation documents: The international pharmacopoeia). (2014). *WHO Drug Information*, 28(2), 162.
- Renzi, C., Cattaneo, Z., Vecchi, T., & Cornoldi, C. (2013). Mental imagery and blindness in *Multisensory Imagery* (pp. 115-130). Springer New York.
- Rieser, J. J. (Ed.). (2008). *Blindness and brain plasticity in navigation and object perception*. New York, NY: Taylor & Francis.
- Saha, G. B., MacIntyre, W. J., & Go, R. T. (1994, October). Radiopharmaceuticals for brain imaging. *Seminars in Nuclear Medicine*, 24(4), 324-349. WB Saunders.
- Salton, J. (2009, August 19). *BrainPort for the visually impaired - 'seeing' with the tongue*. Retrieved September 4, 2017, from New Atlas: <http://newatlas.com/brainport-sight-device/12551/>
- Striem-Amit, E., & Amedi, A. (2014). Visual cortex extrastriate body-selective area activation in congenitally blind people “seeing” by using sounds. *Current Biology*, 24(6), 687-692.
- Striem-Amit, E., Bubic, A., & Amedi, A. (2012). Neurophysiological mechanisms underlying plastic changes and rehabilitation following sensory loss in blindness and deafness. In Murray, M.M. & Wallace, M.T. (Eds.), *The neural bases of multisensory processes* (pp. 402-404). Boca Raton, Florida: Taylor & Francis Group.
- Verrillo, R. T. (1966). Specificity of a cutaneous receptor. *Attention, Perception, & Psychophysics*, 1(3), 149-153.
- White, B. W., Saunders, F. A., Scadden, L., Bach-Y-Rita, P., & Collins, C. C. (1970). Seeing with the skin. *Perception & Psychophysics*, 7(1), 23-27.
- Zwislocki, J. J., & Relkin, E. M. (2001). On a psychophysical transformed-rule up and down method converging on a 75% level of correct responses. *Proceedings of the National Academy of Sciences*, 98(8), 4811-4814.



Descartes on Other Minds: Human and Animal

Hao Yu

Author Background: Hao Yu grew up in China and currently attends Beijing National Day School in Beijing, China. Her Pioneer seminar topic was in the field of philosophy and titled "Descartes' Meditations."

1. Introduction

In his work *Meditations on First Philosophy*, Descartes attempts to prove that knowledge of the mind and God is more transparent and solid than knowledge of the corporeal world. However, there appears to be one issue that is in between our mind and the outside world and it is assumed to be true without proof in Descartes' work: the existence of other minds. It is evident that Descartes did not consider this a particular issue, since in his *Second Replies*, he acknowledges that "in my Meditations...my supposition was that no other human beings were yet known to me" (CSM II 102), but he breaks out of the lonely world he constructs for himself in the First Meditation in the following Meditations, using first-person plural several times without explanation.

The objective of this paper is to examine how Descartes responds to the skepticism about other minds and their implications, and how we can understand his definition of mind, body, and soul.

In the first section of the paper, I take the important definitions and conclusions in Descartes' previous works (mainly from the *Meditations* and the *Discourse on the Method*) related to the issue to construct an argument on Descartes' grounds to prove the existence of other human minds and discuss his position that animals do not have minds. In the next section, I will discuss some possible objections and weaknesses of the possible proof of Descartes' and his other arguments presented in the first section, and will try to offer responses from Descartes' perspective. In the last section, I will examine the implications of multiple minds on Descartes' works, and suggest that we should keep an open mind as to the question of whether animals have minds.

2. Descartes' Stance on Other Minds

2.1 Human Minds

In order to find a stable foundation for his beliefs, Descartes, in his *Meditations*, poses three sets of doubts that directly challenge them. He believes the ideas that remain intact under these doubts shall be certain and indubitable (CSM II 12). Although sense perceptions are not trustworthy, and even mathematical theories can be originated from a deceiving demon, there still seems to be one certain thing for Descartes – that he is thinking (doubting). Since a subject is needed for an action, Descartes is then entitled to say that "I think, I am." (CSM II 17) He then uses his own existence as a basis and proves the existence

of a benevolent and non-deceiving God in *Meditations III*.¹ He then derives a rule by which he could prove the existence of other things, a rule that says since God does not deceive, whatever we clearly and distinctly understand must be true. This proposition is clearly stated in part 4 of his *Discourse*:

I observe that there is nothing at all in the proposition ‘I am thinking, therefore I exist’ to assure me that I am speaking the truth, except that I see very clearly that in order to think it is necessary to exist. So I decided that I could take it as a general rule that the things we conceive very clearly and distinctly are all true; only there is some difficulty in recognizing which are the things that we distinctly conceive (CSM I 127).

He again summarizes this point in *Discourse*:

...our **ideas** or **notions**, being real things and coming from God, cannot be anything but true, in every aspect in which they are clear and distinct (CSM I 130).

Thus, in *Meditation VI*, Descartes argues, since he is led to believe that the ideas of other objects really do come from the outside world, and that God is no deceiver, there must be real material objects in the world:

...God has given me no faculty at all for recognizing any such source for these ideas; on the contrary, he has given me a great propensity to believe that they are produced by corporeal things. So I do not see how God could be understood to be anything but a deceiver if the ideas were transmitted from a source other than corporeal things. It follows that corporeal things exist (CSM II 55).

Following this line of argument, in order to prove other minds’ existence, the question becomes: since other minds are not directly perceivable, how do we conceive clearly and distinctly that there is a mind residing in the human body? Although Descartes did not answer this question directly, in his *Discourse on the Method*, he offers ways that we can distinguish machines and animals from humans, and those methods in turn offer us insights as to how he may make it clear that a human body possesses a mind.

The first criterion he gives is the capability of producing meaningful language:

We can certainly conceive of a machine so constructed that it utters words, and even utters words which correspond to bodily actions causing a change in its organs (e.g. if you touch it in one spot it asks what you want of it, if you touch it in another it cries out that you are hurting it, and so on). But it is not conceivable that such a machine should produce different arrangements of words so as to give an appropriately meaningful answer to whatever is said in its presence, as the dullest of man can do (CSM I 140).

He thinks that a machine could simulate all the physical movements and even reactions a human body has, but it cannot assemble words and phrases to produce an

¹ I remain doubtful about his proofs but will discuss this issue under the presumption that God exists.

appropriate and meaningful response to what's said to it. He believes the reactions from the machine that we observe can be produced merely by the specific changes "in its organs," which is conceivable since we are dealing with a robot designed completely like a human body and can do all physical responses. Descartes also believes that a human's bodily functions depend on its organs, but the organs cannot account for every human action. That is, the humans end their similarities with machines when meaningful language begins.²

The second criterion states that even though the machine or animal may outperform a human being in some tasks, it inevitably fails in others:

[E]ven though such machines might do some things as well as we do them, or perhaps even better, they would inevitably fail in others, which would reveal that they were acting not through understanding but only from the disposition of their organs. For whereas reason is a universal instrument which can be used in all kinds of situations, their organs need some particular disposition for each particular action; hence it is morally impossible for a machine to have enough different organs to make it act in all the contingencies of life in the way in which our reason makes us act (CSM I 140).

From the fact that machines outperform us only in some areas, Descartes believes it is obvious that they do not have any intelligence, because otherwise they would have higher intelligence than the human and outperform us in every task. For example, a machine may do particularly well in its preset areas, but cannot do as well outside of this range. It is also concluded that machines do not have thorough understanding 'but [they act] only from the disposition of their organs' (CSM I 140). Descartes is essentially reminding us that the rearrangement of bodily organs has a limit, while the mind is a "universal instrument" which is useful in performing all kinds of tasks.

To apply the two criteria to proving the existence of other minds is to say that if a person can convey his (or her) ideas clearly with language or signs so that another person (a human being with a mind) understands it,³ and that the person acts out of understanding but not the disposition of organs (both need the subjective judgment from the observer), he or she is then clearly conceived of as having a mind. With the proven presumption from Descartes' *Meditations*, God is no deceiver and what is clearly and distinctly conceived in this case — that the other people have minds — is true.

2.2 Minds of Other Animals

Before examining the topic of Descartes' opinions about whether animals have minds, an important distinction must be made regarding Descartes' definition of mind and soul.

It is a common reading that Descartes says that animals do not possess souls. This kind of reading starts as early as in the *Objections*, in which one of Descartes' opponents writes, "as far as the souls of the brutes are concerned, M. Descartes elsewhere suggests

² It recalls us to the Turing Test widely known nowadays, and there have actually already been scholars who work to point out that Alan Turing may have got his idea for the test from Descartes.

³ "...it is quite remarkable that there are no men so dull-witted or stupid — and this includes even madmen — that they are incapable of arranging various words together and forming an utterance from them in order to make their thoughts understood" (CSM I 140).

clearly enough that they have none. All they have is a body which is constructed in a particular manner, made up of various organs..." (CSM II 144). Even recent scholars hold similar notions; as Anita Avramides writes in her article "Descartes and Other Minds", she thinks that:

He wants to restrict the term "mind" to cover this thinking part of man. According to Descartes, mind is "the whole of the soul which thinks". There is to be no more equivocation: *mind is soul*. As a result, the **lower animals** who grow, are nourished, and move in various ways are considered by Descartes to *have no soul*, it having been established that they *have no mind* (*Teorema* Vol. XVI/1, 1996, pp. 33).

She bases her notion on a very important line — "...the whole of the soul which thinks" — from a passage in Descartes' Reply to Objection V, also cited in her article:

Thus, because probably men in the earliest times did not distinguish in us that principle in virtue of which we are nourished, grow, and perform those operations which are common to us with the brutes apart from any thought, from that by which we think they called both by the single name soul... But I, perceiving that the principle by which we are nourished is wholly distinct from that by means of which we think, have declared that the name *soul* when used for both is equivocal... *I consider the mind not as part of the soul but as the whole of that soul which thinks* (Descartes (1969), p. 210.) (*Teorema* Vol. XVI/1, 1996, pp. 32).

Notice how Avramides equates the mind with the soul and concludes that by stating animals have no minds, Descartes is saying that they have no souls as well. However, it is important to offer a different reading of the line: instead of saying the mind is equal to the soul, what Descartes really means is that there is a kind of "thinking soul" which is distinct from other forms of souls, and that thinking soul is equivalent to the mind. In his works, Descartes has many times the difference between the mind and "soul"; for example, in his Second Set of Replies, he offers definitions of mind as:

The substance in which thought immediately resides is called mind. I use the term 'mind' rather than 'soul' since the word 'soul' is ambiguous and is often applied to something corporeal (CSM II 114).

Here he makes clear distinctions between souls and minds. Again, in the *Sixth Set of Replies*, he directly refers to the same distinction when he is discussing animals' minds:

My critics go on to say that they do not believe that the ways in which the beasts operate can be explained 'by means of mechanics without invoking any sensation, life or soul' (I take this to mean 'without invoking thought'; for I accept that the brutes have what is commonly called 'life', and a *corporeal soul* and organic sensation) (CSM II 288).

It is then manifest that instead of saying that animals have no souls, Descartes is more committed to proving that there is a distinction between the animals' souls and human souls — one is merely a corporeal soul, the other is a "rational and immortal soul," using the

language test and the capability test mentioned in the above section. He generalizes one of his fundamental assertions that “*all and only human animals have [rational] souls*”⁴ to every kind of animal and every man:

... it is quite remarkable that there are no men so dull-witted or stupid — and this includes even madmen — that they are incapable of arranging various words together and forming an utterance from them in order to make their thoughts understood; where as there is no other animal, however perfect and well-endowed it may be, that can do the like. This does not happen because they lack the necessary organs ... that is, they cannot show that they are thinking what they are saying (CSM I 140).

He is certain about his belief that animals have no minds, that is, have different kinds of souls from humans, when writing this passage in his 1637 work *Discourse on the Method*:

... This shows not merely that the beasts have less reason than men, but that they have no reason at all.... it would be incredible that a superior specimen of the monkey or parrot species should not be able to speak as well as the stupidest child ... if their souls were not *completely different in nature from ours* (CSM I 140).

His tone is so assured that it makes me wonder why it is that Descartes is so determined to fight against these “things we have become convinced of since our earliest years” (CSM II 162).⁵ A possible answer is found in Avramides’ work “Descartes and Other Minds” in which she states that “there is evidence that Descartes was driven by theological and moral considerations to deny souls (that is, minds) to all non-human animals”⁶. As he writes at the end of part V of the *Discourse*:

For after the error of those who deny God, which I believe I have already adequately refuted, there is none that leads weak minds further from the straight path of virtue than that of imagining that the souls of the beasts are of the same nature as ours, and hence that after this present life we have nothing to fear or to hope for, any more than flies and ants (CSM I 141).

He then concludes:

But when we know how much the beasts differ from us, we understand much better the arguments which prove that our soul is of a nature entirely independent of the body, and consequently that it is not bound to die with it. And since we cannot see any other causes which destroy the soul, we are naturally led to conclude that it is immortal (CSM I 141).

⁴ (*Teorema* Vol. XVI/1, 1996, pp. 37). I made changes based on different interpretation of Descartes’ definition of ‘soul’, but I appreciate Avramides’ overall point here.

⁵ He is referring to the common belief that animals have souls that are the same as that of human beings.

⁶ This is also Gunderson’s conclusion in Gunderson (1971), p. 15

Descartes is dismissive of the idea of animal immortality; in a letter to More he writes:

... it is more probable that worms, flies, caterpillars and other animals move life machines than that they all have immortal souls (CSM III 366).

How much impact the desire to prove the immortality of the human soul had on Descartes' belief that animals do not possess rational soul is another topic. Here, I would like to point to a change in Descartes' assertiveness from his 1637 work *Discourse* and his six sets of Replies in 1641 where he states clearly that "...beast...have no reason at all" (CSM I 140), to his agnostic statement in his letter to More in 1648, where he seems to have softened his stance:

But though I regard it as established that we cannot prove there is thought in animals, I do not think it can be proved that there is none, since the human mind does not reach into their hearts (CSM III 365).

What causes his agnostic statement shall be an interesting question to investigate since he used to firmly state that animals do not have mind and used the distinction of human and animal soul as an evidence for assurance that humans' souls are immortal, and we would not completely cease to exist as mere caterpillars would after death. Another piece from Avramides further extends a sense of agnosticism on the minds of humans. Avramides changes the above quote to:

But though I regard it as established that there is thought in human animals, I do not think it is thereby proven that there is, once the human mind does not reach into their hearts (*Teorema* Vol. XVI/1, 1996, pp. 33).

3. Objections and Descartes' Possible Replies

With Descartes' ideas that God is no deceiver and what we can clearly and distinctively perceive is true, there seems to be no doubt that once we clearly know that the other person passes the language test and can act out of understanding, the person is proven to have a mind. The doubt, however, is in the criteria. What is to be perceived as having linguistic capability? Should a feral child who doesn't speak human languages be considered as having a mind? Further, does Descartes' firm attitude against animals having minds that is supported mainly by the fact that they cannot "speak as well as the stupidest child" (CSM I 140) imply that he thinks language is the necessary outcome, if not necessary component, of thought? This section will briefly discuss these questions and Descartes' possible replies.

3.1 The Language Test on Human Minds

3.1.1 Feral Child

There have been numbers of reports about children that are raised by or have grown up with only animals and could not communicate with humans after they were found; some of them even fail to acquire language in their later life,

An example would be the Russian "bird-boy," who was found living in an aviary with only cages containing dozens of birds and had no access to human interactions. It was reported that he "does not understand any human language and communicates instead by chirping and flapping his arms" (Cockcroft, 2008).

The child would most likely not be able to pass a human language test, as he has no means by which to be understood by other people. Should he then be considered to have no

mind? A similar case would be an infant who has not yet acquired linguistic capabilities. Considering the case of an infant, Descartes once replied in his letter to More:

Infants are in a different case from animals: I should not judge that infants were endowed with minds unless I saw that they were of the same nature as adults; but animals never develop to a point where any certain sign of thought can be detected in them (CSM III 374).

It seems that since infants always develop to the point that they acquire language and thus show themselves as having the same nature as adults, Descartes believes that they are different from animals and have human minds. However, the questions that cases of feral children pose are slightly different: should a child be considered to have a human mind if s/he never acquire linguistic capabilities? If the answer is positive, should their communication with animals in their early ages be seen as an evidence that animals have minds too?

3.2 On the Statement That Animals Don't Have Minds

3.2.1 Animal Communication

Many would argue that the waggle dance of bees can be used as evidence that animals have language and thus do have minds. As can be observed, honey bees that have successfully found food sources for the hive describe the direction and distance to the flowers by performing waggle dance to the members of the colony. The difference in distance is illustrated by performing either a round dance – indicating shorter distance – or a waggle dance – indicating longer distance.⁷ Even though the system seems primitive, what happens in this system of communication is actually a process of one side encrypting the facts into a symbol, displaying the symbol, and the other side demonstrating comprehension. The ability to communicate a message through a symbol seems to indicate that both parties have at least some kind of intelligence. Descartes is going to argue that the kind of communication that uses only signs and no complicated grammatical structures is completely imitable by machines, just as what the computers can now do. There should thus be no indication of the ability to think from those kinds of communications. His argument here, despite the fact that it's counterintuitive, is considered persuasive.

3.2.2 Should We Think That Animals May Develop Languages That Humans Can Understand?

Even if some animals display higher levels of linguistic skills than bees, Descartes discards them as evidence of minds.

Nor should we think, like some of the ancients, that the beasts speak, although we do not understand their language. For if that were true, then since they have many organs that correspond to ours, they could make themselves understood by us as well as by their fellow (CSM I 141).

In this passage, he proposes that if animals can indeed devise language, then because of the similarities between animal organs and human organs, it must be true that their language can be understood well by humans. His argument here seems like an easy way out of the

⁷ Grüter and Farina, (May 2009). "The honeybee waggle dance: can we follow the steps?"

potential objections brought by the different kinds of animal speech. Descartes is arguing that since we couldn't understand any of them, the animals are not actually producing meaningful language.

However, two flaws can be found in this argument. First, from the mere fact that animals have similar organs to humans it does not necessarily follow that they must develop similar language systems to our own. Similar to humans, cats and dogs have four limbs, but our limbs are used in different ways; for example, they walk with four limbs and we walk erect with two. Therefore, similar organs do not indicate same functions. Secondly, it is not necessarily true that their languages cannot be, or do not have the potential to be, understood by humans. There are animal trainers who can communicate with animals better than other people could, and they can even distinguish the animals' different needs; for example, they can tell whether the animals are sick or hungry from their cries. Those people excel at communicating with animals because they have spent a significant amount of time with the animals trying to understand them. Humans have different kinds of languages and can't understand people who speak different languages at first, until they spend time learning each others' languages. We also observe that people have great difficulty mastering a new language once they pass their critical period of language learning. There is the possibility that the reason humans cannot understand animals' languages is just that when we tried to learn animal languages, we have already passed the critical period of language learning.

3.2.3 The Ape Language Experiment

While there are few cases in which humans have been able to acquire animal language and translate it for other humans, there have been experiments that have shown that apes can learn human language. In an experiment done by Savage-Rumbaugh, a young bonobo named Kanzi learned to communicate with humans with non-pictorial lexigrams. The researchers first attempted to teach Matata, Kanzi's adoptive mother, to use lexigrams on a keyboard to communicate, but that was no success. During his childhood, Kanzi often accompanied Matata when she was taught lexigrams, and to the researchers' surprise, when Matata was sent away to mate one time, Kanzi, in her absence, began to use the keyboard spontaneously. He had learned by watching attentively when Matata was taught, and had displayed far more competence than Matata had. One of the first things Kanzi did with the keyboard was to activate "apple", then "chase". Kanzi's behavior was recorded by the experimenter in his book *Kanzi: The Ape at the Brink of the Human Mind*: "He then picked up an apple, looked at me, and ran away with a play grin on its face" (*Kanzi: The Ape at the Brink of the Human Mind*, 1994, p.135). It appeared that Kanzi not only understood what was meant by each symbol, but could actively use the symbols to communicate what he wanted. The researchers also noticed that nearly all of Kanzi's multiword utterances were spontaneous. These spontaneous utterances, different from responses to questions, give insight as to what Kanzi is thinking. More importantly, Kanzi displayed the ability to understand spoken English, which is unique among other apes tested. For example, he would turn off the lights if he heard someone mentioned turning the lights off, even though there was no direct instruction asking him to do so (*Kanzi: The Ape at the Brink of the Human Mind*, 1994, p148).

With the assistance of Patria Greenfield, a linguist at the University of California, Los Angeles, the group was able to show that Kanzi's use of lexigrams was grammatical: the five criteria for presence of language structure: symbols must have independent use, there must be a semantic relationship between symbols, the relationship needs to be specified by rules between categories of symbols, categories must be related by a formal device, and the rule must be productive (pp158-9) were all satisfied by Kanzi's use of

lexigrams, according to Savage-Rumbaugh. The researchers believe there is evidence for at least rudimentary syntactic structure in Kanzi's use of symbols and they refer to the syntax as "protogrammar" as the rules are rather simple.

Descartes takes the use of language as evidence for the possession of thought and takes animals lack of genuine language use to support that they have no minds. If there is evidence to show that bonobos have the ability to acquire a rudimentary form of language, it certainly undermines Descartes' argument for animals' lack of minds which relies on a factual claim that no animals have been observed to possess language and, as would follow, thought. The Kanzi experiment should make us question the belief that humans have a unique and supreme capability of thinking and open us to the possibility that we are not all that different from other animals, regardless of whether this may mean that "after this present life we have nothing to fear or to hope for, any more than flies and ants" (CSM I 141).

There is insight to be gained by comparing the feral children and the ape experiment. Before they were found, the feral children never needed to learn human language; they survived by learning from and communicating with the animals in their "primitive language." And after they were found, most of them had passed their critical period of language learning and could not fully grasp human language in their later lives. Infants have the ability to produce all kinds of sounds. It is only after they have learned their mother tongue that they lose the ability to produce sounds belonging to foreign languages because they have no need for those sounds and have gone through no significant practice. (For example, it is difficult for an adult second language learner to grasp the Spanish alveolar trill). It could be a similar case with the apes. Apes may have the ability to grasp complex language, but since they have no use for it, the ability is lost during their childhood. As illustrated by the experiment, Kanzi's mom, who was already an adult ape at the time of the experiment, did poorly learning the lexigrams, while Kanzi grasped the use of lexigrams quickly. Is language then, a necessary component or outcome of thought, or is it the other way around, that language only comes when there is a need to communicate individual thoughts among the group? How language use may serve as a criterion for whether one has a mind or not may need further examination.

4. Conclusions and Implications

This paper constructs a possible proof of the existence of other minds within Descartes' framework. The paper also notes Descartes' position that human minds are unique, that the human soul differs from the animal soul and thus we have a different fate from animals in the afterlife. Objections and questions are raised related to both the proof of the existence of other human minds and the animals' lack of mind.

While it is reassuring to know that an individual is not a lonely island floating in space with only "illusions" of other minds like one's self's using the language test proof, it seems that the language test is flawed when seen as proof that animals do not have minds. Although it is tempting to believe in human minds' uniqueness, we must not let emotions deter our clear judgments and be open to the possibility that animals have minds just as we do, until further research is conducted and questions arising from the contradictions between current evidence and theories are answered.

It is also important to notice, however, the implication of the existence of other minds on the credibility of one important supposition of Descartes' which the paper takes to be true during the construction of the proof of other human minds. If feral children shall be considered as having minds, how could Descartes know that they have a same idea of God as he, or even be sure that they have an idea at all of God? Can one person's idea of God be

used as a substantial evidence for God's existence when it cannot be known whether other persons, or say, other minds, have the same idea? Then, if God's existence is not certain, where should the proof of the existence of other minds, which depends on the non-deceiving character of God, stand?

As the superiority of human and the existence of God are brought into question, there seems to be an even greater unknown lying ahead. What is truly important, however, as philosophers throughout time have believed, is not the mere feeling of certainty but the courage to step back from whatever comfort that sense of human superiority brings and to keep doubting in order to get closer to the truth.

References

- Avramides, Anita (1996). "Descartes on Other Minds" in *Teorama*, Vol.XVI/1, pp. 27-46
- Cockcroft, Lucy (Feb 2008) "Russian 'bird-boy' discovered in aviary". *The Telegraph*. telegraph.co.uk/news/worldnews/1580159/Russian-bird-boy-discovered-in-aviary.html. Retrieved 2017-9-21.
- Descartes, René (1986-91) *The Philosophical Writings of Descartes*, Vol. I, II and III, eds. and trans. John Cottingham, Robert Stoothoff, Dugald Murdoch and (for vol. III) Anthony Kenny. Cambridge: Cambridge University Press.
- Grüter, Christoph; Farina, Walter M. (May 2009). "The Honeybee Waggle Dance: Can We Follow the Steps?". *Trends in Ecology & Evolution*.
- Savage-Rumbaugh, S., and Lewin, R. (1994), *Kanzi: The Ape at the Brink of the Human Mind*, John Wiley & Sons, New York.



Deviation and Integration: How *Nüshu* Serves as a Centrifugal and Centripetal Force for Women in Changing Rural China

Yilin Chen

Author Background: Yilin Chen grew up in China and currently attends The Experimental High School Attached to Beijing Normal University in Beijing, China. Her Pioneer seminar topic was in the field of anthropology and titled “Communication and Culture.”

1. Introduction

多少红颜薄命死	How many beautiful women die sad;
多少终身血泪流	How many of them shed tears throughout their lives...
新华女子读女书	We read <i>nüshu</i>
不为当官不为名	Not for power, not for fame,
只为女人受尽苦	But because we suffer.
要凭女书诉苦情	We need <i>nüshu</i> to lament our grievances and our bitterness (Liu, 2015).

Nüshu (女书), translated as “women’s script,” was a script devised and practiced by local women in Jiangyong County, Hunan Province in the south of China for several centuries in the late imperial to early modern period. Regarded as the world’s only gender-specific script, *nüshu* was only revealed to the outside world in the 1980s. By then, it was already on the brink of extinction. *Nüshu* is a phonetic system of writing derived from the local dialect (Lo, 2012). Writings in *nüshu* are found on paper, paper fans, booklets, and some are embroidered on handkerchiefs (Lo, 2012). Their social functions take the form of prayers, letters between friends and sworn sisters (explained in Section 4.2), wedding missives, biographies or autobiographies, and folk songs. The *nüshu* writings were composed almost entirely in highly formulaic verse form, because most of them were intended to be sung or chanted.

Since the discovery of *nüshu*, anthropologists, ethnologists, and linguists have studied this topic with great interest. Because *nüshu* is a gender-specific language phenomenon, it provides invaluable insight for the study of women’s status in the androcentric society of pre-contemporary rural China. Moreover, researching *nüshu* sheds light on the largely unexplored inner-voices of Chinese peasant women. This has important implications, as other texts that explored women’s lives were often written by male or upper-class writers using the official Chinese language. These are mostly detached accounts and do not necessarily reflect social reality.

This paper aims to examine how *nüshu* simultaneously serves as a centrifugal and centripetal force in its historical and demographic context. Here, in accordance with mainstream research papers on the subject, *nüshu* has three meanings: 1) the script itself, 2) works of art and literature created using the script, and 3) objects which have the script inscribed on them (Nie, 2006; Tian, 2004). In using the terms “centrifugal” and

“centripetal,” I aim to discuss how *nüshu* divided and unified people, namely, how it deviated from orthodox Chinese culture (i.e. acting on its centrifugal force) while strengthening the group identity and social participation of peasant women (i.e. exhibiting its centripetal force).

In this paper, I will first discuss women’s overall status in pre-contemporary rural China. I will then go on to analyze the centrifugal and centripetal effects of *nüshu* respectively. Lastly, I will explain how *nüshu* gradually approached its extinction as the need for both forces diminished. The primary sources I have surveyed are all documented in Xie’s collection of written *nüshu* works (1991), and translated by Liu in her research paper (2004b).

2. Historical Context and Women’s Status in Pre-Contemporary Rural Jiangyong

Scholars have asserted that *nüshu* was practiced exclusively in Jiangyong County, Hunan Province, South China. The known geographic *nüshu* area, with a rice-farming economy and a population of roughly twenty thousand, was surrounded by mountains on three sides and therefore isolated (Lu, Jia, & Helsey, 2002). Ethnically speaking, Jiangyong can be seen as a boundary land where Han Chinese and members of the Yao ethnicity intersect (Liu, 2004b). This demographic gave rise to a peculiar intermixing of regional customs. Some features were consistent with Han ideologies, such as the Confucian patriarchal system, but other cultural practices, such as the marriage residence pattern, were in accordance with Yao traditions (Lu et al., 2002). These features contributed to Jiangyong’s distinctive cultural norms, establishing the foundation for *nüshu* to flourish.

The Yao ethnic group, with some of its branches still retaining elements of matrilineal societies, has a high regard for female deities and celebrates numerous festivals organized by women and centered around women. However, the Han’s patriarchal system dominates in the Jiangyong region. The Confucian ideals dictate that a virtuous woman should possess “three obediences (三从)”: obey her father before marriage, her husband when married, and her sons in widowhood. This is a system of relations which pre-contemporary Chinese women depended on for social status, identity, and subsistence (Liu, 2004b). Under this overarching principle, women were subordinate, considered inferior, and powerless. Ruled by the “cult of womanhood,” women were treated as property that could be traded into their affinal homes (Lu et al., 2002).

The most common marriage tradition for Yao people is *buluofujia* (“不落夫家”), referred to as “delayed transfer marriage” (McLaren, 2001) or “delayed patrilocal residence” (Liu, 2004b), with its literal meaning being “not falling into the husband’s home.” This means that a woman does not move into her husband’s residence until the birth of her first child. This directly contradicts the Han marriage system, which “posits a definite rupture at the time of marriage” (McLaren, 2001). According to Liu (2004b), a mass migration of Han Chinese from the north to the Hunan region occurred in the seventh century, leading to a gradual sinicization¹ among the Yao people. The Han customs of arranged marriage and patrilineality slowly became accepted by sinicized Yao, as the women in the region became increasingly dependent on men.

In pre-contemporary rural China, women in the peasant class were almost universally denied access to education. As an old Chinese idiom states, “an unaccomplished woman is a virtuous woman (女子无才便是德).” In China, literacy has long been regarded as a fundamental step on the ladder of success and a facilitator of social mobility (Liu, 2004a). In contrast, illiteracy was often equated with inferiority and lack of intelligence. Therefore, through the restriction of women’s access to education, the Confucian ideology of female inferiority to their male counterparts was sustained. The lack of literacy forever restricted

these women to their domestic spheres and ruled out any political or social activism. In addition to this, women were restricted by the age-old tradition of foot-binding, which physically limited their mobility. This practice advocated the idea of confinement embedded in Confucian principles, which states that a woman's rightful place is in the "inner quarters," doing needlework and embroidery. In part due to this practice, women did not work in the fields unless they were severely impoverished. Instead, unmarried girls were referred to as *loushangnü* (楼上女, or "upstairs maidens") because they spent most of their time doing needlework with their peers in the upstairs chambers of a house (Nie, 2006). Because they were forbidden to interact freely with men, a segregated female society was created, which also contributed to the emergence of *nüshu*.

3. *Nüshu* as a Centrifugal Force

3.1 *Nüshu*'s Deviation from *Hanzi*

Standard Chinese, or *hanzi* (汉字), is one of the oldest continuous writing forms in the world. Among the myriad of differences between *nüshu* and *hanzi*, two characteristics are most prominent, the first being character shapes. The Chinese often describe *hanzi* characters as "square-block characters" (方块字) for their uniform size and shape, while *nüshu* was variously referred to as "tadpole-text" (蝌蚪文), "mosquito-leg script" (蚊脚字) and the like, largely because it resembled the shapes of insects and tiny animal forms (Liu, 2004b). *Nüshu* consists of rhomboid-shaped characters with slanted lines, arcs, circles, and dots.

The second difference lies in syntax. Unlike the texts written in Standard Chinese, *nüshu* writings have no punctuation and an inverted grammar. In Chinese, most modifiers come before the modified element (the head), acting as premodifiers. However, in *nüshu*, it is very common for modifiers to appear after the head, acting as postmodifiers (Xie, 1992). Here are two examples of this phenomenon. The first example is "biological mother" (Fig.1)




<i>Nüshu</i> :			Standard Chinese <i>hanzi</i> :		
					
母	亲	生	亲	生	母亲
mu	tsai	soi	qin	sheng	mu qin
mother	blood-related	give birth	blood-related	give birth	mother

Figure 1. The *nüshu* characters are taken from Zhou (2002), and this comparison of grammar is based on the work of Xie (1992). In accordance with mainstream research on the subject, I present the phonetics of *nüshu* in the International Phonetic Alphabet (IPA). The phonetics of *hanzi* is represented by pinyin, the official romanization system for Standard Chinese.

The second example is "dark night" (Fig.2).

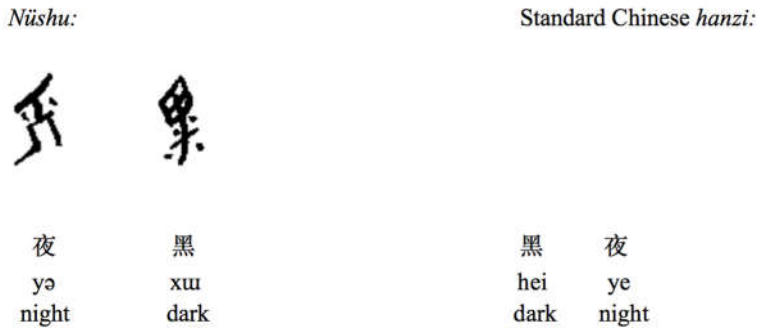


Figure 2. The nüshu characters are taken from Zhou (2002), and this comparison of grammar is based on the work of Xie (1992). The use of the IPA and pinyin is consistent with Figure 1.

Why did the women in Jiangyong develop a script so utterly different and foreign, both in shape and syntax? First, it should be realized that, from the beginning, *hanzi* was a device for men, developed by men, and intended to enact and maintain a Confucian patriarchal social structure. It was created through male ideals and their superiority in language use. According to Ma (2007), in the process of creating and standardizing Chinese characters, women had always been a form of property in the patriarchal structure. The radical² “女” (female) gave rise to countless characters, describing everything from a woman’s beautiful complexion to her gossipy nature, while “男” (male) only generated three words to describe family relations. As time went by, many of these female-related characters became evaluative terms for both sexes, further valorizing women. It is clear then that women, by using the official language, are unconsciously submitting to a set of symbols in which their voice remains unheard.

In 1975, anthropologists Edwin and Shirley Ardener developed the “Muted Group Theory.” They contend that the arena of public speech is typically male-dominated, and that in order to be heard, an individual (in this case, the woman) must use the dominant mode of language (Wall & Gannon-Leary, 1999). Similarly, Virginia Woolf lends insight to this problem by lamenting that “before a woman can write exactly as she wishes to write, she has many difficulties to face. ...the very form of sentence does not fit her. It is a sentence made by men” (Woolf, 1979: 48). In fact, Jiangyong women called official *hanzi* characters *nanshu*, or “men’s script” (Liu, 2004b). *Nüshu* was not the only effort in history for women to counter their muteness, but other such phenomena did not have a comparable impact. Although both Japanese *hiragana* and Korean *hangul* were once considered to be feminine, over time they were incorporated into the official script and became accepted by both sexes (Tian, 2004). In contrast, rather than “appropriating the subjugated language” (Lo, 2012), women in the Jiangyong region were able to speak up for themselves. Through their fine penmanship and reversed grammar, they found a way to express their sentiments without the interference of a male-dominated script. This deviation from *hanzi* has enabled them to detach themselves from the dominant androcentric discourse.

3.2 The Extent to which *Nüshu* Alienates Women

The gender specificity of *nüshu* brought with it its reputation of being a “secret language” among women, yet this is erroneous. According to Liu (2004a), *nüshu* was

widely visible and audible in a variety of social gatherings and occasions; men could always learn to read and write the script if they wanted. However, women's chanting of *nüshu* was characterized as "mad" women's articulation, resting on the edge of culture where they were the "sorceresses" (Lo, 2012: 385). It was due to men's disdain for the script, not women's deliberately exclusive efforts, that *nüshu* became male-illegible. In fact, these women longed for the day when their script could be appreciated, and their personal struggles be told. In other words, though *nüshu* has empowered women to expand female connections beyond the confines of male-derived familial ties (Liu, 2004a), it also serves as evidence that women had failed to be heeded or understood by men. Just like *hanzi* excluded one sex, *nüshu* segregated women from men.

With few exceptions, *nüshu* practitioners were born in the first half of the nineteenth century. Those who were literate in *nüshu* stopped using it in the Cultural Revolution (1966-76), during which *nüshu* was condemned as "witch's script," because residents of Jiangyong burned *nüshu* writings when their owners passed away (Liu, 2004b). As suggested by its name, a significant part of the Revolution included ransacking cultural sites and destroying all artifacts and historical remnants that reflected the traditional China before Mao's regime. As the movement swept through the country, *nüshu* was relegated to "the Four Olds": namely, old customs, culture, habits, and ideas. Letters and embroidery artifacts were deliberately burnt (Song, 2013), which dealt a severe blow to women, as they were publicly humiliated and discouraged from using the script. The additional connotation of witchcraft further degraded Jiangyong women back to their original state, in which they could not find a way to communicate their miseries.

4. *Nüshu* as a Centripetal Force

4.1 Formation of Strong Bonds between Sworn Sisters

In 1993, a coin that was made during the Taiping Rebellion was discovered. It was stamped with a *nüshu* phrase: women in the world are sisters in a family (天下妇女姊妹一家). Two major genres exist in *nüshu* writings between sworn sisters: *jiebai* sisterhood letters and *sanzhaoshu* wedding literature. Because the tone and content of *sanzhaoshu* vary according to the relationship between the sender and the recipient, to maintain a contrast with sisterhood letters, the *sanzhaoshu* presented in this subsection will mainly include examples of texts written by peers of brides, while the messages from senior relatives will be discussed in Section 4.2.

Jiebai ("结拜", sworn sisterhood), mainly practiced in southern China, was a distinctive custom of Jiangyong. Two kinds of non-kin *jiebai* relationships were most common: general *jiebai* and *laotong* (老同, "very same"), or *tongnian* (同年, "same year"). The category of general *jiebai* was more inclusive, with no specific qualifications required to start the relationship (Liu, 2004b). This *jiebai* sisterhood could be found between two or more girls of different ages, preferably with similar socioeconomic backgrounds, and could be initiated in both childhood and adulthood. *Laotong* or *tongnian*, on the other hand, involved two girls of the same age. Unlike other formalized relationships (husband/wife, father/son, elder/younger brother, teacher/student) in Confucian ideology, the sworn sisterhood was "the only non-hierarchical dyad grounded in sameness and equality" (Silber, 1995: 88).

In sisterhood letters, *nüshu* could be used to express sorrow and protest, as evidenced by a letter Cizhu Hu (胡慈珠) received from her sworn sister, Nianhua Yi (义年华);

The two women in the example below still used the script in the 1980s when scholars began to formally study *nūshu* (Xie, 1991; Liu, 2004b).

到他三年上四载，	Three or four years after I married him
见下女儿是一人。	I had a daughter.
丈夫出乡入书院，	[At that time] my husband went to study in the township;
我在堂前奉双亲。	I stayed home and waited upon my in-laws.
三餐茶水多端正，	I served them three meals a day and brought them cups of tea with propriety.
孝顺父母尽我心。	I did my best to be as filial as possible.
三从四德也知礼，	I observed the Three Obediences and Four Virtues; ³
忠孝两全父心欢。	My father-in-law was satisfied with my loyalty and filiality.
父亲出门墟场去，	[One day] when my father-in-law went to the periodic market,
谁知母亲说丑言。	My mother-in-law spoke ill of me:
枉我暗中煮蛋吃，	She accused me of cooking eggs for myself on the sly.
哪个神仙来证明。	Which god would prove me innocent?
日夜与娘同房睡，	I slept in the same room with her at night.
时刻不离娘的身。	I never left her during the daytime.
哪有何尝有此话，	How could I possibly do such a thing?
无人伸冤来证明...	But no one could prove my innocence...
自想自叹伤心哭，	I sobbed and I grieved.
几夜天光愁断肠。	For several nights I stayed up with pent-up sadness.
又想放中自缢死，	I wanted to hang myself.
难舍红花女一人。	But I couldn't bear to part with my daughter.

To an outsider, “cooking eggs on the sly” may seem trivial. However, it is considered an unacceptable conduct because an egg, with its nutritious and symbolic value, was a rarity in pre-contemporary China reserved for illness and giving birth (Liu, 2004b). Nianhua understood this tacit logic, and could only resort to her *jiebai* because no one, not even a god, could appeal on her behalf. Her only spiritual outlet rested in her sworn sister, whom she depended on for emotional support.

In return, *nūshu* could also be used to comfort the grieved sender of the letter. Cizhu Hu was also in a *jiebai* relationship with Baozhen Tang (唐宝珍). In 1974, when Baozhen's husband died, Cizhu wrote a *nūshu* letter on a handkerchief to comfort her. This letter, presented below, expresses her grief and solidarity with her sworn sister (Xie, 1991; Liu, 2004b).

把笔写书双流泪，	Holding a pen, I write with double-flowing tears,
急跨回家劝妹娘。	Write to comfort my sister, anxiously.
你夫落曹几个月，	Your husband has been dead for several months.
不得回程疼惜声...	I, however, was unable to go to pay my condolences...
句句实言来劝你，	Each word I am advising here is truthful.
知理妹娘听言章。	My reasonable younger sister, please listen to me:
不气丈夫缘分浅，	Don't complain that you had a shallow relationship with your husband;
落地三声注定来...	It was predestined, before we were born...
不气丈夫寿已过，	Don't be mad that your husband is no longer alive;
六十有余正终身。	He lived up to his sixties.
夫死阴曹免得虑，	He died with no worries at all;
子女个个交全啦。	All of his children are married...

At the end of this correspondence, Cizhu lamented her own pitiful situation, a typical practice in *jiebai* sisterhood letters (Xie, 1991; Liu, 2004b).

劝声妹娘将完了，	After giving these words to my younger sister,
再讲姐娘一段情...	Let the older sister talk about her own situation...
非常回家无出气...	That she has no home to return to to release her sentiments...
依其结交几姐妹。	That is why she made sworn sisters.

As we can see, *nüshu* letters between *jiebai* sisters can be better described as an echoing of despair. When young brides grew up often to become the in-law that they once dreaded, only sworn sisters could understand each other's sufferings and emotional turmoil. By reverberating their *jiebai*'s misfortunes, they were drawn closer together into a tightly-knit social subgroup to re-establish a sense of belonging and spiritual support.

The second category of sisterhood correspondence was *sanzhaoshu* or *hesanzhao* (literally "third morning letter" or "congratulation on the third day"). Usually taking the form of booklets, *sanzhaoshu* was produced in the month leading up to a wedding when all the bride's *jiebai* sisters gathered in the upstairs chambers. *Sanzhaoshu* narrated the departure scene, the relationship between the authors and the bride, the bride's new role, and instructions to help her transition (Lu et al., 2002). Contrary to what the name suggests, *sanzhaoshu* written by *jiebai* sisters were more of a lamentation than a congratulation: exogamous marriage invariably renegotiated sisterhood relationships as the girls became separated (Lu et al., 2002). Below is an example of a *sanzhaoshu* between two *jiebai* sisters (Xie, 1991; Liu, 2004b).

人家要虑着，	Please remember that even though you are now in someone's family,
照归在以前。	Our relationship should not be, as such, changed.
透想眼泪落，	The more I think, the more my tears flow.
独自冷哭愁…	I sob, cold and alone…
也要搁开做，	But still I have to put everything aside and write;
书本记千年。	Write this book as our testimony for thousands of years.

Instead of terminating *jiebai* relationships, marriage of a sister actually led to a pledge of eternal fidelity. As the spatial distance between two sisters increased upon marriage, spiritual distance decreased with this *sanzhaoshu* testimony. In Liu's 2004(b) essay, she quoted a Jiangyong man who remarked, "in sum, a *sanzhaoshu* was a woman's letter to break off relations," in an effort to help her understand the overall picture. However, this assertion is one-sided and only reflects the dominant male perspective. Quite on the contrary, *sanzhaoshu* actually facilitated future correspondence. A significant portion of sisterhood letters came after separation, when the sender lamented her misfortunes to the recipient. This was exactly the aim of the script — to offset geographical distance, thus maintaining *jiebai* ties even after marriage (Liu, 2004b).

4.2 The Strengthening of Women's Group Identity

Group identity refers to a person's sense of belonging to a certain group. *Nüshu* succeeded in strengthening Jiangyong women's perceived identity in four ways. It provided more occasions for women to gather, documented collective memories in the female sphere, contributed to group history through biographies and autobiographies, and maintained female obedience from generation to generation.

First of all, activities that included the circulation of *nüshu* works brought women together more often and more effectively. In the creation process of *nüshu*, women (especially unmarried girls) would gather in the upstairs chamber of a house during group needlework, spinning, weaving, and embroidering. For them, this was an important part of socializing. *Nüshu* authors often referred to this kind of group activity through such phrases as "For every thread and every string, we consult with one another" (Zhao, Zhou, & Chen, 1992: 92). When one member of a group left to be married, her peers often lamented, "I have the needle and thread ready at hand, but no one to ask" (Zhao et al., 1992:51).

Nüshu texts written on paper or paper fans or embroidered on handkerchiefs used to be valuable gifts exchanged between Jiangyong women. A good *nüshu* work would be much contested as eager girls all wanted to be the first to read or recite the text (Xie, 1991). In other words, *duzhi* (读纸, reading *nüshu* from paper), *dushan* (读扇, reading *nüshu* from paper fans), and *dupa* (读帕, reading *nüshu* from handkerchiefs) all became favorite pastimes of Jiangyong women (Xie, 1991). At social events like *chuiliangjie* (吹凉节, "cooling days") or *nüerjie* (女儿节, "girls' day"), which were festivals celebrated exclusively by women, they gathered to exchange embroidery techniques and share *nüshu* works. Because there was no official school to instruct *nüshu*, women took advantage of such occasions to teach each other and transmit *nüshu* orally.

This centripetal force brought together peasant women of humble origins and loose organizations together into a cohesive group. Through singing, chanting, embroidering, and writing, they formed a unified body that could find solace despite their hardships in marriage or widowhood.

Nüshu was also used to record memories of past events in the region, including traumatic memories of war and suffering, and the tumultuous years of the early socialist period of the 1950s (McLaren, 2013). One such text recorded the Taiping Rebellion (1851-64), describing how difficult it was during the war to flee with bound feet and how much they suffered from the death of husbands (Xie, 1991; Fan, 1996). They also had to take care of children and cultivate the lands to feed the family. According to Fan (1996), this was the only history of Taiping written from a woman's personal perspective. Jiangyong women also documented numerous major events of the People's Republic of China, 1949 to 1958.

The women in this tiny village in China, like many of their sisters all over the world, had long been excluded from official history. History is, after all, written by the literate and powerful (Fan, 1996). *Nüshu* provided these women with the tool to record history from their own perspectives, and to narrate their collective stories which had hitherto been omitted from most historical documents.

In addition, innumerable biographies and autobiographies written in *nüshu* contributed to a collection of life narratives of women. In fact, according to Lo (2012), some 80% of the found *nüshu* manuscripts are biographies and autobiographies, with the latter outnumbering the former. Liu Feiwen is a Taiwanese anthropologist who has done sustained research on *nüshu* over the past two decades. In one of her research papers (2004a), she describes her interview with a *nüshu*-literate woman, Tang, during one of her field research experiences. Tang explained the reason for composing her own biography in *nüshu*: "I want people to know that I have suffered" (Liu, 2004a: 432). This reflects the majority of Jiangyong women's aims when creating life narratives – to document their sorrows and miseries. Often, even those women who were illiterate in *nüshu* would recite their accounts for someone to document.

Though both autobiographies and biographies usually consist of life stories of one individual, it is inevitable that they will also include details about the lives of others. One does not live in isolation; therefore, one woman's auto/biographic lamentations reflect similar sentiments present in her surroundings. The collaboration between *nüshu*-literate and illiterate women further accelerated the creation of a group history, transforming individual memories to collective ones.

However, the centripetal force of *nüshu* sometimes affected women negatively. Through the circulation of *nüshu*, especially *sanzhaoshu*, an implicit consensus was ultimately reached within the social group of women – that a woman should be meek, obedient, and polite. It should be noted that though women expressed their sadness in marriage or widowhood through *nüshu*, they never discussed overturning the ingrained Confucian principles. Here, in one *sanzhaoshu* text, an elder woman imparts to a bride the proper behaviors after marriage.

While *sanzhaoshu* texts sent by a bride's *jiebai* express grief upon parting, those written by senior relatives of the bride's natal family tend to endorse the androcentric system. They contain advice aimed at convincing the brides to silently accept their fates and to "give their complete loyalties to their husbands' families in order to demonstrate a proper upbringing" (Liu, 2004b). By passing on similar *sanzhaoshu* from mother to daughter, the notion of wifely submission was enforced from generation to generation. This shared group identity should not be overlooked, because it explains female obedience in feudal China, as this social norm was maintained externally by men and internally by women.

5. *Nüshu*'s Demise

Several factors contributed to *nüshu*'s demise, including the implementation of the New Marriage Law in 1950, the burning of *nüshu* works upon death, and the "Destroying the Four Olds" campaign of the Cultural Revolution (Lu et al., 2002). Because the last two were already mentioned in previous sections, this analysis will focus mostly on the first factor.

Traditionally, most peasant families "put great pressure on the wife to bear a son or face the humiliation of accepting a concubine or other forms of emotional abuse and divorce" (Fan, 1996: 102). Chinese marriage had often been forced or arranged, and women could not seek divorce. The New Marriage Law in 1950 was a radical change from existing Chinese patriarchal traditions. It dictated that the marriage system should be based on the free choice of partners, on monogamy, and on equality between the two sexes (Marriage Law of the People's Republic of China, 1950). The law also liberalized divorce, leading to an unprecedented wave in which six million married couples filed for divorce ("Marriage Law of 1950: a revolution of concepts and regulations," 2012). For the first time, the Chinese realized that women had the right to divorce their husbands, while previously they could only cling to unhappy marriages, with the risk of being divorced at any moment.

This initiated a turn of events marking the gradual decline of the oppressive conditions that fostered *nüshu*. As 20th-century China progressed toward gender equality, foot-binding was terminated, and women joined the workforce. Because *nüshu* was primarily used for bridal lamentations and recounting marriage misfortunes, as these progressive movements chipped away at extreme patrilineality, the need for the centrifugal and centripetal effects of *nüshu* diminished, leading to its demise.

6. Conclusion

This paper analyzes the tangible effects *nüshu* had on women in the mountainous region of Jiangyong County, south China. As a centrifugal force, *nüshu* deviated significantly from Standard Chinese and challenged prevailing social norms, creating boundaries between the two sexes which further alienated women. However, this unprecedented women's script also helped them communicate their previously unexpressed selves. As a centripetal force, *nüshu* wove the women into a tightly-knit community with strong bonds such as *jiebai* sisters, providing them with an outlet to express their pent-up sorrows. The script documented unheard-of stories written from female perspectives, revealing sufferings within this neglected social group. *Nüshu* enabled these women to create, write, embroider, and chant their personal narratives, which strengthened their group identity.

As presented in the verse at the beginning of this paper, women read *nüshu* "...because [they] suffer." Studying *nüshu* sheds light on peasant women's struggles in a changing rural China before the modern period: they were alienated from society, yet this only created stronger bonds between themselves.

It is a great pity that this beautiful script is not likely to be circulated anymore. However, viewing this phenomenon more positively, it indicates that in contemporary China, women are incorporated into greater social interactions where *nüshu* is no longer needed. While women in China merge to become "sisters in a family" and gradually acquire gender equality, the struggles of Jiangyong women never cease to reverberate thanks to the remarkable script of *nüshu*.

Notes

1. Sinicization refers to the process whereby non-Han Chinese societies come under the influence of Han Chinese culture and society.
2. Radicals (部首, literally “section header”) are graphical components of Chinese characters under which the characters are traditionally listed in a Chinese dictionary. The radical is often a semantic indicator, meaning that it generally tells the reader what the character is related to. For example, in the character “妈 (mother)”, which is composed of “女 (woman)” and “马 (horse)”, while “马” is a phonetic component, the radical of “女” indicates that the character has associations with femininity.
3. The Four Virtues (四德) refer to morality, proper speech, modest manner and diligent work, and is one of the most widely accepted spiritual fetters on Chinese women in a feudal society.

References

- Fan, C.C. (1996). Language, Gender, and Chinese Culture. *International Journal of Politics, Culture, and Society*, 10(1), 95-114.
- Liu, F. (2004a). From Being to Becoming: *Nüshu* and Sentiments in a Chinese Rural Community. *American Ethnologist*, 31(3), 422-439.
- Liu, F. (2004b). Literacy, Gender, and Class: *Nüshu* and Sisterhood Communities In Southern Rural Hunan. *Nan nü: men, women, and gender in early and Imperial China*, 6(2), 241-282.
- Liu, F. (2015). *Gendered Words: sentiments and expression in changing rural China*. Oxford: Oxford University Press.
- Lo, Y. (2012). She Is One and All - Writing *Nüshu*, Women's Script in an Autobiography. *New writing*, 9(3), 374-395.
- Lu, X., Jia, W., & Helsey, D.R. (2002). *Chinese Communication Studies: Contexts and Comparisons*. Westport: Greenwood Publishing Group, Incorporated.
- Marriage Law of the People's Republic of China. (1950). Retrieved from http://www.law-lib.com/law/law_view.asp?id=43205
- “Marriage Law of 1950: a revolution of concepts and regulations.” (2012). Retrieved from <http://roll.sohu.com/20120216/n334962420.shtml>.
- Ma, J. (2007). “女人与”汉字的男性化 [“Female” Person and Virilization of Chinese Characters]. *甘肃联合大学学报 (社会科学版) [Journal of Gansu Lianhe University (Social Sciences)]*, 23(3), 96-99.
- McLaren, A.E. (2001). Marriage by Abduction in Twentieth Century China. *Modern Asian Studies*, 35(4), 953-984.
- McLaren, A.E. (2013). Women's Script Compositions in China: Recording Collective Memories. *International Journal of the Book*, 10(1), 51-61.
- Nie, C. (2006). 论女书流传地女性习俗的传播学意义及社会功用 [On the communicative significance and social functions of womanhood traditions in *nüshu*'s birthplace]. *湖南科技学院学报 (Journal of Hunan University of Science and Engineering)*, 27(2), 33-35.
- Silber, C. (1995). *Nüshu (Chinese women's script) literacy and literature*. Unpublished doctoral dissertation, University of Michigan.
- Song, Y. (2013). 生态博物馆：江永女书文化的保护与传承 [Eco-museum: Jiangyong *nüshu* protection and inheritance]. Unpublished masters thesis, Guangxi Teacher's Education University, Guangxi, China.

- Tian, L. (2004). “江永女书及其女性文化色彩” [*Nüshu in Jiangyong and its women’s culture*]. 中华女子学院学报 [*Journal of National Women’s University of China*], 16(4), 23-27.
- Wall, C. J., & Gannon-Leary, P. (1999). A sentence made by men: Muted group theory revisited. *European Journal of Women's Studies*, 6(1), 21-29.
- Woolf, V. (1979). *Women and Writing*. London: Women’s Press.
- Xie, Z. (1991). 江永“女书”之谜 [The Mysteries of Jiangyong *nüshu*]. Henan: 河南人民出版社.
- Xie, Z. (1992). “女书语法中的百越语底层” [The substratum of Baiyue language in the syntax of *nüshu*]. 民族语文 [*Ethnic languages*], 4, 16-24.
- Zhao, L., Zhou, S., & Chen, Q. (eds.) (1992). 中国女书集成 [*The collection of Chinese nüshu*]. Beijing: 清华大学出版社.
- Zhou, S. (2002). 女书字典 [*Nüshu Dictionary*]. Hunan: 岳麓书社

Note: “Marriage Law of the People’s Republic of China” (1950) and “Marriage Law of 1950: A Revolution of Concepts and Regulations” (2012) are translated from the original document/article.



Divided by Discrimination: An Analysis of Stereotyping and Differential Behavior during a Shopping Expedition

Gabrielle Battle

Author Background: Gabrielle Battle grew up in the United States and currently attends The College Preparatory School in Oakland, California. Her Pioneer seminar topic was in the field of anthropology and titled "Nonverbal Communication."

Abstract

The purpose of this paper is to examine racial discrimination through nonverbal communication. Black people are often more likely to encounter negative outcomes as a result of implicit bias and stereotypes than Whites. For this study, Black and White teenagers were sent to an upper end department store in a predominately White community to shop for items and interact with people in the community. Their interactions with store clerks and adult shoppers were examined in order to determine if there were any differences in the way teens were treated based on the race of the teens, the adult shoppers with whom they interacted, and the department store clerks. The results demonstrated that Black teenage boys and girls were often ignored and treated with hostility while their White counterparts were treated pleasantly by shoppers and were welcomed into the store where they received excellent customer service. This paper shows that Black people are discriminated against based on preconceived notions and stereotypes.

Discrimination

Discrimination comes in different forms and has many impacts. Discrimination can display itself when the only African American student in a classroom has her hair continually touched and played with without permission, or it can manifest itself by a disproportionate number of African Americans in jail for crimes that both White and Black people commit at similar rates. As a young African-American woman, I find that racial discrimination is extremely pervasive in my everyday life. The discrimination that I have faced, however, has not always been verbal. Discrimination can manifest itself in many forms, such as through nonverbal communication. By examining White Americans' nonverbal communication with African-Americans and White teenagers, I concluded that the implicit biases of White Americans in shopping environments often result in hostile and frightening treatment of African Americans that can create and perpetuate a culture of division.

Compared to situations that can be understood verbally, nonverbal communication allows people to have a more profound understanding of both the people and environment around them. Edward Hall, ground-breaking anthropologist, lays the framework for analysis of the hidden meanings of nonverbal communication presented in this paper. Mark Knapp and Hall, professors and authors of *Nonverbal Communication in Human Interaction*, a book that analyzes nonverbal communication and its effects, state that

“the process of receiving nonverbal messages, including our own (“Why is my fist clenched when he’s around?”), includes giving meaning to or interpreting those messages. (This process will be defined later as *decoding* a nonverbal message). As a receiver of nonverbal messages, you may focus on one particular nonverbal cue or several in an attempt to understand the message that another person has sent to you”.¹

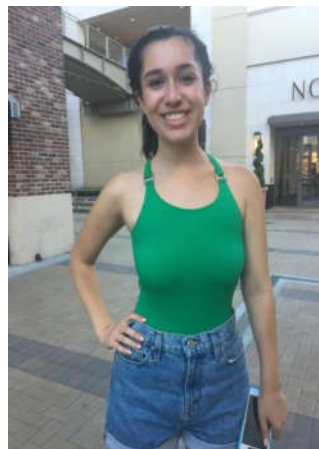
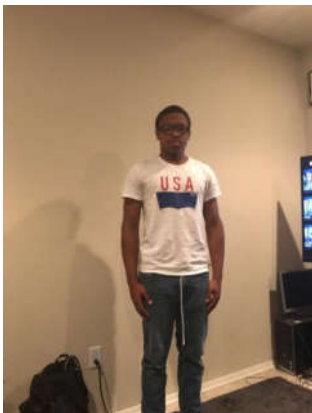
This framework fosters discussion on the meaning of messages that are conveyed through people’s actions. This paper specifically analyzes racial discrimination through nonverbal communication. According to Stanford’s Encyclopedia of Philosophy, discrimination is defined as “any viable account of what discrimination will regard it as consisting of actions, practices, or policies that are—in some appropriate sense—based on the (perceived) social group to which those discriminated against belong. Moreover, the relevant groups must be ‘socially salient’ as Kasper Lippert-Rasmussen puts it; i.e., they must be groups that are ‘important to the structure of social interactions across a wide range of social contexts” (2006: 169).² This definition creates a lens for examining discrimination.

Methodology

I created a field study in which I used proxemic, emotional, and kinesic markers such as eye contact, body distance, posture, tone, etc. to analyze the messages that people sent to Black and White teenagers when interacting with them. I worked with African-American and White teenagers (ages 15 to 16) to conduct different field studies in a popular shopping center in Walnut Creek, California. Walnut Creek was the perfect backdrop for my field study because according to the 2010 census, the city is 73.5% White. In this field study, people’s actions allowed me to assess their implicit biases; I wanted to conduct my field study in a non-diverse area to truly examine people’s implicit biases towards a group of people with whom they would not commonly interact. In my paper, I analyze several interactions of two basic field studies, and I examine how simply changing the test subjects’ race led to different outcomes of interaction. A city comprised mainly of White people who do not frequently interact with people of color allowed me to more deeply analyze and identify their implicit beliefs and natures.

The teenagers in the field study had to complete several tasks such as asking for directions and shopping for items in stores. Certain factors were controlled such as the time of day the teenagers went to the Walnut Creek shopping center and the clothing they wore. Clothing is one of the obvious external nonverbal factors of communication allowing people to express themselves and their beliefs, so it was crucial that all test subjects were conveying the same message. All boys wore white t-shirts and blue or grey pants. The first time the field study was conducted, the teenage girls wore blue shirts and shorts, and the second time the girls wore yellow shirts and jeans. I also tried to find volunteers with the same relative height and weight. The teenagers were also instructed on specific things to say and do with their bodies when interacting with people to ensure that they were also conveying the same message to people with whom they spoke. The field studies that were most influential and successful were those in which the youths asked for directions, as well as when they asked for help in a specific department store.

Volunteers



Field Observations 1

The first field study required the teenagers to approach White men and women who were alone as well as in groups and ask them for directions to a nearby popular coffee shop. When a White teenage boy (WB) interacted with a White man, he was well-received: the boy approached the man, who was sitting down, and the man stood up and moved closer to the boy to instruct him on how to get to his destination. The man cupped his ears to listen to the boy's question and maintained direct eye contact with him. The man only broke eye contact when looking around to give the WB directions. The test subject also leaned his body closer to the WB to engage with him for the brief time that they conversed. Similarly, the WB had a positive experience when interacting with a group of White women. One of the women wrinkled her nose and cupped her ears to better hear what the WB was asking her. As the boy asked for directions, she reciprocated with a brief nod and used her hands, even though they were full, to point towards the location that the WB had asked about. As the WB was talking, the listener also changed her body position, so her stance more openly faced the boy.

Likewise, the White girl (WG) had positive experiences when asking directions from a White man and a White woman; similar to the interactions the WB had, the White woman used her body language to engage with the WG. The woman faced her body towards the teenage girl and pointed her in the direction of the store, maintaining relative eye contact with the teenager. The White woman also maintained little distance between the WG and herself, demonstrating her willingness to engage in a conversation with the WG. When the White man interacted with the WG, he used the same kinesic, proxemic, and emotional markers as the White woman.

The interactions that the Black girl (BG) and Black boy (BB) had with the shoppers were not as pleasant in comparison to the interactions of their White peers. The experience that the Black teenage girl (BG) had with a White woman showed hostility and an unwillingness to engage with the girl. When the BG approached the White woman, the woman sped up to walk more quickly than the speed the BG had used to approach her. Even when the BG was able to catch up to the White woman, the woman never stopped walking. Throughout the conversation, the woman took small steps away from the girl and never fully committed to the conversation. The woman never smiled, even though the BG attempted to smile at the woman to diffuse the tension between them. When the woman was asked where a particular store was, she initially dismissively replied "I do not know." When asked again, she gave a vague but correct location of the store, thus implying the woman lied the first time she was asked about the location of the store. The BG did, however, have a positive experience with a White man. When the BG approached the White man to ask him a question, he leaned his body inward toward the BG to listen to her question and continued to maintain eye contact with the girl. He used his arm to represent a street and his hands to demonstrate the route the girl would take to get to the store.

The Black teenage boy (BB) also had negative encounters with the shoppers with whom he spoke. When the BB approached a White woman to ask for directions, she looked down to avoid eye contact and started to walk away quickly. The woman moved a shopping bag on her right shoulder, which was closer to the BB, to her left, and held her purse closer to her body. When the BB asked for directions from a White man who was sitting down, unlike the experience the WB had when the man walked closer to him to engage in conversation, this man continued to maintain a large distance between the BB and himself. The man did, however, use his arm to indicate the direction of the store.

Field Observations 2

In the second field study, Black and White teenagers were tasked with walking into Store X to complete several different tasks. The objective of the field study was to try to determine a covert culture within Store X that resulted in inequitable treatment of shoppers. The boys were tasked with asking the cashiers in different departments for a shirt for vacation, tennis shoes, and cologne. Similarly, the girls asked different cashiers for undergarments, jeans, and tennis shoes. Boys and girls of both races also asked for ties for their fathers for Father's Day. In this field study, there was a more equal treatment of shoppers regardless of their race. In retail, salespeople are often taught to treat every shopper as a potential money spender, and this was shown in the continual asking of the teenagers if they needed help or assistance. However, not all shoppers were treated as kindly when given the same services. When the BG interacted with a non-Black sales clerk, the woman was rude to her. Her tone expressed irritation by the pauses in her speech and the low tone of her voice. Her eyes were slightly narrowed with a questioning glare. The saleswoman stood several feet away from the prospective shopper, and was not thorough when helping the client with her questions. Another BG as well as a WG did not receive any help from any of the non-Black salespeople, and were never checked on. However, one WG had a positive experience with the White saleswoman she encountered. The woman smiled, maintained eye-contact, and kept a regular conversational distance. In Store X there was an inconsistency between how girls were helped in comparison to their male counterparts.

All of the boys reported having positive experiences in Store X. The non-Black and Black salesmen that they interacted with all attempted to make the boys feel comfortable in the store. They used vocabulary such as "bro" and "dude" to cater to the clients with whom they were interacting. The non-Black and Black salesmen working in the store would nod in response to comments made by the teenage boys, and would maintain a conversational distance with the boys.

Analysis

Spatial Distance

When comparing the different nonverbal messages that were sent to the Black and White teenagers by the shoppers in the Walnut Creek shopping center, the only controllable change in the field study was the race of the teenagers asking the shoppers for directions. That factor, however, was enough to drastically alter the interactions the teenagers had with the White shoppers. In closely examining the different space dynamics, it becomes apparent that different messages were sent to the boys. Distance is one way a person expresses their level of interest in interaction as well as comfort level in a situation: if a person chooses to maintain a greater distance, a signal is sent to the person with whom they are communicating that the conversation is not comfortable. This signal fosters a lack of engagement by the person who does not choose to come closer to the person they are talking with. In the conversations that the WB had with the White shoppers, the distance that was maintained was about one and a half to two feet, which Hall, anthropologist, researcher, and author of *The Hidden Dimension*, sights as being labeled the Personal Distance, which "might be thought of as a small protective sphere or bubble that an organism maintains between itself and others."³ This signifies that this distance allows for the person to control who is in their close proximity and who is not, and for them to symbolically create a miniature world that only allows who and what they desire in their "world" or their "bubble"—keeping the bad things out of their "bubble." Entrance into this person's bubble is something that is granted, and we can see that in the distance that was maintained between

the BB and the White shoppers—what Hall labels the “social distance.” Hall cites the social distance between people as usually four to seven feet. Unlike the WB, the BB had several feet separating him from the people with whom he interacted because he was not granted “access” to the small world that the personal distance fosters. When the BB encountered a White man, the man chose to maintain a long distance that was somewhat awkward for conversation, in comparison to another White man who chose to step closer to the WB. Why would a White shopper choose to maintain such a distance? As mentioned before, the personal distance allows for the shoppers to invite someone into their bubble, which, in this field study, is a segregated one. The bubble allows WBs and WGs in, but keeps the BBs and BGs out. Why keep the Black teenagers out? The BBs interaction with the White woman provides insight into the culture that has been created after years of division and segregation in the United States. Not only did the White woman ignore the BB’s question, but she started to increase her speed when walking by him to create greater distance between herself and him. She moved her metaphorical “world” so that he was nowhere near it. She then demonstrated that she felt threatened by the boy because she grabbed her purse tighter and moved her shopping bag.

A common implicit bias surrounding Black men is that they are often considered dangerous and animalistic, especially towards White women. Young Black men are often labeled as “thuggish,” dangerous, and violent, a narrative that has been prevalent for several years. An article in the *Journal of Black Studies* by Wayne M. Blake and Carol A. Darling argues that “Blacks are frequently labeled as immoral, lazy, violent, and mentally deficient, along with being sexual super studs, athletes, and rapacious criminals (Hare & Hare, 1984). This is how he is perceived in the public consciousness.”⁴ This can be seen from the depiction of Black men in *Birth of a Nation* (1915), to the way Black men were portrayed during the war on drugs, to the many unarmed Black men dying at the hands of police officers who are not indicted for the unlawful killings. Black men are seen as predators to be afraid of and to avoid. “The ‘Black male predator,’ though, is not a stereotype that exists in a vacuum; he is part of a broader constellation of racist stereotypes about Black people and Black culture in which the family is surely intertwined.”⁵ This cultural norm and bias provides an explanation for why the BB was not invited into the inside of the White shopper’s personal distance: he was seen as a predatory criminal to be avoided. This further explains why the White woman held her possessions closer to her and why the White man maintained a relatively large distance between himself and the BB.

The BGs involved in the field study, like their Black male peers, also had a difficult time when interacting with the White shoppers. The situations that the White and BGs experienced almost mirrored the way the White and BB were treated. The BGs were treated the most hostilely by the White woman they encountered; though the White woman that interacted with the BGs did not clutch her bags more closely, she did continue walking while talking to the BG and did not maintain eye contact. One reason that she may have acted negatively and dismissively towards the BG is that Black women are constantly labeled as hostile: Patricia Hill Collins, who wrote an article on the sociology of discrimination entitled *Social Problems*, states “King suggests that stereotypes represent externally-defined, controlling images of Afro-American womanhood that have been central to the dehumanization of Black women and the exploitation of Black women's labor. Gilkes points out that Black women's assertiveness in resisting the multifaceted oppression they experience has been a consistent threat to the status quo. As punishment, Black women have been assaulted with a variety of externally-defined negative images designed to control assertive Black female behavior.”⁶ Since Black girls are often perceived as hostile and aggressive, this provides a plausible explanation for why the White woman wanted to keep

her distance and would not invite the BG into her personal zone, in comparison to the interaction with the WG.

Spatial Distance Conclusions

Implicit biases pertaining to race perpetuate unequal treatment. In this field study, different space zones depended heavily on the teenagers' race, and the field study honed in on the messages those space zones sent. We were able to examine how race allowed the teenagers to be invited into a metaphorical "world" that the personal space zone fosters. If the teenagers were Black, this zone did not extend to them. When examining reasons why Blacks were not able to be a part of the White shoppers' personal zones, we referred to stereotypes such as the perception of Black men as violent criminals as well as the belief that BGs are hostile.

If we continue to let only certain people into a bubble that we have created for ourselves, we will never learn to listen to one another and accept each other. The only way to defy the implicit biases that we *all* have is to interact with one another and be willing to let people into space zones that are a little more intimate in order to foster trust and conversation. Continued discrimination will hinder open communication and perpetuate a cycle of friction and divide.

Paralanguage

Paralanguage or lack thereof can serve as a mode of communicating the emotions behind the actions people take. Professor Pennycook of McGill University cites the Education Resources Information Center when defining paralanguage: "The 1984 ERIC definition of paralanguage is the 'study of those aspects of speech communication that do not pertain to linguistic structure or content, for example, vocal qualifiers, intonation, and body language' "⁷. This can be seen by the woman who interacted with the BB choosing not to say anything or acknowledge him. This showed that her actions, especially moving her items away from the BB, were instinctive rather than planned. In contrast, the tone that the man used when interacting with the WB was a tone that expressed conveyed interest. His tone exhibited a level of engagement in the boy's question that allowed for the WB to feel more relaxed and comfortable in the conversation. The interest and friendliness of the tone fostered a connection that quickly allowed the WB to become a part of the man's space bubble, which prompted the man to stand up and stand closer to the boy to give him directions. Knapp and Hall argue that when people are communicating with one another, "Young children rely much more on verbal content, older children show a mixed pattern, and adults rely much more on nonverbal tonal qualities."⁸ This indicates how much impact the tone of the man's voice had on the WB, and how tone served as an alternate means of including WBs while rejecting BBs. However, sometimes there are exceptions applicable to both races. In Store X, the cashier used tone and language to relate to the teenage boys to make them feel more comfortable. The employee used a more relaxed tone to fit in and accommodate and cater to the client. This ultimately led to the boys feeling accepted in the store, which increased their chance of wanting to buy something there.

This phenomenon can also be seen in the way that the non-Black saleswoman in Store X interacted with the BG. Though many factors could have led to the BG being treated poorly by the saleswoman in comparison to the treatment of the WG, the BG was treated as if she were a burden to the salesclerk. The sales clerk provided the girl with short, curt answers in an irritated tone that showed her unwillingness to engage in conversation and interaction with the client. This greatly differed from the experience the WG had of having light and easy conversations with the cashier while they looked for items. The ability to

understand the hostility towards Black teenage girls was also evident in the way a White woman interacted with one of the BGs on the street when giving her directions. The woman used the same tone and speech pattern as the woman in the store. This, accompanied with the failure of either woman to smile, led the BG to understand this was not a safe space. Tone of voice has the ability to define what kind of space people are in. The tone sets a precedent for what is deemed okay, and the tone used by both women led the BG to understand that that space was not a positive one.

Paralanguage Conclusion

In summary, paralanguage is an important emotional marker because it indicates the motives for people's thoughts and actions. It allows for people to understand the spaces they are in. The tone of several White people showed that Black people were not always welcome in their spaces. They were seen as a nuisance, bothering the people they interacted with by asking them questions. Though people do have different ways of responding in terms of their temperament and the way they express themselves, we can see several examples of Black people experiencing a harsher tone than their White peers. We can, however, see exceptions to this rule. For example, almost all the boys in Store X received equal treatment in terms of the tone and language that the salesclerk used, which we can most likely attribute to the culture and requirement of a salesperson to treat all customers as money-spenders.

Eye Contact and Facial Expression

Facial gestures, most importantly eye contact, in both field studies one and two, demonstrate people's willingness to interact and engage with the people they are talking to. In the field studies, eye contact and interaction supported that Black teenagers were more often discriminated against, because they were less likely to receive facial expressions that conveyed interest in interaction. Different races have different beliefs pertaining to appropriate levels of eye contact, however. Professor Ruth Davidhizar, who wrote a paper analyzing interpersonal communication through eye contact, categorizes appropriate levels of eye contact for different races. She argues that "Most Caucasians, however, value eye contact as indicative of positive self-concept, interest, and openness. Avoidance of eye contact is considered by some to be rude, to indicate lack of attention, or to show mental illness. African-Americans and Mexican-Americans appear to desire and engage in eye contact even more frequently than some Caucasians."⁹ This idea of who values eye contact is critical when examining who *received* eye contact when participating in the study.

In the first study, there was an extreme inequality in the maintaining of eye contact and facial expression between different parties. When interacting with a BG, a White woman refused to maintain eye contact, suggesting an unwillingness to interact with and acknowledge the BG and, by extension, an unwillingness to acknowledge the BG's presence. The BG was somewhat "dismissed" by the woman because the woman did not use eye contact throughout the conversation; in order to maintain an equal relationship throughout the conversation, there needed to be an equal exchange of eye contact. In addition, the woman did not smile during the whole time the pair interacted despite the BG's attempts to smile at the woman. This is similar to the experience that the other BG had in Store X. In Store X, the woman's lack of facial expression created a stern expression that manifested in the cold and curt interactions the two had. Similar to the first study, the lack of eye contact exhibited the lack of interest the woman had in helping her client. This, paired with the woman not paying full attention to the client and busying herself as the BG was trying to speak to her, fostered the feeling of non-acceptance or lack of value to the store.

Professors from the Graduate School of Education at Columbia University describe this phenomenon as a type of micro-insult. According to them, “A micro-insult is characterized by communications that convey rudeness and insensitivity and demean a person’s racial heritage or identity.”¹⁰ They further argue that “Micro-insults can also occur nonverbally, as when a White teacher fails to acknowledge students of color in the classroom or when a White supervisor seems distracted during a conversation with a Black employee by avoiding eye contact or turning away (Hinton, 2004). In this case, the message conveyed to persons of color is that their contributions are unimportant.”¹¹ This demonstrates how such small nonverbal actions can have such large implications in terms of creating safe spaces, spaces where people feel acknowledged and respected. What is interesting to note is that in society, Whites, Blacks, and Hispanics all deem eye contact as something that is necessary and respectful, but not everyone was shown the same relative amount of eye contact. Therefore, this sends the message that the spaces where eye contact was maintained (the WG’s interaction) was deemed a space where she was going to be respected, because they were treating her as they would anyone they thought worthy of being treated by the status quo.

This contrasted to the WG who maintained an average level of eye contact with the person she talked to and engaged with. The level of engagement in that pair ultimately fostered a more equitable relationship. This was exhibited by the White woman using her hands to interact more with the girl, and using her hands to point in the direction of the store. Likewise, when one of the WGs went into Store X, the woman maintained an appropriate and professional amount of eye contact with the girl. This made the girl feel more comfortable in the space, because she would then smile at the saleswoman and the saleswoman would reciprocate.

The Boys’ results were slightly varied, but the same principle applied. Similar to paralinguage, the lack of eye contact that the BB encountered when asking the White woman for directions demonstrates how this can lead to feelings of “unimportance” or invisibleness, as mentioned in earlier quotes. The lack of eye contact not only ties directly into making young Black men feel unimportant and invaluable, but also ties directly into the perception of young Black men as thugs, as mentioned earlier. This perception also probably led to the woman ignoring the boy when he spoke to her. The woman allowed her implicit biases to influence her actions and to treat the young boy as someone with whom she did not want to associate or engage. However, the BB did have positive experiences with the White woman he encountered in Store X. When asked questions, she continued to answer him by maintaining eye contact and nodding in response to the boy’s comments. This exemplifies her ability and willingness to engage with the boy and help him find the products that he needed. In this specific situation, the BBs and WBs were treated with relatively equal demeanors. This demonstrates that there are always exceptions to the rules.

Eye Contact and Facial Expression Conclusion

Eye contact and facial gestures helped to convey the way that the adults viewed and valued the teenagers with whom they were interacting. Their smiles, use of eye contact, and nods of their head helped to exemplify engagement in the conversation. What often occurred was that Black teenagers did not encounter the positive kinesic markers and, instead, experienced being ignored and shown stern faces and squinting, scrutinizing eyes. Though the divide is not as always clearly obvious, the different forms of treatment also convey different messages. Unlike the White teenagers, who received the message that people were ready to engage, Black teenagers received messages that made them feel invisible, unequal, and undervalued.

Conclusion

The purpose of this paper is to shed light on a pressing issue in today's society. It is easy to deny that racism or sexism, or any system of oppression, exists in today's society. One might want to believe that American society is making great strides towards fairness and equality, that prejudice, stereotyping, and discrimination are disappearing. But it is my hope that my small study shows an example of how this issue continues to bedevil our society. In both studies, it was apparent that Black teenagers were not treated the same as their White peers. The stereotypical labels such as Black men being criminals or Black women being aggressive appeared to be a leading reason why they were not let into the "bubble" that White people create around themselves where people feel safe. The results of this study are not to say that one race is good or bad, because there are good and bad people of all races. However, this study serves as some evidence that in today's society, Black people are not treated equally.

As a result of writing this paper, I ask that instead of building walls, and creating division within the community, we listen to one another. We must engage with one another, and try to understand each other. The narrative that has been painted of Black people in society is a negative one, so, I encourage you to talk, to learn, and to listen. Then maybe next time, a Black young woman or man might have a positive experience in which they feel valued and safe.

Endnotes

¹Hall, Judith and Knapp, Mark *Nonverbal Communication in Human Interaction*, (Boston, MA: Monica Eckman) Page 4.

²Altman, Andrew "Discrimination" *Stanford Encyclopedia of Philosophies*, February 1, 2011 revised August 30, 2015, <https://plato.stanford.edu/entries/discrimination/#ConDis>

³Hall, Edward *The Hidden Dimension*, (United States: Anchor Books, 1966) Pg. 119

⁴Blake, Wayne and Darling, Carol "The Dilemmas of the African American Male" *Sage Publications, Inc.*, June 1994, Pg. 402 <https://www.jstor.org/stable/pdf/2784561.pdf?refreqid=search:2e489cf0369c86780dbeab9a601e859c>

⁵Lane, Alycee "'Hang Them if They Have to be Hung': Mitigation Discourse, Black Families, and Racial Stereotypes," *University of California Press*, Pg.190 <https://www.jstor.org/stable/pdf/10.1525/nclr.2009.12.2.171.pdf?refreqid=search:2808005d2e897060f8b3e757f24ff4be>

⁶Collins, Patricia *Learning from the Outsider Within the Sociological Significance of Black Feminist Thought*, (United States, University of California Press)

⁷Pennycook, Alastair "Actions Speak Louder Than Words: Paralanguage, Communication, and Education." McGill University, TESOL QUARTERLY, 1985, Pg.26, https://www.researchgate.net/profile/Alastair_Pennycook/publication/264373117_Actions_Speak_Louder_Than_Words_Paralanguage_Communication_and_Education/links/56048e6308aeb5718ff00719/Actions-S

⁸Hall, Judith and Knapp, Mark *Nonverbal Communication in Human Interaction*, (Boston, Ma: Monica Eckman) Pg. 325

⁹Davidhizar, Ruth, "Interpersonal Communication: A Review of Eye Contact," Cambridge University Press on behalf of The Society for Healthcare Epidemiology of America, <https://www.jstor.org/stable/pdf/30147101.pdf?refreqid=search%3Aea445b028bf537e92ab6e040e1478bff,03-08-2017>

¹⁰Sue, Derald, Capodilupo, Christina, Torino, Gina, Bucceri, Jennifer, Holder, Aisha, Nadal, Kevin, Esquilin, Marta, "Racial Microaggressions in Everyday Life, Implications for Clinical Practice", *Teachers College, Columbia University*, May-June 2007, Pg.274 <http://world-trust.org/wp-content/uploads/2011/05/7-Racial-Microaggressions-in-Everyday-Life.pdf>

¹¹ Sue, Derald, Capodilupo, Christina, Torino, Gina, Bucceri, Jennifer, Holder, Aisha, Nadal, Kevin, Esquilin, Marta, "Racial Microaggressions in Everyday Life, Implications for Clinical Practice", *Teachers College, Columbia University*, May-June 2007, Pg.274 <http://world-trust.org/wp-content/uploads/2011/05/7-Racial-Microaggressions-in-Everyday-Life.pdf>

Bibliography

1. Altman, Andrew. "Discrimination." *Stanford Encyclopedia of Philosophies*, plato.stanford.edu/entries/discrimination/#ConDis.
2. Blake, Wayne M., and Carol A. Darling. "The Dilemmas of the African American Male." *Journal of Black Studies* 24.4 (1994): 402-15. Web.
3. Collins, Patricia H. "Learning from the Outsider Within: The Sociological Significance of Black Feminist Thought." *Social Problems*, vol. 33, no. 6, 1986, pp. S14-S32.
4. Davidhizar, Ruth. "Interpersonal Communication: A Review of Eye Contact." *Infection Control and Hospital Epidemiology*, vol. 13, no. 4, 1992, pp. 222-225.
5. Hall, Edward T. *The Hidden Dimension*. Peter Smith Pub, 1992.
6. Knapp, Mark L., et al. *Nonverbal Communication in Human Interaction*. Wadsworth Cengage Learning, 2014.
7. Lane, Alycee. "'Hang Them if They Have to be Hung': Mitigation Discourse, Black Families, and Racial Stereotypes." *New Criminal Law Review*, vol. 12, no. 2, 2009, pp. 171-204.
8. Pennycook, Alastair. "Actions Speak Louder Than Words: Paralanguage, Communication, and Education." *TESOL Quarterly*, vol. 19, no. 2, 1985, p. 259.
9. Sue, Derald W., et al. "Racial microaggressions in everyday life: Implications for clinical practice." *American Psychologist*, vol. 62, no. 4, 2007, pp. 271-286.



Ethical Considerations of State Responsibility toward Refugees: Analyzing China's Refugee Capacity from a Socio-Economic Perspective

Xiangyu Zheng

Author Background: Xiangyu Zheng grew up in China and currently attends the International Department of the Affiliated High School of South China Normal University in Guangzhou, China. His Pioneer seminar topic was in the field of international relations and titled "Globalization and International Migration."

Abstract

Syrian refugee crisis has posed the timely question of what responsibility states, especially those geographically distant and culturally distinctive, have towards refugees. Specially, this paper argues for China's moral obligation of accepting Syrian refugees from a cosmopolitan perspective, i.e. the universal human rights of refugees should be protected by any states. Other perspectives such as utilitarianism and communitarianism are also considered. I either refute them or find them irrelevant based on China's economic capacity and cultural foundations. China's socio-economic situations are taken into consideration when giving recommendations on China's future relations with Syrian refugees.

1. Introduction

Since the start of the Syrian conflict in 2011, more than five million Syrian refugees have fled the country and around 6.5 million have been internally displaced (UNHCR). The responsibility to provide sanctuary for these refugees has fallen unequally on a few countries: on the one hand, tiny Lebanon, with a pre-conflict population of four million people hosts one million refugees, while Germany, a country with a population of 81 million, has accepted more than one million refugees from the Middle East (mainly Syria). The wealthy Gulf countries on the other hand have accepted only a handful of refugees. Likewise, China, with a population of 1.3 billion, has accepted fewer than 30 Syrian refugees (BBC; Borgen Project). Such a contrast raises questions regarding China's moral obligations in the face of this humanitarian catastrophe.

In this paper, I will seek to answer the following three questions: Does China have an ethical obligation to accept refugees? Do China's social and economic conditions enable the country to take in Syrian refugees? How can China fulfill its ethical responsibilities towards Syrian refugees in specific terms?

I will use Joseph Carens's theory of cosmopolitanism to support my argument that China is ethically obliged to accept refugees fleeing the conflict zone. Carens embraces a cosmopolitan right, which he develops based on Kant's Third Definitive Article on Perpetual Peace which itself holds that the cosmopolitan right shall be limited to universal hospitality. In particular, Carens argues that moving freely across the border is an inherent human right,

just like individuals have the equal freedom to move within a state (Gibney 171). As Carens himself argues, however, political realities dictate that a more pragmatic approach regarding accepting refugees should be taken; in this instance, public order and the maintenance of liberal institutions determine whether an open border is possible (251-264).

Since these more practical concerns generally have socio-economic roots, in this paper I will examine whether China's economic and social capacities allow it to accept Syrian refugees. I will focus in particular on China's economic power and cultural foundations, and will argue that China's current economic prosperity, compared to the minor needs of individual refugees, should allow China to take in a large number of Syrian refugees—I estimate around two million (Gibney 171). The low cost of taking in refugees also meets the communitarian standard advanced by Michael Walzer, who articulates that the only case in which the state's right to exclude refugees does not apply is when refugees are in dire need and the cost to accept them is relatively low (Carens 32). Yan Xuetong, one of China's leading scholars of international relations, points out the cultural context of this proposition, saying that Confucianism claims that any state has a universal moral responsibility—"the sphere of concern for any humane ruler should be the whole world, not just the people of one state" (Yan 1). Therefore, one can argue that China also has moral obligations to these foreign refugees based on Confucian principles. Additionally, China has a long history of cultural interactions with domestic Muslim communities – the Uighurs and the Hui. According to Seyla Benhabib, such cultural interactions tend to promote social progress (Young 28). Therefore, China's obligation to accept refugees stands on solid ground culturally, and in light of its extensive experience in cultural exchange with domestic and international Muslim communities, the country will likely benefit from the culture intake.

Iris Marion Young's argument about the pervasiveness of structural injustice and Martha Nussbaum's famous "capabilities approach" will be used to develop an argument for what China should do, specifically, to integrate the refugees socially, economically, and politically. Young differentiates between a positional difference and a cultural difference (Young 4). Facing structural injustice as newly arrived aliens, refugees in host countries might suffer again from economic challenges, such as poor employment. China should therefore proactively give refugees access to the labor market, providing them jobs that match their professional qualifications. An absence of a Muslim community or Muslim places of worship may lead to the marginalization of the refugee community, but in light of China's current successful protection policies for its Muslim ethnic minorities in the Xinjiang and Ningxia Provinces, applying these policies to Syrian refugees could also help mitigate structural injustice caused by cultural differences. Nussbaum's capabilities approach argues for the advancement of individual capabilities, calling for measures such as equal access to education and psychological treatments (Gasper and Truong 342). China, with its vast economic resources, is more than equipped to provide a suitable environment for refugees as well as employment and access to education.

The research is based on primary and secondary sources, including academic works on refugee ethics and state rights, China's official economic reports, NGO reports on Syrian refugees, and China's official documents relating to its current refugee policies and economic capacity.

It is imperative that China start to accept refugees from Syria. Not only would China's decision to accept refugees alleviate the suffering of Syrians fleeing danger at home, but it would significantly reduce pressure on the global community. Additionally, as a country that aspires to step onto the international stage as a major power, for China to take

in Syrian refugees would be a good chance to assert its moral soundness and humanitarian commitment.

2. The Syrian Refugee Crisis—A Snapshot

The violently suppressed anti-Assad movement lies at the roots of the Syrian conflict. Since the Syrian conflict erupted in 2011 as part of the so-called Arab Spring, millions of Syrians have fled the brutal civil war to European countries such as Germany and Sweden, and neighboring non-Gulf Middle Eastern countries, such as Jordan and Lebanon (See Figure 1 & 2). Countries outside the region, including the wealthy Gulf countries and the U.S., have made few efforts to relieve the burden by accepting refugees.

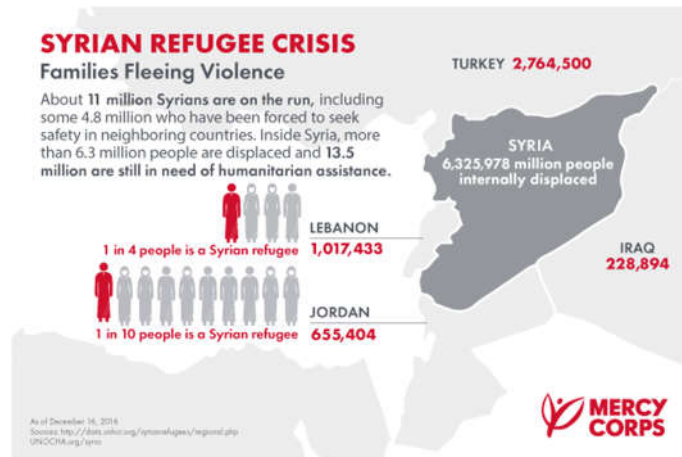


Figure 1. <https://www.mercycorps.org/articles/iraq-jordan-lebanon-syria-turkey/quick-facts-what-you-need-know-about-syria-crisis>

The figure shows that currently non-Gulf Middle East countries take in the largest portion of Syrian refugees. In particular, Lebanon, a small and economically weak country with a population of less than 6 million, has taken in over one million Syrians, almost a quarter of its own population. Adding that to its relatively weak economy (when compared to the wealthy Gulf countries), the country is clearly carrying more than its share of the burden.

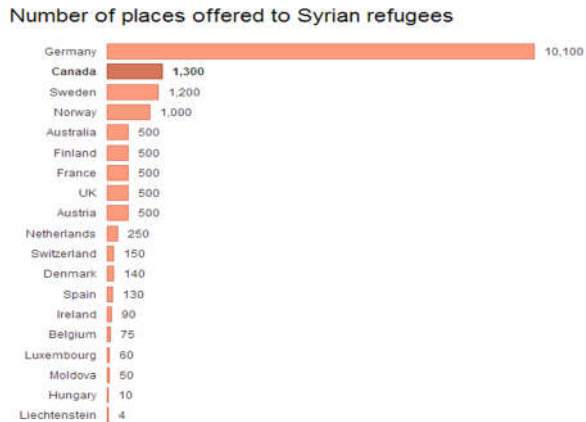


Figure 2. <https://www.theguardian.com/news/datablog/2014/jan/29/where-are-the-syrian-refugees-going>

This figure shows that in Europe, Germany is hosting the largest number of refugees (the number in 2017 is now over one million). For a country with a population of 81 million, taking in over one million refugees is clearly an unequally large burden, especially when compared to other European countries. Out of pure economic consideration, it seems that the wealthy Gulf countries should also bear refugees fleeing Syria. The reasons why they are not are borne of purely political consideration and are outside the scope of this paper.

The Syrian conflict thus raises all the concerns about a gigantic humanitarian crisis that cannot be locally contained – vast destruction of infrastructure, massive displacement and massive movements of people. Additionally, according to a 2014 UN report, both sides of the conflict–government forces and non-state armed groups–have committed serious human rights violations against civilians. Their atrocities include murder, rape, torture, sexual violence, hostage-taking, indiscriminate aerial bombardment, the use of chemical weapons, execution without due process, massacres, and forcible displacement (UN Human Rights Council 1). Since 2015, when outside forces began to launch airstrikes, the death toll of Syrian civilians has rose even more (Mercy Corps). All of this makes the return of refugees even less likely. As a result of the ever-increasing violence of the war, Syrians refugees have fled their home nation to escape such humanitarian violations, yet for many refugees the situation en route or in the camps, even the UN camps, is no better. On the way to receiving countries, Syrian refugees face physical and emotional threats such as sexual violence, human trafficking, rising child marriage rates, serious threats to health, and lack of basic needs being met. For example, it was reported that a human trafficker deliberately sank a boat full of Syrian refugees in the Mediterranean, and children were easily manipulated by these human traffickers and even sold as sex workers (Mercy Corps; The Guardian). Despite best efforts made by the international community and humanitarian organizations, for refugees staying in camps, living conditions still tend to be harsh. In Za'atari, Jordan, the largest refugee camp in the world and the temporary home of 80,000 Syrians, supported by the Jordan government and UN, although basic life is maintained, the living conditions and opportunities for human development are meager. Many Syrians hold illegal jobs and live on scanty humanitarian aids (Oxfam). Children have little access to local education–in Lebanon, over half of the 500,000-registered school-age Syrians are out of school, though in some camps, camp schools have been set up (Human Rights Watch).

An often-ignored issue is the abuse of women's rights among displaced Syrian refugees. According to UNHCR, 48.5 % of the 5,106,934 registered Syrian refugees are female (UNHCR). The rights of such a large number of Syrian refugees are not only often ignored but are even abused. According to the International Rescue Committee, girls and women in Syrian refugee communities suffer from sexual abuse, domestic violence, and early marriage-worse conditions than those they face in their home communities back in Syria (International Rescue Committee). Also, sanitary facilities for women are inadequate. In some camps, women have no separate toilets, bathrooms, or sleeping areas, increasing the chance of being sexually abused by men (Amnesty International).

Keeping refugees in camps or on roads is not the solution to these humanitarian issues; rather, more countries need to be willing to accept these refugees into their national communities and offer them normal—if not better—lives.

3. Ethical Theories About the Responsibilities Towards Refugees: An Overview

Due to the unequal burdens that fall upon certain countries and the human suffering involved, the Syrian refugee crisis has given rise to a hot debate over countries' moral and legal obligations to Syrian refugees. Over the centuries, and in particular in the decades since World War II and the Cold War with its massive streams of displaced human beings, for example from Eastern Europe, Cuba, or South East Asia, philosophers have developed sophisticated theories on the moral obligations individuals and states owe to refugees. Several of these theories, in particular those with roots in cosmopolitanism, communitarianism, and utilitarianism, will be considered when outlining possible ways to deal with the Syrian refugees. Below I will advance a moral argument that is based on cosmopolitanism.

3.1. The UN Legal Framework on Refugees

The UN 1951 Convention Relating to the Status of Refugees and the later revised 1967 Protocol Relating to the Status of Refugees are currently the most important legal frameworks guiding the international community and countries on refugee protection. The Convention defines a refugee as "someone who is unable or unwilling to return to their country of origin owing to a well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group, or political opinion"(United Nations). In this regard, Syrians who have fled their homeland are considered refugees as they may be persecuted because of their religious beliefs and political affiliations at home.

The long-held principles in the Convention and the Protocol include non-discrimination, non-penalization, and non-refoulement. In terms of non-discrimination, countries should not reject refugees based on their race, religion, or country of origin. The more important principles are non-penalization and non-refoulement, which together state that refugees cannot be charged with the crime of illegal immigration, and that once they cross the border of a country, the country cannot return the refugees to their origins, where their safety or lives might be threatened (United Nations). The UN Refugee Convention therefore provides important guidelines for treatments of Syrian refugees, i.e. they should not be penalized and discriminated based on their identity, and should not be returned involuntarily once they enter the border. As the Refugee Convention is in part rooted in cosmopolitanism, I will further illustrate the ideas of cosmopolitanism.

3.2. Kant's Cosmopolitanism

Immanuel Kant claimed that liberty, equality, and property are inherently-held rights. He argues that each human being has an inherent worth, possessing dignity inherent to human nature (Bayefsky 811). Kant's ideas on human rights are the theoretical foundation of the bulk of the UN's human rights framework: the Universal Declaration of Human Rights proclaims that "recognition of the inherent dignity and of equal and inalienable rights of all members of the human family is the foundation of freedom, justice, and peace in the world" (qtd. in 810). Therefore, an important implication of the inherent human rights recognized by the international community is that they should not be transcended by any other forces. The protection of human rights should be prioritized in dealing with the Syrian refugee crisis.

Kant also develops what he calls a "cosmopolitan right" or *jus cosmopolitanicum*, which was further developed by contemporary philosopher Joseph Carens and upon which my main argument is based. Kant developed his concept of cosmopolitan rights from the understanding that all human beings belong to a single human community, and thus individuals and individual states have a moral duty to protect others from harm (war): "originally no one has more of a right to be at a given place on earth than anyone else" (qtd. in Kleigeld). Therefore, every individual can claim universal rights or "hospitality" into any state: "the right of a stranger [should] not be treated in a hostile manner by another upon his arrival on the other's territory" (qtd. in Doyle). This right, however, is not unrestricted, according to Kant. In his Third Definitive Article, he states that the cosmopolitan right "shall be limited to the conditions of universal hospitality." His version of cosmopolitan rights stipulates that a stranger should not be treated with aggression, but that the right to exclude still remains in the hands of the state. The only exception is when expelling the stranger could lead to the person's death (Kleigeld). In this regard, the UN definition of a refugee expands upon Kant's (Kant's cosmopolitan right may reasonably be interpreted to imply the scenario of refugees) to include more areas such as political and religious persecution, which do not necessarily lead to death but certainly to human suffering.

3.3. Cosmopolitanism After Kant

3.3.1. Joseph Carens' Cosmopolitan Argument

Carens further builds on Kant's concept of cosmopolitan rights but, unlike Kant, Carens' cosmopolitanism transcends the state's right to exclude. To formulate his own argument for an open border, he mainly draws inspiration from Robert Nozick's theory that state rights should not transcend individual rights and from John Rawls' concept of freedom to move within borders (Carens 252). According to Carens, Nozick argues that "the state has no right to do anything other than enforce the rights which individuals already enjoy in the state of nature" (253). Under Nozick's definition, the range of individuals includes both citizens and aliens. Therefore, once within the border of a state, the aliens' rights should also be enshrined by the government. Based on this reasoning, since individuals possess the inherent right of freedom as proposed by Kant, countries should not limit the entrance of refugees into their territory. Carens also draws heavily from John Rawls' Original Position to derive his argument for an open border. He argues that according to Rawls' Original Position, individuals will not relinquish their inherent rights of equal liberties to move, which also include the liberties of migration. Extending the Original Position from a national level to the global community therefore supports opening the borders to foreign refugees (Gibney 171).

Nevertheless, Carens also contemplates a non-ideal perspective, which takes "historical obstacles and the unjust actions of others" into consideration (Carens 255). In this regard, a clearly delineated ethnic culture or historically fixed biases against a people may create conditions that hinder free migration. For example, for two countries currently at war or historically often at war, it might be impossible for one to accept refugees from the other. Nevertheless, Carens contends that equal liberties should still be prioritized in the long run (261). Besides the obstacles forwarded by a pragmatic approach, Carens also voices several exceptions to his open border argument, including threats to public order and to the maintenance of liberal institutions (Gibney 171). According to John Rawls, Carens says that the liberties of immigration should be restricted in the case of damaging public order to reinstall the more fundamental liberties provided by a sound public order. However, such a "public-order argument" should not be overused, and "a need for some restrictions would not justify any level of restrictions" (269). The argument for maintenance of liberal institutions seems dubious though: if a nation does not even observe the liberal principle of a cosmopolitan right, to what extent can it preserve its liberal ideals?

3.3.2. Kwame Anthony Appiah's Cosmopolitanism

In his book *Cosmopolitanism: Ethics in a World of Strangers*, Appiah outlines the two conflicting strands of his cosmopolitanism—"one that stresses global obligations, [and] one that celebrates local differences" (Ikenberry 151). Although his argument is far more complex than these two points, they can nevertheless be distilled and viewed through a more practical lens than that of Carens. Under Appiah's cosmopolitanism, the global obligations echo Carens' call for open borders, but it adds on the more pragmatic concern of cultural distinctiveness that resonates with communitarians such as Michael Walzer's argument for some degree of cultural homogeneity. Drawing inspiration from Appiah, I find it necessary to include some communitarian elements into the overarching cosmopolitan framework to make my argument more applicable. Therefore, besides cosmopolitanism, I will also embrace other arguments, including utilitarianism, represented by Peter and Renata Singer, and communitarianism, mainly advanced by Michael Walzer.

3.4. Peter and Renata Singer's Utilitarian Approach

Singer and Singer take a different approach—utilitarianism—but come to a similar conclusion, advocating for open borders.

In response to the UN definition of refugees, which suggests that countries only have a non-refoulement responsibility to those who have reached their borders either through legal or illegal means, the Singers ask an intriguing question: Why should someone who is able to travel to another country have priority over others who are in refugee camps and unable to travel (Singer)? Alternatively, the Singers suggest a utilitarian "neediest approach," calling countries to turn away the relatively well-off asylum claimants who have reached their borders but take in those facing the harshest situations, who are often still in refugee camps (Martin 990). Such suggestions echo classical utilitarianism, which advocates the greatest marginal benefits for the greatest number of people.

The Singers also favor "the principle of equal consideration of interest," a utilitarian method of weighing the interests of aliens against those of citizens and prioritizing them accordingly (qtd. in Martin 990; Gibney 171). In this way, they argue for an acceptance of whoever, asylum claimants or refugees, is in the most dire need, up to the point when "the negative effects on current residents would outweigh the positive effects on the refugees" (Carens 36).

3.5. Refutation of Garrett Hardin's Lifeboat Ethics

Numerous countries in Europe and around the world are rejecting refugees based on the argument that "the boat is full," meaning that the country should reserve its fortune only for its own citizens. Drawing on an analogy of the sinking of an overcrowded lifeboat, Hardin warns "rich countries" to close their borders. Hardin's main argument lays on the welfare of a rich country's future generations-"Every life saved this year in a poor country diminishes the quality of life for subsequent generations" (qtd. in Callahan 3). This argument may be true, but it ignores the basic consensus that life comes first, and further, is in a sense racist because it claims citizens of rich countries deserve a better life. Therefore, I strongly reject Hardin's argument because it ignores the basic human rights. Furthermore, in fact, according to Foreign Policy, taking in refugees would be actually beneficial to China, for refugees complement a lack of labor force, fueling the country's further economic growth (Pan).

3.6. Michael Walzer's Communitarian Approach

The need to preserve cultural homogeneity is also heard in the arguments against taking in Syrian refugees. The argument of European countries such as Poland and Hungary - that the Islamic culture brought along with Syrians threatens their cultural homogeneity-echoes communitarians such as Michael Walzer, who put forth the argument that since national culture shapes important aspects of our daily life, states should have the right of self-determination to ward off foreign refugees who could potentially threaten the fundamental national culture (Walzer 49). They argue for the homogeneity of a society, as people "have the right to demand that others respect whatever is indispensable to... [their] being full human subjects" (Gibney 172). Walzer's view echoes a common variant of Westphalian-type state sovereignty, which respects self-determination and is seldom subject to challenge. When facing the inherent human right proposed by cosmopolitans, however, the legitimacy of such communitarian framework needs to be reconsidered.

Nevertheless, Walzer does make an exception-the mutual aid: when the cost of accepting refugees is low for the receiving state and the refugees are really in urgent need, the state should take in refugees. This means for the purpose of my argument that I will examine if China is both economically strong and culturally welcoming to refugees.

4. China's Obligations Towards Refugees

4.1. A Need for a More Nuanced Approach

What I term "pragmatic cosmopolitanism" incorporates aspects of cosmopolitanism, utilitarianism, and communitarianism. In addition to Kant and Carens' ideal cosmopolitan approach, the central framework of my argument that China has a moral responsibility to Syrian refugees, I will consider the more practical utilitarian and communitarian concerns-the utilitarian's socio-economic considerations, which include the number of refugees China's economics realistically allow, popular opinion towards refugees on a societal level, and the communitarian's contemplation of cultural homogeneity and cultural foundations, about which I will discuss the long history of China's flourishing Muslim communities and the Confucian foundation of a universal moral obligation.

The next portion of the paper entails detailed analysis of the theories mentioned in the previous section, illustrating why cosmopolitanism is the best overarching framework, to which the other, more practical approaches should be added.

In accord with the three principles of the UN Convention, despite that Syrian refugees come from an entirely different ethnic background and that their faith—Islam—is different from Confucianism, Buddhism, and the mainstream Han culture of China, China should not make a distinction between which refugees to accept—for example, more culturally similar Vietnamese versus Syrians.

Before moving on to the discussion of my "pragmatic cosmopolitanism," an important yet often neglected point deserves at least a brief discussion, i.e. that UN legal obligation to assist refugees is limited to those refugees who have reached the host country by crossing its borders. This, however, due to the long distance that prevents Syrian refugees from reaching China, clearly does not apply to the case of China's associated responsibilities to Syrian refugees. Therefore, to address the urgency of the current Syrian refugee crisis, we must find an additional set of moral frameworks to oblige other countries, specifically China, to actively open their borders.

Like Carens, I agree with cosmopolitanism, which focuses primarily on the inherent human right to move freely. I regard it as my fundamental framework because it is a relatively humanitarian approach, focusing on individual rights rather than the state's determination, which plays a vital role in both utilitarian and communitarian's approaches.

I find that Singer and Singer's utilitarian approach offers an easily acceptable and normative framework towards refugees, but it lacks both ethical soundness and pragmatic concern for a state's social environment. Instead, I agree with Carens that deporting asylum claimants might be more morally culpable than failing to admit refugees from distant camps, and that it is highly impractical for countries to attempt to identify refugees in the neediest state (Carens 39). The Singers also advance a marginal analysis in determining the number of refugees to be accepted, proposing a breaking point when "the negative effects on current residents would outweigh the positive effects on the refugees" (Carens 36). Such brinkmanship, however, is too strict, for before reaching this "breaking point," domestic turbulence caused by severe harms to original citizens would already have bailed out the policy of accepting a considerably large number of refugees. Having said that, a moderate number of refugees (compared to a country's population—in China's case, the number might be significantly larger) is generally more acceptable. A neediest-based open border that calculates marginal benefits is also impractical because it is impossible for a country to deny entry to those who have already reached the border and accept those still in distant refugee camps. In a nutshell, Carens' cosmopolitan approach tends to be more reasonable than the highly idealistic utilitarian approach.

As for Michael Walzer, at first glance, prioritizing the right to exclude based on national culture seems perfectly justifiable for China, whose mainstream Han culture is both distinct and completely different from that of Syrian Muslims. Upon close examination, however, I cannot but challenge Walzer. Inspired by Carens, I use the inherent right to move freely within borders to challenge Walzer's priority of a state's right to exclude: since the right to move is inherent in our nature, a person entitled to move freely inside a country should also be able to move freely across borders. Although China is not always considered a fully liberal state, freedom to move is constitutional. Therefore, the principle should also be applied to the country's refugee policies. Carens also points out that Walzer does not make clear the relationship between "the low-cost proviso of the mutual aid principle" and "a concern for communal self-determination," for which Walzer argues that if the cost is really low for the national community, refugees can be accepted (Carens 33). The two seem to be in conflict, and it is hard to determine what a low cost refers to. In addition, the excuse of preserving national culture gives "states virtually unlimited discretion in entrance"

(Gibney 173). Which is integral to a culture and which is not is also subject to additional investigations. Therefore, cosmopolitanism is better suited to the issue of refugees.

Nevertheless, utilitarianism and communitarianism provide inspirations for more practical socio-economic and cultural considerations. In light of the utilitarian marginal analysis, I will analyze China's economic capacity and social acceptance level in determining the number of Syrian refugees China should accept. In light of the communitarian emphasis on national culture, I will examine China's national culture to see if it justifies the taking in of Islamic refugees.

4.2. China's Economic Capacity

China is now considered an economic giant, boasting the second-largest economy in the world. According to 2017 data, its reported GDP is 19.4 trillion dollars; its growth rate is 6.9%; its unemployment rate is only 4.6%. Compared to the United States' 2.4% growth and 5.3% unemployment, China seems to be on a better track in an overall sense (Heritage.org).

Meanwhile, according to the UN Convention, host states have the responsibility to grant refugees rights including "access to the courts, to primary education, to work, and the provision for documentation, including a refugee travel document in passport form" (United Nations). Adding to this, I consider that at the initial stage of arrival, basic living supplies, food, and housing should also be provided to Syrian refugees. Therefore, I will consider primary education, housing, food, and basic supplies in calculating the costs China should cover for a Syrian refugee. Additionally, assuming that China will accommodate Syrian refugees to its mid-range cities instead of large cities or rural areas, I will use Shenyang city as a standard. In Shenyang per month, a large apartment outside the city center costs 306 US dollars, a basic diet is around 40 US dollars, and basic living supplies cost around 25 US dollars (Numbeo.com). Based on China's current education policy, the government offers around 83 US dollars for a child's primary education per year, and at the same time, the student's family needs to pay almost nothing. Since a large department generally accommodates six people and consists of two children, we can calculate the cost of a refugee for a year accordingly:

$$\frac{[306 + 40 * 6 + 25 * 6]}{6} * 12 + 83 * \frac{2}{6} = 1420$$

Therefore, a refugee costs China US\$1,420.00 per year with a basic living standard in China's mid-range cities. When compared to China's current economic strength, I find the cost of accepting Syrian refugees relatively low, enabling China to take in a fairly large number of Syrian refugees. Such a number should not be as large as what the Singers propose would have caused the negative influence on domestic citizens to outweigh the benefits received by refugees, for under such circumstances China's internal chaos could already be extremely serious. Moreover, since 82 million people in China still lived under one dollar per day in 2014, it does not make sense for China to accept too many refugees (The Wall Street Journal). The number shouldn't be as small as what Walzer proposes, for such a take-in will not suffice for the inherent right of free movement and will have little impact on relieving the Syrian refugee crisis. A more moderate standard that suits China should therefore be found.

I recommend taking in two million refugees, which based on my calculation, means that China has to pledge 2.84 billion USD to maintain Syrian refugees' lives in China. I propose this number because this 2.84 billion USD accounts for around 10 percent of

China's 2017 budget spending on education or public security—25.3 billion USD and 30.6 billion USD, respectively (Ministry of Finance of the PRC). Despite the subjectivity of my proposed number, in my opinion 10 percent might be a good measure. It is highly practical for a country as large and prosperous as China to pledge 2.84 billion USD to save human lives—a mere \$2 per person given China's 1.4 billion population. For the international community, a two million take-in can already significantly reduce the intensity of the Syrian refugee crisis—according to the UN, in 2017, the number of registered Syrian refugees was 5.1 million (Aljazeera).

I also contend that refugees should only stay in China temporarily—ideally less than 10 years, with exceptions such as marriages and financial hardships that prevent them from returning to Syria. Part of the reason is that after the conflict, refugees should return to Syria to rebuild their home and continue their national culture; otherwise, Syria will suffer a huge loss in its labor force and cultural practices. I suggest China let refugees decide on their own whether to stay or not, as enshrined by the inherent right of free movement. At the same time, however, China should end all additional financial subsidies to Syrian refugees. In this way, those who have already flourished and enjoy their lives in China might stay—and China should grant them the status of permanent residence—while those who have not will leave. Therefore, China has only to provide financial supports for refugees for a short period of time, which puts little negative influence on the country's economy.

The question of enough jobs for refugees also raises doubt about China's acceptance capacity. Here I will provide a brief analysis. Since China is currently entering the stage of an aging society, in the long term the country faces a shortage of labor. According to Foreign Policy, the arrival of refugees injects new laborers to keep China's economic growth in the long run (Pan). Therefore, China does not lack jobs for refugees.

4.3. Social Acceptance

China also ranks high for welcoming attitudes towards refugees among its people. According to a 2016 Amnesty International Report, China is ranked in the top-three refugee-welcoming countries, along with Germany and Britain, with 86% of the survey respondents showing a favoring attitude towards accepting refugees (Aljazeera).

China currently holds around 300,000 Vietnamese refugees, who fled to China during the 1979 China-Vietnam conflict. According to UNHCR, these Vietnamese refugees have been well integrated into Chinese society. Many of them are economically and socially well-off. The younger generation of refugees commonly identifies more with China than with Vietnam. António Guterres, UN High Commissioner for Refugees, on a visit to China in 2006 said the country has “one of the most successful integration programs in the world” (UNHCR). Thus, China has proven its capacity to successfully integrate refugees into local communities and provide them with an appropriate standard of living, and I believe a policy similar to that which enabled Vietnamese refugees to flourish can also be applied to Syrian refugees.

4.4. Cultural Justification

4.4.1. Confucianism

Traditional Confucianism plays a key role in the cultural justification of taking in refugees. According to Yan Xuetong, a leading Chinese scholar of international relations, Confucianism justifies China's moral responsibilities around the globe. Based on Xunzi and Mencius' works, an influential state has a universal moral responsibility to the world (Yan 1). As an aspiring state which wishes to gain more influence on the international stage,

China is therefore obliged by its own traditional culture to take on the responsibility of refugees. What's more, since Confucianism justifies this take-in, there will probably not be too many domestic objections to the issue.

Interestingly, while Kant claims individual rights, Confucianism seeks to advance a state's moral responsibility. There seems to be an inherent conflict between the two frameworks, but in regard to the issue of refugees, both theories point to the same solution—taking in refugees, either to respect individual rights or to fulfill the state's responsibility.

4.4.2. Historical Muslim Communities in China

Historically, several Muslim communities, including the Hui and the Uighurs, have been well integrated into Chinese society and culture. The Hui, with a population of more than 10 million, mainly reside in the Ningxia Province. Their dietary preferences are similar to those of the Han and their Islamic traditions incorporate Han elements—their Mosques are built in Han style (The Economist). Their presence marks a well-integrated Muslim community in China, and at the same time, their unique traditions constitute an important element of Chinese culture—for example, Han Chinese enjoy the ethnic food of the Hui and Uighur community. According to Foreign Policy Magazine, given Hui's cultural integration and political allegiance, China's Communist government also grants the Hui enormous freedom to practice their religion (Foreign Policy). Islamic practices are well preserved and flourish in the Hui community. The Uighur, although persecuted by the government under the excuse of terrorism and secessionism, have built fabulous communities for hundreds of years in the distant Xinjiang region (Foreign Policy). The two Muslim communities show that Muslims can flourish in China; therefore, we can expect a prosperous Syrian refugee community.

Concerning accusations of terrorism against the Uighur, the Chinese government might consider whether taking in Syrian refugees poses a threat of terrorism. Although some Islamic extremists do exist, China can implement close inspections both before and after the acceptance of refugees to ensure national security. As Carens contends, such "public order argument" against refugees should not be overused and our fundamental cosmopolitan right should be prioritized most highly (Carens 259). Therefore, China should not use an ambiguous "public order argument" to refuse taking in refugees, and meanwhile, measures can be taken to ensure the refugees are not terrorists.

4.5. Recommendations

4.5.1. Martha Nussbaum and Iris Marion Young

Martha Nussbaum argues for the Capabilities Approach, which seeks to maximize an individual's capability to achieve what he or she desires. In light of Nussbaum's principle, China needs to take measures to ensure that refugees' capabilities are maximized. Iris Young speaks of a structural injustice, which includes a positional difference and a cultural difference. The positional difference refers to the socially entrenched bias against certain groups of people, and the cultural difference refers to the discrimination originating from a difference in cultural values. Syrian refugees, both socially and culturally (especially in terms of language) disadvantaged, might suffer from both. Therefore, China also needs to implement certain measures to minimize such structural injustice and advance refugees' capabilities. Below are measures worth considering.

4.5.2. Buddy System

A buddy system, in which historical Muslim communities—the Uighur and the Hui—help newly arrived Syrian refugees to integrate into Chinese society, can be developed. With similar Islamic traditions, these domestic Muslims can more easily communicate with the Syrian Muslims, and therefore pass knowledge and cultural values necessary for living in China to them.

There are also established Muslim communities which have migrated to China in recent years. For example, in the flourishing commercial city Yiwu, there is already a well-established Muslim community, which includes Muslim migrants from Yemen, Iraq, and Syria, etc. (Chen). These newly arrived Muslims have a similar experience to that of Syrian refugees. They can therefore deliver more direct help.

4.5.3. Education (Language and Jobs)

A language education centered on both Chinese and Arabic should also be provided to all Syrian refugees. An education in Chinese helps them better integrate into Chinese society; an education in Arabic helps them preserve their Syrian cultural identity. Career education is also necessary to further sharpen refugees' career skills and prepare them for Chinese employment.

4.5.4. Housing, Medical Care, Employment

In order to help Syrian refugees quickly set up their life, I suggest governmental subsidies for housing and medical care, and relatively prioritized positions in employment should be provided. Since Syrian refugees arrive in China with few financial resources, payments for housing and medical care should be entirely or partially, depending on the actual situations of both the refugees and the local economics, exempted at the initial stage, but extending the deadline of payments to a few years later when Syrian refugees are believed to have gotten stable jobs. At the same time, in order to reduce the burdens on Syrian refugees, subsidies should also be provided for housing and medical care, i.e. exempting some portion of the fees permanently. The Chinese government should also ensure a favorable labor market for Syrian refugees in order to help them be employed and quickly earn their living in China. This should at least include laws that prevent Syrian job applicants from being discriminated against by potential employers.

5. Conclusion

In this paper, I have examined China's moral obligations towards Syrian refugees, and whether China is socio-economically eligible to fulfill its responsibilities, also outlining actions that China can take to better the life of Syrian refugees once they take refuge in China. I have provided an approach I call "pragmatic cosmopolitanism," which mainly draws inspiration from Kant and Carens' cosmopolitanism but also embraces considerations of the Singers' utilitarianism and Walzer's communitarianism. The approach argues that countries should respect the fundamental human right to move across borders freely, but that a country's socio-economic conditions should also be taken into pragmatic considerations. Under this approach, as I argue, China not only has a moral obligation to accept Syrian refugees but is also socio-economically empowered to do so. Noticeably, China's taking in Syrian refugees both relieves their sufferings and also, probably realistically more important for China, boosts China's status as a major power that undertakes responsibilities towards the international community. Currently China is still taking a conservative approach towards refugees, but has already offered a huge amount of financial help; President Xi Jinping proclaimed in January of this year that China will provide \$29.1-million-worth of additional

humanitarian aid to Syrian refugees and the homeless (China Daily). We can therefore expect China to take a major step forward and open its borders to Syrian refugees.

Works Cited

- “2017 年中央一般公共预算支出预算表.” *Ministry of Finance of the PRC*, Ministry of Finance of the PRC, yss.mof.gov.cn/2017zyys/201703/t20170324_2565779.html. Accessed 15 Aug. 2017.
- An, Baijie, and Jingxi Mo. “Xi says more help on way for Syria refugees.” *China Daily*, *China Daily*, 20 Jan. 2017, www.chinadaily.com.cn/china/2017-01/20/content_28005354.htm. Accessed 15 Aug. 2017.
- “Are We Listening? Acting on Our Commitments to Women and Girls Affected by the Syrian Conflict.” *International Rescue Committee*, International Rescue Committee, www.rescue.org/report/are-we-listening-acting-our-commitments-women-and-girls-affected-syrian-conflict-0. Accessed 15 Aug. 2017.
- Bayefsky, Rachel. “Dignity, Honour, and Human Rights: Kant's Perspective.” *Political Theory*, vol. 41, no. 6, 2013, pp. 809–837. *JSTOR*, www.jstor.org/stable/24571373.
- Callahan, Daniel. “Doing Well by Doing Good: Garrett Hardin's ‘Lifeboat Ethic.’” *The Hastings Center Report*, vol. 4, no. 6, 1974, pp. 1–4. *JSTOR*, www.jstor.org/stable/3560586.
- Carens, Joseph H. “Aliens and Citizens: The Case for Open Borders.” *The Review of Politics*, vol. 49, no. 2, 1987, pp. 251–273. *JSTOR*, www.jstor.org/stable/1407506.
- Carens, Joseph H. “Realistic and Idealistic Approaches to the Ethics of Migration.” *The International Migration Review*, vol. 30, no. 1, 1996, pp. 156–170. *JSTOR*, www.jstor.org/stable/2547465.
- Carens, Joseph H. “Refugees and the Limits of Obligation.” *Public Affairs Quarterly*, vol. 6, no. 1, 1992, pp. 31–44. *JSTOR*, www.jstor.org/stable/40435795.
- Chen, Shanshan. “Middle Eastern Migrants Live the Chinese Dream.” *Sixth Tone*, 6 June 2017, www.sixthtone.com/news/1000297/middle-eastern-migrants-live-the-chinese-dream. Accessed 15 Aug. 2017.
- Chalabi, Mona. “Where are the Syrian refugees going?” *The Guardian*, Guardian News and Media, 29 Jan. 2014, www.theguardian.com/news/datablog/2014/jan/29/where-are-the-syrian-refugees-going. Accessed 14 Aug. 2017.
- “China.” *China Economy: Population, GDP, Facts, Trade, Business, Inflation, Corruption*, *Heritage.org*, www.heritage.org/index/country/china. Accessed 14 Aug. 2017.
- “China's other Muslims.” *The Economist*, The Economist, 6 Oct. 2016, www.economist.com/news/china/21708274-choosing-assimilation-chinas-hui-have-become-one-worlds-most-successful-muslim. Accessed 15 Aug. 2017.
- “Chinese most welcoming to refugees, Russians least.” *News from Al Jazeera*, www.aljazeera.com/news/2016/05/china-uk-welcoming-refugees-russia-160519044808608.html. Accessed 15 Aug. 2017.
- Copywriter, Filed by: Kelly Montgomery Fundraising. “Refugee crisis: What's happening on the ground in Greece.” *Mercy Corps*, 11 Dec. 2015, www.mercycorps.org/articles/iraq-jordan-lebanon-syria-turkey/refugee-crisis-whats-happening-ground-greece. Accessed 15 Aug. 2017.
- “Cost of Living in Shenyang.” Jul 2017. Prices in Shenyang, www.numbeo.com/cost-of-living/in/Shenyang. Accessed 15 Aug. 2017.
- Doyle, Michael. “Kant and Liberal Internationalism.” *Toward Perpetual Peace and Other Writings on Politics, Peace, and History*, Yale University Press, 2008.

- “Female refugees face physical assault, exploitation and sexual harassment on their journey through Europe.” *Amnesty International*, Amnesty International, www.amnesty.org/en/latest/news/2016/01/female-refugees-face-physical-assault-exploitation-and-sexual-harassment-on-their-journey-through-europe/. Accessed 15 Aug. 2017.
- Gasper, Des, and Thanh - Dam Truong. “Movements of the ‘We’: International and Transnational Migration and the Capabilities Approach.” *Journal of Human Development and Capabilities*, 10 May 2010. *Taylor & Francis*, doi: 10.1080/19452821003677319. Accessed 14 Aug. 2017.
- Gibney, Matthew J. “Liberal Democratic States and Responsibilities to Refugees.” *The American Political Science Review*, vol. 93, no. 1, 1999, pp. 169–181. *JSTOR*, www.jstor.org/stable/2585768.
- Growing Up Without an Education*. Human Rights Watch, 2016, www.hrw.org/report/2016/07/19/growing-without-education/barriers-education-syrian-refugee-children-lebanon. Accessed 15 Aug. 2017.
- Ikenberry, G. John. “Foreign Affairs.” *Foreign Affairs*, vol. 85, no. 3, 2006, pp. 151–152. *JSTOR*, www.jstor.org/stable/20031977.
- Kleigeld, Pauline. “Editor's Introduction.” *Toward Perpetual Peace and Other Writings on Politics, Peace, and History*, Yale University Press, 2008.
- “Life in Za’atari refugee camp, Jordan’s fourth biggest city.” *Oxfam International*, Oxfam, www.oxfam.org/en/crisis-syria/life-zaatari-refugee-camp-jordans-fourth-biggest-city. Accessed 15 Aug. 2017.
- Martin, David A. “The American Journal of International Law.” *The American Journal of International Law*, vol. 84, no. 4, 1990, pp. 987–991. *JSTOR*, www.jstor.org/stable/2202862.
- “Migrant crisis: Migration to Europe explained in seven charts.” *BBC News*, BBC, 4 Mar. 2016, www.bbc.com/news/world-europe-34131911. Accessed 14 Aug. 2017.
- Pan, Liang. “Why China Isn’t Hosting Syrian Refugees.” *Foreign Policy*, 7 Mar. 2016, foreignpolicy.com/2016/02/26/china-host-syrian-islam-refugee-crisis-migrant/. Accessed 15 Aug. 2017.
- “Quick facts: What you need to know about the Syria crisis.” *Mercy Corps*, 7 Apr. 2017, www.mercycorps.org/articles/iraq-jordan-lebanon-syria-turkey/quick-facts-what-you-need-know-about-syria-crisis. Accessed 14 Aug. 2017.
- Report of the independent international commission of inquiry on the Syrian Arab Republic**. UN Human Rights Council, www.ohchr.org/Documents/HRBodies/HRCouncil/CoISyria/A.HRC.27.60_Eng.pdf. Accessed 14 Aug. 2017.
- Singer, Peter. “Escaping the Refugee Crisis.” *Project Syndicate*, 1 Sept. 2015, www.project-syndicate.org/commentary/escaping-europe-refugee-crisis-by-peter-singer-2015-09. Accessed 15 Aug. 2017.
- Su, Alice. “Meet China’s State-Approved Muslims.” *Foreign Policy*, 2 Nov. 2016, foreignpolicy.com/2016/11/02/meet-chinas-state-approved-muslims-hui-linxia-beijing-compromise/. Accessed 15 Aug. 2017.
- “Syria Regional Refugee Response.” *UNHCR*, UNHCR, data.unhcr.org/syrianrefugees/regional.php. Accessed 15 Aug. 2017.
- “Ten Facts about Refugees in China.” *The Borgen Project*, 11 Jan. 2017, borgenproject.org/ten-facts-about-refugees-in-china/. Accessed 14 Aug. 2017.
- United Nations High Commissioner for Refugees (UNHCR). “UNHCR Syria Regional Refugee Response.” UNHCR,

- data.unhcr.org/syrianrefugees/regional.php#_ga=2.65204776.723186520.1502721430-1063613521.1489759619. Accessed 14 Aug. 2017.
- United Nations High Commissioner for Refugees. "Convention and Protocol Relating to the Status of Refugees." *UNHCR*, www.unhcr.org/3b66c2aa10.html. Accessed 15 Aug. 2017.
- "UN: Number of Syrian refugees passes five million." *Al Jazeera*, Al Jazeera, 30 Mar. 2017, www.aljazeera.com/news/2017/03/number-syrian-refugees-passes-million-170330132040023.html. Accessed 15 Aug. 2017.
- "Vietnamese refugees well settled in China, await citizenship." *UNHCR*, UNHRC, www.unhcr.org/en-us/news/latest/2007/5/464302994/vietnamese-refugees-well-settled-china-await-citizenship.html. Accessed 15 Aug. 2017.
- Walker, Peter. "Migrant boat was deliberately sunk in Mediterranean Sea, killing 500." *The Guardian*, Guardian News and Media, 15 Sept. 2014, www.theguardian.com/world/2014/sep/15/migrant-boat-capsizes-egypt-malta-traffickers. Accessed 15 Aug. 2017.
- Walzer, Michael. *Spheres of justice: a defense of pluralism and equality*. New York, Basic Books, 1983.
- Wong, Chun Han. "More Than 82 Million Chinese Live on Less Than \$1 a Day." *The Wall Street Journal*, Dow Jones & Company, 15 Oct. 2014, blogs.wsj.com/chinarealtime/2014/10/15/more-than-82-million-chinese-live-on-less-than-1-a-day/. Accessed 15 Aug. 2017.
- Yan, Xuotong. "The Sources of Chinese Conduct." *Project Syndicate*, 28 Mar. 2011, www.project-syndicate.org/commentary/the-sources-of-chinese-conduct?barrier=accessreg#0RxoGLL4Xc2R3xF0.99. Accessed 14 Aug. 2017.
- Young, Iris. "Structural Injustice and the Politics of Difference." May 2005, www.kent.ac.uk/clgs/documents/word-files/events/young.paper.doc. Accessed 14 Aug. 2017.



Exposing Heat Stress Differential within the Urban Heat Island

Olivia R. Colombo

Author Background: Olivia R. Colombo grew up in the United States and currently attends Sacred Heart High School in Kingston, Massachusetts. Her Pioneer seminar topic was in the field of environmental science and titled "The Nexus of Buildings, Energy and the Environment."

1.0 Abstract

The urban heat island affects urban health and quality of life, as heat stress is measurably higher within the heat island due to blackbody radiation from urban building materials, pollution, and more concentrated population. Within this island, however, the heat stress is not equally dispersed due to the presence of green and blue space intermingled with the urban landscape. Because the proximity of living space to green and blue space increases the related real estate prices, those with a higher income are able to live in areas with access to parks, trees, better landscaping, and bodies of water. By utilizing Particle temperature and humidity sensors and novelly moving throughout space with a singular sensor, we were able to collect heat stress data about a set geographic course that spanned several areas with varying levels of income. By using a color correlation feature on ArcGIS online, it can be visually concluded that higher income areas have lower heat stress due to lower temperature and slightly higher humidity, while lower income areas tolerate a dry heat from urban cement and asphalt. These heat stress differences within the overall island, when reported, have the ability to motivate urban planning changes to increase access to green space in lower income areas, improving the quality of life.

2.0 Introduction

Climate change, a complex and dynamic result of various causes including greenhouse gases heating the atmosphere, plays a role in the socioeconomic and population density aspects of society. Cities, as highly populated areas with high levels of greenhouse gas emissions from vehicles and factories, are often centers of negative impact on the global climate. Additionally, the emissions and the heat-absorbent surfaces, like asphalt and buildings, cause cities to be an "island" of higher atmospheric temperature in comparison to surrounding areas. This concept, known as the urban heat island effect, causes the average temperature in cities to be higher. Additionally, the temperature within the island may differ due to access to green space and blue space. Higher income areas have increased access to these spaces of lower temperature, and thus lower income urban areas have higher heat stress due to a lack of green and blue space.

3.0 Background

3.1 Motivation

It is established that climate change is severely impacting global health and human safety, and these risks are only increasing as the effects of climate change increase. These effects are particularly impacting the populations in cities due to the urban heat island. However, we are working to measure the differences in temperature and humidity within the heat island to better report these metrics. This data could be used to make informed decisions about city planning and landscape to decrease the heat stress that is concentrated in lower income areas and improve overall health and quality of life.

3.2 Literature Review

As the effects of climate change increase, a result of urbanization, known as the urban heat island effect, is causing high levels of heat stress in cities in comparison to the surrounding suburban or rural areas. The urban heat island effect occurs when “buildings, roads, and other structures in cities absorb heat from sunshine and slowly release it.” (“Urban Heat-Island Effect in Boston | City of Boston” 2017) Consequently, urban areas can be 1-3 °C warmer than the greener, less densely populated areas outside. Due to *blackbody radiation*, the buildings and other urban surfaces absorb the sun and are slower to release that energy as heat in the evenings, leading to temperature differences as high as 12 °C at night. (Epa et al. 2014) This dome of high temperature encompassing a city can lead to higher demand and cost for energy for cooling systems, particularly in the summertime.

Air pollution, greenhouse gas emissions, and the related rise in temperature that make up the urban heat island all contribute to health issues, including:

1. Heat Stress on Respiratory and Cardiovascular Health: Elevated air temperatures raise the ozone and pollutant levels, increasing cardiovascular and respiratory diseases. In modern history, each century has been increasingly warmer than the previous, as the average temperature increase from 1906 to 2005 was 0.9°C. (Riebeek 2010) It is estimated that the heat stress and malnutrition caused by the global warming aspect of climate change will cause about 250,000 deaths between 2030 and 2050 alone. (“WHO | Climate Change and Health” 2017)
2. Increased Asthma Triggers: Over 300 million people are affected by asthma worldwide and extreme heat intensifies triggers, such as pollen and air pollution.
3. Natural Disaster and Extreme Weather: Since 1960, reported weather-related natural disasters have tripled, causing an average of 60,000 deaths annually.
4. Drought and Famine: Extreme temperatures are expected to create variable rainfall patterns and affect the fresh water supplies. (“WHO | Climate Change and Health” 2017) Unpredictable rainfall will likely decrease food supplies and the ability to grow staple crops, especially in developing countries, where 3.1 million people already die from undernutrition each year. (“World Child Hunger Facts - World Hunger Education - World Hunger News” 2017)

Despite the mortality and health effects from climate-induced phenomena, heat stress “kills more people in the United States than hurricanes, earthquakes, tornados, and floods combined.” (Golden et al. 2008) However, even within cities heat-related mortality is not evenly distributed. (Klein Rosenthal, Kinney, and Metzger 2014) A study performed in New York City by Harvard researchers demonstrated that lower income neighborhoods experience more heat stress related deaths, due to a combination of lack of efficient air

conditioning systems, more physical labor related jobs, and buildings of heat-retaining material with a simultaneous lack of trees. This lack of green space, such as parks, trees, and gardens, creates a mostly cement and asphalt environment, which lends itself to black body radiation, when the solar energy is deflected and then absorbed between taller buildings. However, in areas that benefit from a green space, such as an urban park, the immediate surrounding neighborhood and the area up to 100 meters from the park experiences a lower temperature of up to 7°C. (“ASLA 2010 Student Awards|The Cooling Ability of Urban Parks” 2017) Blue space, such as urban rivers and bodies of surface water, can also help reduce the heat island through the cooling effects of evaporation and water absorption of solar energy. (Hathway and Sharples 2012)

Both urban blue space and green space are “positive amenities” that improve the quality of life in a neighborhood, through community pride, decreased obesity and cardiovascular health issues due to increased exercise, and less heat stress-related medical issues. (Haeffner et al. 2017) As real estate with access to green and blue space typically comes at a higher price, higher income neighborhoods have the means to benefit from the cooler temperature difference, which when combined with access to efficient cooling systems lends itself to overall less heat stress and the related medical issues.

Given this, a lack of blue and green space creates heat stress in low income areas where there is already oftentimes lower access to cooling systems. The heat stress difference within the overall heat island is unreported, and with this knowledge, city planning and urban landscape could be redesigned to improve the quality of life for lower income urban neighborhoods.

4.0 Methods and Tools

4.1 Tools

1. FlirOne Thermal Imaging Camera
2. iPhone with Particle, FlirOne, and MapMyRun iOS applications
3. Income distribution map (<https://arcg.is/0PneCu>)
4. Particle Electron with temperature & humidity sensor
5. Particle Photon with temperature & humidity sensor

4.2 Procedure

The first data collection was August 23, 2017 from 2:22 PM to 3:41 PM in Boston, Massachusetts, surrounding the campus of Boston University. Before going to collect data in Boston, we plotted a course in ArcGIS Online that circled the Boston University campus and went through three levels of average household income on the ArcGIS online “2017 USA Average Household Income” public map layer.

Once at Boston University, we drove the course with the Photon and Electron Particle temperature and humidity sensors propped out the windows. The car was stopped to take a series of photos with a FlirOne thermal imaging camera to be stitched into a panorama image in Flir Tools. A total of 8.65 miles were driven and recorded with the mobile application MapMyRun, through both heavily cement lower income neighborhoods as well as more affluent areas that contain more parks and were closer to the Charles River.

The data was collected initially by the CHAOS lab CHAOS-BOX server running a python script to capture the RestAPI data posted by Particle from the photon and electron, which was then stored in an InfluxDB database and displayed in Grafana data visualization tool, and downloaded as text file for evaluation in Excel and other tools.

The MapMyRun route was exported as a .gpx file and layered over the income map layer on ArcGIS online. The data collected by Particle, including two sensors' data of temperature and humidity, and a timestamp, was opened as a .csv in Microsoft Office Excel. The location data and corresponding time stamps from MapMyRun were converted from a gpx to a .csv file and opened in Excel as well. In order to match temperature and humidity data with their appropriate GPS coordinates, both data sets were copied in the same Excel file.

The date and time from the .gpx file were split using text to columns feature, and then the vertical lookup function was used to go through the timestamps and match them using approximate match. The function pulled the corresponding coordinates with the matching timestamp and outputted the column of coordinates next to the matching Particle data.

The resulting .csv file that contained matched timestamps, coordinates, temperature, and humidity data was then upload as a layer into the ArcGIS map that already contained the income distribution Living Atlas layer. Two layers were created from the .csv file: one named "temp CHAOS_PH017" that plots the coordinates and temperature data, and one named "humidity CHAOS_PH017" that plots coordinates and humidity data. The "count and amounts" mapping tool was used to display the value of the data in a corresponding color scale.

4.3 Thermal Imaging

Using a FlirOne thermal imaging camera, we took pictures of select points along the course near Boston University, highlighting areas of higher income, lower income, and areas with high levels of cement or green space. The five locations are shown as pinpoint on the map below.

Several photos were taken in one stationary location while pivoting the camera, making sure there was overlap between the images. The images were then exported from the Flir iOS application and imported to Flir Tools+. The panorama tool was used to stitch together each series of images to make a full panorama visual.

5.0 Results and Discussion

5.1 Data

In Figure 1, the green blocks refer to average household income, with the darkest green indicating the higher income and the lightest green correlating to lower income areas. The pink dots are each a singular data point recorded by the Photon temperature sensor (accurate to ± 0.1 degrees Celsius). The lighter pink is warmer temperature, which was hypothesized to correlate to the lighter green, the lower income areas.



Figure 1. Temperature ArcGIS Map. Temperature data collected by the Particle Photon while driving through neighborhoods encompassing a range of average household income.

Figure 2 contains the green average household income Living Atlas ArcGIS layer, as well as the humidity data collected by the Particle sensors while moving through space, rather than calibrating multiple sensors at multiple locations. The lighter orange and lower humidity percentage was hypothesized to correlate with the light green and lower income areas.



Figure 2. Humidity ArcGIS Map Layer
Humidity data collected by the Particle Photon while driving through neighborhoods encompassing a range of average household income.

At multiple locations in each income level, we took a series of images with a Flir One thermal imaging camera that were later stitched into panoramas. These images

provided insight and a second source of the temperature in particular locations. The temperature ranges from 16.6°C (dark purple) to 45.8°C (light yellow, almost white).

Figure 3. Thermal Image Panoramas

Figure 3A. Low Income, High Cement and Asphalt

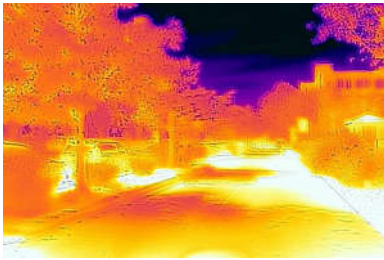


Figure 3B. Higher Income, Green Space (park)

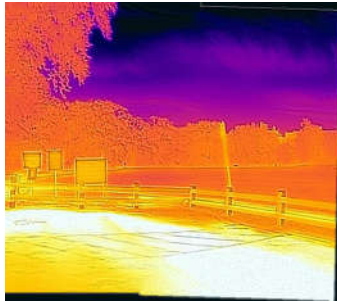
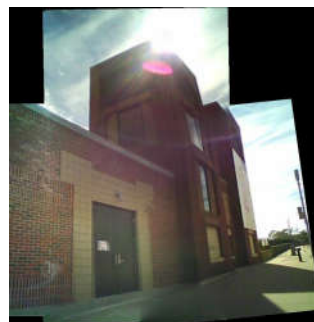
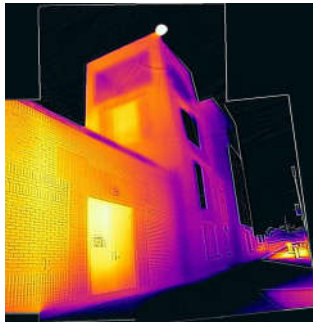


Figure 3C. Lower Income, Main Road



Figures 3D. Right View of Figure 3C.



Figure 3E. Upward Shots between Building (black body radiation)



Figure 3F. Wooded Area, Pavement

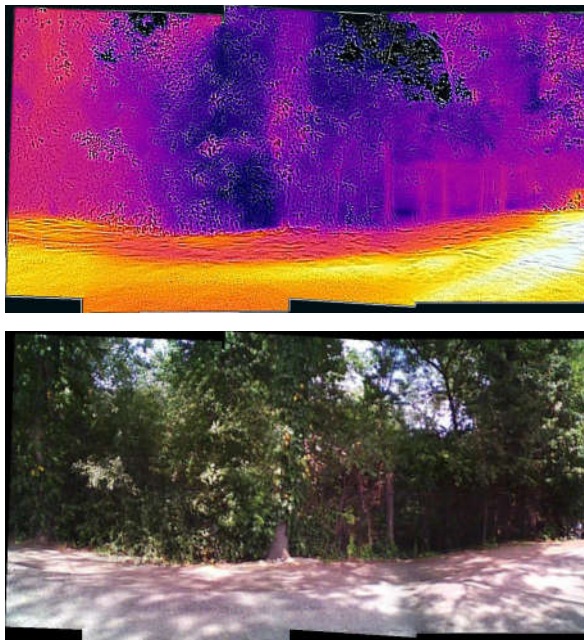


Figure 3G. *Upward Shots between Building (black body radiation)*



5.2 Discussion of Data

After superimposing the ArcGIS Living Atlas Income Distribution layer with the data collected by Particle, it becomes clear that the colors indicate a definite correlation between the two layers. We also observed this correlation while driving with the sensors and watching the temperature drop and humidity rise when we drove into a neighborhood with higher average income. These initial field observations were clearly supported by the visual correlation of colors in the ArcGIS map.

5.2.1 Temperature

From visual observation and research on income-dependent access to green and blue space, there was much more cement and asphalt in lower income areas. Higher income areas were populated with more brick buildings, gardens, trees, green space, parks, and were located closer to the Charles River. Black body radiation between taller buildings and the abundance of heat-absorbent building materials lead to a noticeable increased temperature difference in those areas, as supported by the temperature map above and the pink temperature data (lighter pink for a higher temperature and dark pink for cooler temperature).

5.2.2 Humidity

In the humidity map above, the higher humidity (displayed as darker pink) clearly increases in areas of higher income. However, the Particle Photon measured relative humidity, not absolute humidity. When consulting a psychometric chart and taking into account that the humidity sensors have an accuracy of $\pm 3\%$, little to no change is seen in the absolute humidity when moving through the different income areas. Despite this, the temperature sensor with a more precise reading recorded a definite change. This is due to the abundance of observed green and blue space that this more expensive real estate has access to. The trees, green areas, and parks allow for evapotranspiration, when water evaporates into the area through plants, to increase the humidity. This higher level of humidity combined with the cooler temperature make the air feel cooler, leading to a higher quality of life and less heat stress in these areas.

6.0 Conclusion and Future Work

The hypothesis that income distribution within the urban heat island has a direct impact on temperature due to access to green and blue space is supported by the data. By utilizing the Particle sensors and using ArcGIS to map the color correlation between temperature, humidity, and income areas, we are able to visually conclude that income areas correspond to the Particle data. Higher income areas have cooler temperature and a higher humidity due to green spaces, which ultimately leads to a more comfortable climate with much less heat stress. On the other hand, black body radiation and the absorption of heat into building materials like cement cause lower income areas with less access to green and blue space to have higher heat stress. This underreported heat stress differential within the heat island is currently causing unnecessary health and quality of life issues for those living in lower income neighborhoods. With the support of the above data, we can conclude that the heat stress in these areas could be greatly improved, by several degrees Celsius, by the installation of green space. By planting trees, working on adding green space to median and sidewalks, as well as the addition of parks, heat stress and the overall urban heat island will be reduced, leading to an improvement in urban heat and quality of life.

7.0 Works Cited

- “ASLA 2010 Student Awards | The Cooling Ability of Urban Parks.” 2017. Accessed August 24. <https://www.asla.org/2010studentawards/169.html>.
- Epa, U. S., OAR, OAP, and CPPD. 2014. “Heat Island Effect,” February. <https://www.epa.gov/heat-islands>.
- Golden, Jay S., Donna Hartz, Anthony Brazel, George Luber, and Patrick Phelan. 2008. “A Biometeorology Study of Climate and Heat-Related Morbidity in Phoenix from 2001 to 2006.” *International Journal of Biometeorology* 52 (6): 471–80.
- Haeffner, Melissa, Douglas Jackson-Smith, Martin Buchert, and Jordan Risley. 2017. “Accessing Blue Spaces: Social and Geographic Factors Structuring Familiarity With, Use Of, and Appreciation of Urban Waterways.” *Landscape and Urban Planning* 167 (November): 136–46.
- Hathway, E. A., and S. Sharples. 2012. “The Interaction of Rivers and Urban Form in Mitigating the Urban Heat Island Effect: A UK Case Study.” *Building and Environment* 58 (December): 14–22.
- Klein Rosenthal, Joyce, Patrick L. Kinney, and Kristina B. Metzger. 2014. “Intra-Urban Vulnerability to Heat-Related Mortality in New York City, 1997-2006.” *Health & Place* 30 (November): 45–60.
- Riebeek, Holli. 2010. “Global Warming: Feature Articles.” NASA Earth Observatory. <https://earthobservatory.nasa.gov/Features/GlobalWarming/page2.php>.
- “Urban Heat-Island Effect in Boston | City of Boston.” 2017. Accessed August 24. <https://www.cityofboston.gov/climate/urbanheatislandeffectboston.asp>.
- “WHO | Climate Change and Health.” 2017, July. World Health Organization. <http://www.who.int/mediacentre/factsheets/fs266/en/>.
- “World Child Hunger Facts - World Hunger Education - World Hunger News.” 2017. *World Hunger News*. Accessed August 24. <http://www.worldhunger.org/world-child-hunger-facts/>.



Fast Video Retargeting Based on Seam Carving with Parental Labeling

Chuning Zhu

Author Background: Chuning Zhu grew up in China and currently attends The Experimental High School Attached to Beijing Normal University in Beijing, China. His Pioneer seminar topic was in the field of computer science and titled "Computers That See: Exploring New Techniques of Computer Vision."

Abstract

Seam carving is a state-of-the-art content-aware image resizing technique that effectively preserves the salient areas of an image. However, when applied to video retargeting, not only is it time intensive, but it also creates highly visible frame-wise discontinuities. In this paper, we propose a novel video retargeting method based on seam carving. First, for a single frame, we locate and remove several seams instead of one seam at once. Second, we use a dynamic spatiotemporal buffer of energy maps and a standard deviation operator to carve out the same seams in a temporal cube of frames with low variation in energy. Last but not least, an improved energy function that considers motions detected through difference method is employed. During testing, these enhancements result in a 93 percent reduction in processing time and a higher frame-wise consistency, thus showing superior performance compared to existing video retargeting methods.

Key Words

Video retargeting, seam carving, spatiotemporal buffer.

1. Introduction

As more and more smart devices are becoming a part of our lives, multimedia content, especially videos, is viewed on a variety of displays with aspect ratios that run the gamut from 1:1 to 21:9. However, videos are usually generated in a specific resolution with a set aspect ratio. Presenting them on different displays without optimization will result in a compromised visual perception. Therefore, video adaptation algorithms are of great importance to present-day multimedia consumption.

Seam carving, or dynamic resizing, is a content-aware image resizing algorithm first proposed by Shai Avidan and Ariel Shamir in 2007 (Avidan and Shamir, 2007). Unlike standard image scaling, which corrupts the ratio of content, or image cropping, which only discards peripheral pixels, seam carving maintains the salient regions of an image by recursively removing or inserting the seam with the lowest cumulative energy. A seam is an 8-connected path that crosses the image either vertically or horizontally. The cumulative energy of a seam is defined via an energy map, or a visual saliency map. The map assigns a value to each pixel according to an energy function (gradient energy, entropy etc.) and the cumulative energy of a seam is the sum of all pixel energy values in its path. Under optimal

conditions, seam removal retains the salient regions of an image while creating barely noticeable artifacts.

Despite its satisfactory performance on a single image, the original seam carving algorithm is not ideal for video retargeting. First, the fact that it is based on dynamic programming renders the algorithm excessively time-consuming for the purpose of video retargeting. Resizing an individual image can be done in a reasonable amount of time; video retargeting, however, requires the adaptation of so vast a number of frames that the length of a video and the time needed for its retargeting is extremely disproportional. Second, the result of applying seam carving to separate frames is visually inconsistent. Since the seam calculation of each frame is independent of others, there are conspicuous fluctuations in the output which greatly sabotage the visual appeal. For instance, an edge may move left or right in different frames since the locations of seam removal may change.

In this paper, we present a seam-carving-based video retargeting algorithm that runs in reasonable time while retaining frame-wise consistency. We propose a multiple seam searching and removing method based on what we call Parental Labeling (SCPL). It can be demonstrated that all the seams are derivations of some parents in the first row or column. By labeling all the children (pixels in the last row or column) with values corresponding to their parents and subsequently removing the minimal child of each parent, multiple seams can be removed in one iteration. Based on the SCPL, we establish connections between frames by creating a dynamic spatiotemporal buffer, whose size is determined by dynamically calculating the average standard deviation of energy map. Once the STD is above a dynamic threshold, the buffer is fixed, and the same seams derived from a uniform energy function are removed across the buffer. Finally, an improved visual saliency map that takes motion into account is employed to stabilize and smoothen the motion.

The rest of the paper is organized as follows. We first review related work concerning improved video retargeting. Then, we elucidate the proposed method in detail. Finally, we evaluate the results and conclude the paper with an analysis of pros and cons and a discussion of future work.

2. Related Work

Previous research has already proposed manifold video retargeting approaches based on seam carving. M. Rubinstein et al. (2008) first refined the original seam carving process to extend its functionality to video retargeting. To achieve this, they introduced a new seam carving method based on graph cut. In this method, a seam is a connected and monotonic S/T cut of the graph representation of an image. Formally, a vertical cut is a continuous function over the relevant domain $S: row \times time \rightarrow column$, and a cut in video is a 2D manifold of a 3D time-frame volume. Another enhancement is the employment of forward energy. The original backward energy calculation, which removes seams with the least energy, may increase the total energy of the image as new neighbors are created. Taking this factor into account, they proposed an algorithm that removes the seam whose removal will insert the minimal amount of energy. There are, however, limitations on this algorithm. Although it protects the structure of the image quite well, it sabotages the shape of content. Furthermore, graph cut has a higher time complexity than dynamic programming, which deviates from what we are exploring in this paper.

As for accelerated video retargeting, Stephan Kopf et al. (2009) proposed a novel seam-carving-based algorithm to adapt videos to mobile screens named Fast Seam Carving for Size Adaptation of Videos (FSCAV). FSCAV combines camera compensation with background aggregation. It also features a criterion for determining robust seams (i.e., seams

in the background that can be removed from frames temporally). Another mechanism proposed by Kopf is the measuring operator of video adaptation. If the resizing factor goes beyond a threshold, other resizing methods such as scaling, and cropping are introduced. As for evaluation, the FSCAV is indeed significantly faster than graph cut methods, but it still requires more than 30 minutes for typical videos with appropriate resolution. A great portion of running time is spent on live analysis of robust seams. The approach is generally limited by camera movements, but the background aggregation technique provides inspiration for our approach.

Other approaches usually modify seam carving for higher parallelism and resort to GPU for multi-threading. Kim et al. (2014) proposed an improved seam carving that is modified to run in parallel on multiple GPUs. Their method has three major novelties. First, they propose Non-Cumulative Seam Carving, an approach that resembles a greedy algorithm, to replace the traditional dynamic programming. The algorithm calculates the optimal path for each pixel simultaneously, depending exclusively on the energy values of its neighbors in the following row or column. The major purpose of this algorithm is to parallelize the calculation of the seam map. Then, the authors proposed the Adaptive Multi-Seam Algorithm. Unlike traditional SC that discards one seam per iteration, this algorithm searches for multiple neighboring seams with distances less than a threshold, and removes several seams at a time. Lastly, they elaborate on the syncing of multiple GPUs.

Another GPU-based method is proposed by C. Chiang et al. (2009) The major contributions of this method are seam-split and JND (Just Noticeable Distortion) based saliency map. The former enables the algorithm to remove multiple seams at a time by splitting “local seams” that branch off from “global seams,” while the latter improves the quality of video to make it more visually appealing. With GPU acceleration, the algorithm can achieve a frame rate ranging from 10 to 30 fps, depending on the resolution of test cases.

Both of the GPU-based methods are impractical for wide application, as the majority of mobile and immobile display devices do not possess such powerful multi-threading capability. Nevertheless, they do enlighten us about possible aspects and directions of optimization.

Recently, work concerning improved visual saliency map optimized for video retargeting has emerged. Duan-Yu Chen et al (2011) proposed a novel approach based on visual saliency cube to retarget videos. First, it detects the salient points in spatiotemporal domain using modified Harris Detector. Then, using these salient points as a seed for searching, the method constructs a motion attention map where consistent motions correspond to higher value. Finally, seam carving is imposed on the motion attention map. Since moving areas have higher values, they are protected from removal. Although this approach prevents discontinuities in motion that result from ordinary seam carving, its limitation lies in the assumption that there is no camera movement. In other words, the camera has to be fixed in order for the searching algorithm to function. Detection of motion is of great importance in video retargeting as it maintains the coherence of moving objects. In our paper, a more simplistic motion-aware visual saliency map is used for the purpose of fast video retargeting.

3. Fast Video Retargeting Based On SCPL

3.1 Seam Carving with Parental Labeling (SCPL)

According to the seam carving algorithm, a seam is defined as an 8-connected path that cuts through the image either vertically or horizontally, and contains one and only one pixel in each row or column respectively. Formally, let I be an $n \times m$ image, and a vertical seam is defined via:

$$S^x = \{s_i^x\}_{i=1}^n = \{(x(i), i) | i \in \{1, \dots, n\}, |x(i) - x(i-1)| \leq 1\}, \quad (1)$$

where x is a mapping $x: [1, \dots, n] \rightarrow [1, \dots, m]$.

Similarly, if y is a mapping $y: [1, \dots, m] \rightarrow [1, \dots, n]$, a horizontal seam is defined via:

$$S^y = \{s_j^y\}_{j=1}^m = \{(j, y(j)) | j \in \{1, \dots, m\}, |y(j) - y(j-1)| \leq 1\}. \quad (2)$$

Since vertical and horizontal seam carving are largely congruent, and horizontal seam carving can be done by rotation and vertical seam carving, for the sake of simplicity, the illustrations below will only address vertical seam carving.

Given an energy function e , the cost of a seam is defined as $E(S) = \sum_{i=1}^n e(P(S_i))$, where $P(S_i)$ is the position of the i th pixel along the path of a seam and e is the energy of a pixel. Typically, we want to search for and remove the seam with the minimal cost, or S_{min} such that:

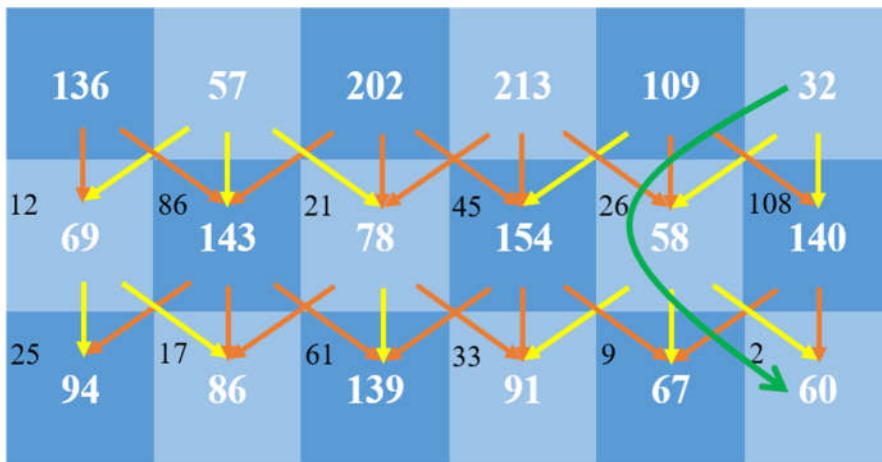


Figure 1. Illustration of the seam searching process. Black numbers are pixel energy; white numbers are cumulative energy. Orange arrows indicate eliminated paths; yellow arrows indicate selected paths; the green arrow indicates the seam with lowest energy.

$$E(S_{min}) = \min\{E(S)\} = \min\left\{\sum_{i=1}^n e(P(S_i))\right\}. \quad (3)$$

The searching process is done through dynamic programming. We traverse the image row by row, and for each pixel (ij) in a row, a selection is made between three possible paths above it, namely $(i-1, j-1)$, $(i-1, j)$, and $(i-1, j+1)$. The pixel is connected to the path with lowest cumulative energy, ensuring that the updated cumulative energy is the least among all the possibilities. Formally, the cumulative energy is updated via the following function:

$$CE(i, j) = e(i, j) + \min(CE(i-1, j-1), CE(i-1, j), CE(i-1, j+1)). \quad (4)$$

At the end of traversal, the least cumulative energy in the last row indicates the least significant seam. During the traversal, an array of indicators is created to record the path. Seam removal can thus be done through tracing back the indicators. Since a seam is monotonic (i.e. there is exactly one pixel in each row or column), seam removal has only a local effect on the image.



Figure 2. Presentation of seams in images.

Note that the collection of seams can be regarded as a mathematical function. We can define the collection of pixels in the last row as set L , and the collection of pixels in the first row as set F . Then the collection of seams can be represented as function $SE: L \rightarrow F$. Recall that a function is a relation between sets that satisfies left-totality and right-uniqueness. Define relation SE between L and F as $\{(l, f) \in L \times F \mid l \text{ and } f \text{ are connected via a seam}\}$, the proofs of the two properties of functions are as follows:

a. Left-totality: During the last iteration of traversal, by formula (4), each pixel from the last row is connected to one in three paths from above, which is connected to some parents. $\forall l \in L \exists f \in F \text{ st } lSEf$. The relation is thus left-total.

b. Right-uniqueness: Each seam connects one pixel from the last row with exactly one pixel from the first row. $\forall l \in L \forall f, g \in F, (lSEf \wedge lSEg) \Rightarrow (f = g)$. The relation is thus right-unique.

Observe that the set containing pre-images of each element in the domain of SE is a partition of domain (pixels in the last row). In other words, each pixel in the last row can be classified by its origin into discrete categories. Specifically, let a be in the domain of SE, L , a partition $P(L)$ of L is defined as:

$$P(L) = \{ \text{Preim}_{SE}(\{a\}) \mid a \in \text{Im}_{SE}(F) \}. \quad (5)$$

Therefore, we can label each element of L (each pixel in the last row) with its image (pixels in the first row), or what we figuratively name “parent”. Each child has one and only one parent, while each parent has multiple children. Since all the elements of partition $P(L)$ are pairwise disjoint, there is no overlap between the children of any two parents. After the removal of a child from one parent, the children of all other parents remain intact. In addition, since for each pixel in the image there is only one set path that can be traversed backwards, no two seams have intersection with each other. (If two seams intersect, then for the intersection point there are two possible paths to trace back to, which is impossible.) It is thus ensured that the output has uniform width. With all the restrictions clarified, it is viable to remove the child with the least cumulative energy from each parent without jeopardizing the overall structure of an image. The indices of target seams can be using through the following function:

$$Ix(I) = \{ \text{argmin}(CE(a)) \mid a \in P(L) \}, \quad (6)$$

where I is the input image and $P(L)$ is the partition defined above. With parental labeling, multiple seams can be removed at a time. The processing time is thus accelerated to various extents depending on the input.

3.2 Spatiotemporal Buffer of Energy

The dynamic programming nature of seam carving casts a limit on optimization of single-frame retargeting. When we expand the scope to video retargeting, however, there exist more possibilities of improvement. Through observation, we notice that videos usually include short periods with slight variation of content. Applying seam carving separately on each frame in such a period is redundant, as the calculations are highly repetitive. Moreover, the barely noticeable differences lead to minor alterations of seam positions, which in turn accumulate to visible displacement of content between processed frames, i.e. jittering. In this paper, we propose a dynamic spatiotemporal buffer whose elements can be applied with seam carving uniformly. Figuratively, a spatiotemporal buffer is a cube whose cross sections are frames. There are two problems to be addressed about this construction: the size of the buffer and its energy map.

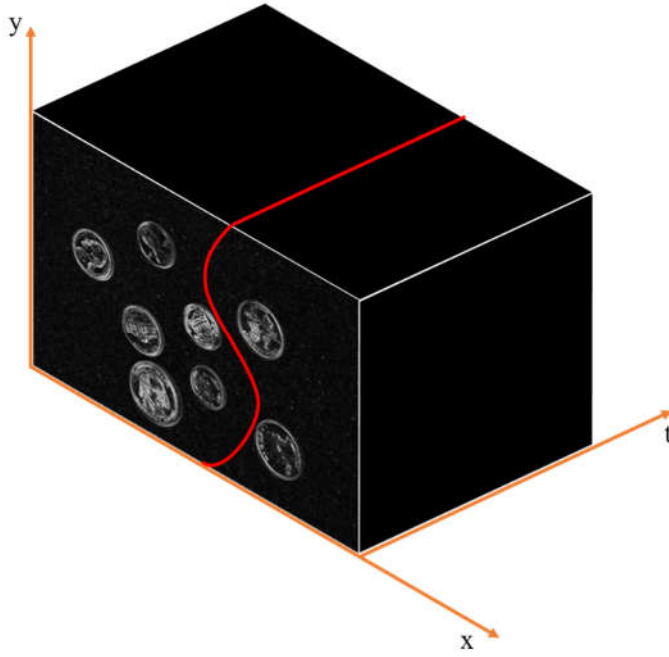


Figure 3. Spatiotemporal buffer of energy maps.

To determine the size of the buffer, we present an operator based on average standard deviation of energy (ASDE). Typically, frame-wise difference is directly proportional to variation of energy, i.e. lower frame-wise difference corresponds to lower variation of energy. Therefore, we construct two arrays, one as a container of color frames, the other as the spatiotemporal buffer of energy maps. The initial lengths of the two arrays are zero, yet they grow simultaneously. With each increment of frame, an ASDE is defined as follows.

$$ASDE(I) = \frac{1}{n \times m} \sum_{i=0}^n \sum_{j=0}^m Std(i, j), \quad (7)$$

where I is an energy map with size $n \times m$.

For each pixel $(ij) \in \{1, \dots, n\} \times \{1, \dots, m\}$, its standard deviation in a spatiotemporal buffer $Std(ij)$ with length T is defined via:

$$Std(i, j) = \sqrt{\frac{1}{T} \sum_{t=0}^T (e((i, j)_t) - \mu_{e(ij)}^T)^2}, \quad (8)$$

where e is the energy of a pixel and μ is the average energy of a pixel across time window T .

Using the ASDE as an indicator of variation, we can thus determine the size of the spatiotemporal buffer by comparing its value to a threshold. However, note that the

threshold is varying, for the same upheaval has less effect on the ASDE of a larger buffer than that of a smaller one. To visualize the relation, we plot the ASDE of a spatiotemporal buffer with first n consecutive frames of lowest energy and one frame of highest energy against the variable n , as shown in Figure 4.

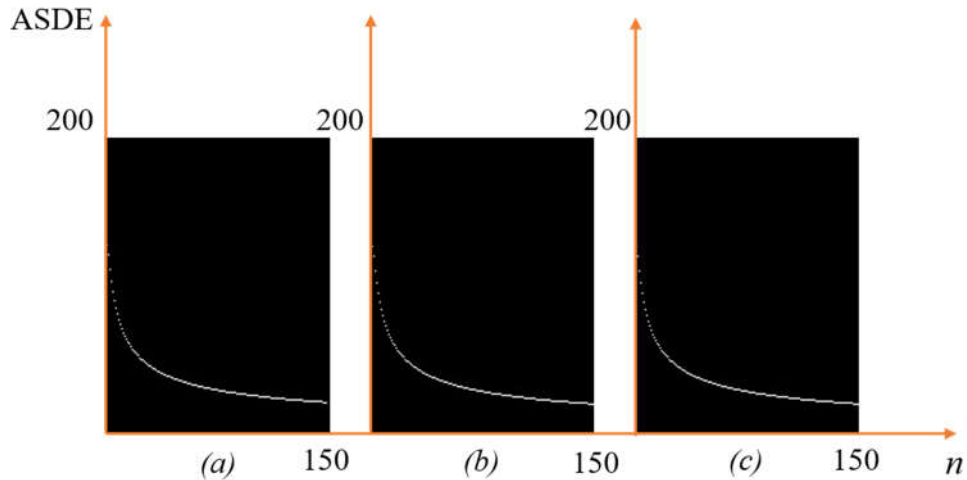


Figure 4. Plots of ASDE against size of spatiotemporal buffer. (a) is a buffer of resolution 640x480, (b) is of resolution 1280x720, (c) is the plot of proposed function (9). As can be observed, there are hardly any differences between the curves, showing that (1) the curve is unaffected by resolution; (2) our proposed function optimally fits with the relation between buffer size and threshold value.

We hereby propose a function of threshold value with respect to the size of a buffer:

$$\varphi(n) = \alpha \sqrt{\frac{(\maxval - \frac{\maxval}{n})^2 + (n-1)(\frac{\maxval}{n})^2}{n}}, \quad (9)$$

where n is the size of buffer, α is a variable parameter with an optimal range of $[0.15, 0.25]$, and \maxval is the maximal value of energy map. Fundamentally, the formula simulates a worse-case scenario where the most radical energy change (from lowest to highest) happens after n consecutive frames. Then we take a fraction (according to parameter α) of its corresponding standard deviation, and use the result to bound the size of our spatiotemporal buffer.

With the spatiotemporal buffer established, the uniform energy map is defined as the average of the collection of energy maps. Same seams are then calculated and removed in sequence from all color frames in the container. The employment of spatiotemporal buffer not only speeds up the seam carving process, but also provides more consistency

across frames. Slight alternation occurs less frequently, making the video less jittery and more visually appealing.

3.3 Improved Visual Saliency Map

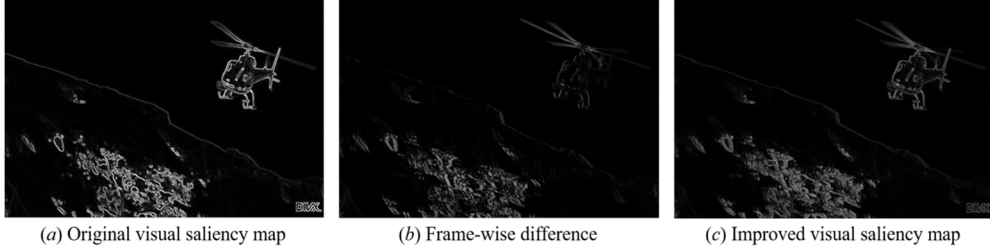


Figure 5. Improved visual saliency map.

Visually, seam removal in video sequences is exceptionally noticeable. The problem is exacerbated with the introduction of spatiotemporal buffer, since motions are more likely to be undercompensated. To protect the consistency of motions, an improved visual saliency map that compensates for frame-wise difference is proposed. Specifically, the energy function takes the absolute difference between the current frame and the previous frame, and blends the result in a weighted manner with the original energy map based on Sobel gradient energy. The weights may vary, but considering the higher priority of motion, more weight should be assigned to difference and less to energy. In our implementation, the weights are 0.6 and 0.4 respectively.

4. Experiment Results

All the experiments are conducted on a PC with Intel Core i7 processor running at 3.4 GHz with 16 Gigabytes of RAM. The methods are implemented with Python 2.7 and OpenCV 3.1.0.

For Seam Carving with Parental Labeling, we conducted two experiments to compare its speed performance with the original seam carving method. First, we ran both seam carving methods recursively, reducing the width of previous results by a factor of 0.8 in each recursion. The results reflect the methods' performances under different resolutions. Second, we ran a direct seam carving, reducing the width of images to various target values in one step. The results reflect the methods' performances when removing different number of seams.

For Fast Video Retargeting, we compared our proposition with raw video retargeting which applies original seam carving to separate frames. We first recorded the time consumption of directly retargeting videos to various widths. Then, instead of recursively retargeting, we retargeted the same video of different resolutions to the same extents. Results are shown as follows.

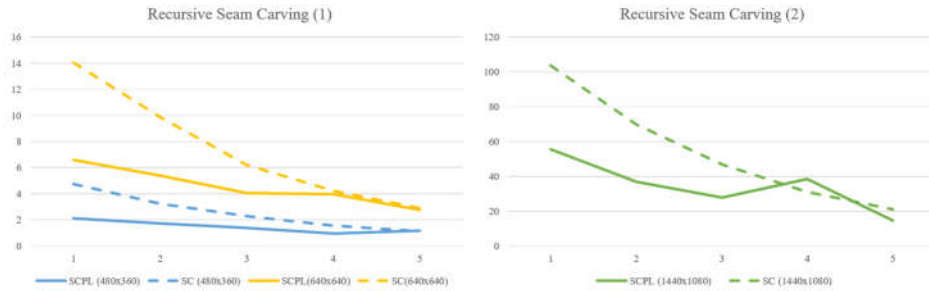


Figure 6. Results of recursive seam carving. The horizontal axis is the number of recursive loops. In each loop the image is resized to 0.8 of its previous width. The vertical axis is the performance of each trial measured in seconds.

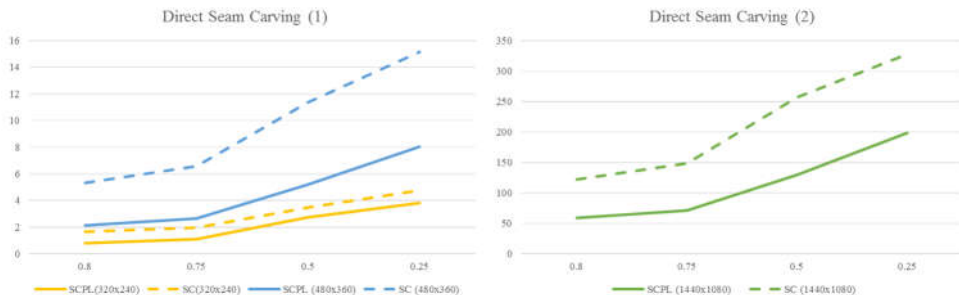


Figure 7. Results of direct seam carving. The horizontal axis is the ratio of target width to original width. The vertical axis is the performance of each trial measured in seconds.

Through an interpretation of data, we discover a universal reduction of processing time using SCPL. The reduction is especially significant when resizing images with high resolution, with SCPL running 50% faster. When the resolution is low, however, the run time of SC approaches SCPL. The reason lies in the fact that in our experiments only the width is resized recursively. When the width-to-height ratio becomes low, there are fewer parents per iteration since more seams derive from the same parents.

Table 1. Performance of Video Retargeting on Various Samples

Name	Resolution	Number of Frames	Target Resolution	Method	Time (seconds)
"Big Belly"	640x480	377	512x480	Proposed	124.36
				Raw	1255.19
			384x480	Proposed	214.24
				Raw	2484.91
"Lion"	640x480	117	512x480	Proposed	80.72
				Raw	600.9
			384x480	Proposed	140.88
				Raw	1225.6
"Lego"	560x320	164	448x320	Proposed	32.28
				Raw	263.89
			336x320	Proposed	56.21
				Raw	532.16
"Helicopter"	720x576	178	576x576	Proposed	143.31
				Raw	1594.19
			432x576	Proposed	282.7
				Raw	3096.67
	320x240		256x240	Proposed	16.32
				Raw	201.71
	192x240		Proposed	31.44	
			Raw	426.67	
	176x144		140x144	Proposed	3.43
				Raw	52.65
			105x144	Proposed	6.73
				Raw	108.34

For video retargeting, our proposed method yields a 90.6% reduction of run time on average. The enhancements are more significant for low resolution videos because the majority of time is consumed by operations on the spatiotemporal buffer, whose time complexity is directly related to the frame resolution. When retargeting sample "Helicopter" from 176x144 to 105x144, for instance, our approach achieves a whopping 93.7% reduction of time. It can thus be concluded that our approach has a universally superior performance over raw image retargeting.

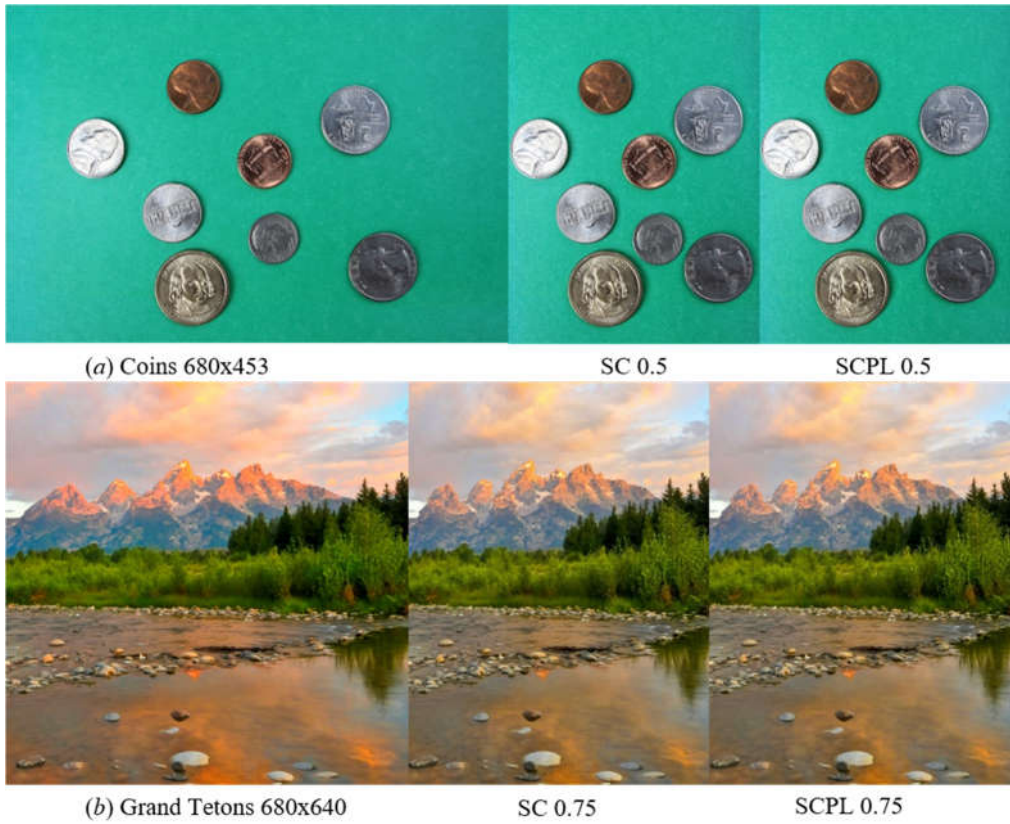


Figure 8. Comparison of image quality between SCPL and SC.

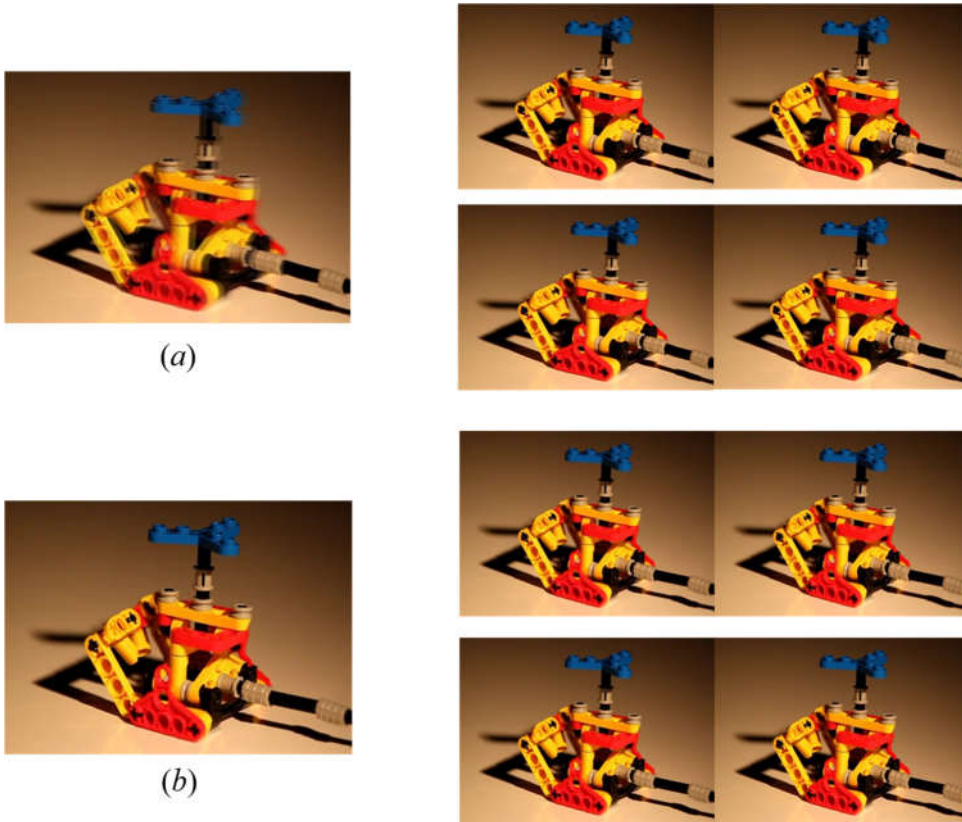


Figure 9. Blending four consecutive frames of both results. (a) is from raw video retargeting, (b) is from Fast Video Retargeting based on SCPL. The original video sequence is nearly static. The result of raw video retargeting is blurry, indicating a displacement of content. Our result yields a sharper image.

Despite the time reduction, it can be observed that SCPL yields qualities similar to if not better than those of original seam carving. Fast video retargeting also achieves similar visual quality to that of raw seam carving. Moreover, due to the spatiotemporal buffers, retargeted videos are significantly more consistent and visually appealing.

5. Conclusion

Fast and high-quality video retargeting methods have always been an intriguing research topic. In this paper, we propose a novel approach that has three major contributions to this field of research. Firstly, we propose an improved single-frame seam carving method, SCPL, that labels each seam with its parent and removes the least significant seam from each parent per iteration. Secondly, we propose a spatiotemporal buffer that gathers a period of frames with low energy variation and applies seam carving on the spatiotemporal cube. We present an operator, average standard deviation of energy (ASDE), to determine the size of buffer. Thirdly, we enhance the energy function specifically for video retargeting by compensating frame-wise motions whose destruction causes conspicuous discontinuities and affects visual perception.

It is demonstrated that our proposition greatly outperforms other seam carving and video retargeting methods. However, due to its low parallelism, its performance is largely confined by the resolution of source and the number of seam removals. Also, although jittering is alleviated in the retargeted video, objects may be slightly distorted as the visual saliency map is less representative of a single frame. In the future, it would be promising to further improve performance by avoiding repetitive process through partial energy update and seam memorization. Distortions can be optimized by the employment of better visual saliency maps.

6. Bibliography

- [1] Avidan, S., & Shamir, A. (2007). Seam carving for content-aware image resizing. ACM SIGGRAPH 2007 papers on - SIGGRAPH 07. doi:10.1145/1275808.1276390
- [2] Chao, T., Leou, J., & Hsiao, H. (n.d.). An Enhanced Seam Carving Approach for Video Retargeting. Retrieved August 27, 2017, from http://www.apsipa.org/proceedings_2012/papers/36.pdf
- [3] Chen, D., & Luo, Y. (2011). Content-aware video resizing based on salient visual cubes. *Journal of Visual Communication and Image Representation*, 22(3), 226-236. doi:10.1016/j.jvcir.2010.12.003
- [4] Chiang, C., Wang, S., Chen, Y., & Lai, S. (2009). Fast JND-Based Video Carving With GPU Acceleration for Real-Time Video Retargeting. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(11), 1588-1597. doi:10.1109/tcsvt.2009.2031462APPENDIX Proof of no intersection.
- [5] Farag, T. G., Xingming, S., & Gaobo, Y. (2013). Video retargeting using matching area with seam carving and object detection. *Future Computer and Information Technology*. doi:10.2495/icfcit130871
- [6] Grundmann, M., Kwatra, V., Han, M., & Essa, I. (2010). Discontinuous seam-carving for video retargeting. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. doi:10.1109/cvpr.2010.5540165
- [7] Kim, I., Zhai, J., Li, Y., & Chen, W. (2014). Optimizing Seam Carving on multi-GPU systems for real-time image resizing. 2014 20th IEEE International Conference on Parallel and Distributed Systems (ICPADS). doi:10.1109/padsw.2014.7097861Kopf, S., Kiess, J., Lemelson, H., & Effelsberg, W. (2009). Fscav. Proceedings of the seventeen ACM international conference on Multimedia - MM 09. doi:10.1145/1631272.1631317Rubinstein, M., Shamir, A., & Avidan, S. (2008). Improved seam carving for video retargeting. *ACM Transactions on Graphics*, 27(3), 1. doi:10.1145/1360612.1360615
- [8] Rubinstein, M., Shamir, A., & Avidan, S. (2009). Multi-operator media retargeting. ACM SIGGRAPH 2009 papers on - SIGGRAPH 09. doi:10.1145/1576246.1531329



Home Alone: How the Lack of Chinese Immigrant Women Caused the Failure of Chinese Immigrants to Integrate into the United States, 1848-1882

Felix Hohne

Author Background: Felix Hohne grew up in China and currently attends St. George's School in Vancouver, Canada. His Pioneer seminar topic was in the field of sociology and titled "Hyphenated Americans."

Abstract

The 19th century was the highpoint of foreign immigration into the United States. Most groups successfully integrated into American society, with a clear exception of the Chinese, who were already excluded by the 1882 Chinese Exclusion Act. Three common theories to explain this are religious intolerance, racism, and economic competition. This essay argues that the generation-long lack of Chinese immigrant women is a more decisive factor to explain the early exclusion of the Chinese in the United States.

Introduction

Chinese immigrants first began to arrive in the United States of America in large numbers following the beginning of the 1848 California Gold Rush. In 1853, approximately eighty percent of Chinese living in the United States had emigrated from Guangdong in southern China, a province with a tradition of emigration and trade in Southeast Asia.¹ By the mid-19th century, the power of the governing Qing dynasty, ruling since 1644, had steadily waned. The ensuing chaos and corruption, coupled with natural disasters in 1846 and 1848 that led to massive starvation and poverty, created a "push" factor leading to massive Chinese emigration to Southeast Asia, South America and California.² In addition, the Chinese immigrants were motivated to come to the US because of the start of the Californian Gold Rush, which provided the hope of striking a fortune and created a strong "pull factor", attracting Chinese immigrants to the US in large numbers for the first time.

Chinese immigrants were able to earn far more money in the United States than in China, which encouraged them to stay in America despite widespread racism towards them. A combination of factors ultimately led the United States government to pass the Chinese Exclusion Act of 1882, effectively blocking Chinese immigration. The act severely curtailed the number of new Chinese arrivals until the mid-1960s.

¹ Lani Ah Tye Farkas, *Bury My Bones in America: The Saga of a Chinese Family in California, 1852-1996: From San Francisco to the Sierra Gold Mines*, 1st ed. (Nevada City, Calif: Carl Mautz Publishing, 1998), 2.

² *Ibid.*, 7.

This essay asks why Chinese immigrants to the United States in the latter part of the nineteenth century did not integrate into American society. Various hypotheses for the Chinese exclusion have been put forward, such as that hostilities toward the Chinese immigrant community were primarily caused by the “general xenophobia of white Protestant Americans”, that “economic considerations were most important”, or that the negative public opinion was caused by “white demagogic wily politicians who... hysterically fanned anti-Chinese sentiment.”³

Although I agree that these factors contributed to Chinese exclusion in the 19th century, I propose that a more decisive factor preventing Chinese immigrants arriving between 1848-1882 from integrating into American society is to be found in the lack of Chinese immigrant women. This generation-long lack of women caused the Chinese immigrants to remain sojourners who usually continued to focus on life in China, where their wives and children remained. This, in turn, led to the rise of vices associated with the Chinese immigrants such as prostitution, gambling, and opium smoking. These licentious behaviors ravaged the immigrant communities. Furthermore, they also ensured that no second-generation Chinese cohort would exist in the United States in the 19th century to defend Chinese ethnic interests and provide an example that Chinese could successfully integrate into American society. In time, the widespread association of those vices with the Chinese led them to be the first deliberately excluded nationality in United States history.

The Lack of Chinese Immigrant Women

From 1848 to 1882, the year of the passage of the Chinese Exclusion Act, the vast majority of Chinese immigrants in the United States were men. Unlike most other immigrant groups in the 19th century, the male dominance of Chinese immigration was continuous for generations, and no sizable second-generation Chinese cohort ever emerged in the 19th century. This occurred because few Chinese women moved to the United States and Chinese men were not permitted to marry white women. While over half of the population of male Chinese immigrants were married before moving to the United States,⁴ very few brought their wives and families with them. They were generally reluctant to do so because in Chinese patriarchal family society it was culturally unacceptable for “decent”, married women to travel abroad.

The dangerously racist nature of American society in general and Gold Rush Californian society in particular, as well as the fact that the high cost of transporting their families and maintaining them in the United States was unaffordable for all but the wealthiest Chinese.⁵ Thus, most Chinese immigrants maintained a split household, in which the men lived and worked in the United States and the women remained in China to raise children. Dr. Lucie Hirata, a scholar focusing on Chinese American immigration, cogently argues why “decent” Chinese women did not immigrate to the United States:

“The patriarchal system required the preservation of the relationship between the men who went abroad to seek work and the families they left behind. It was

³ A more detailed explanation of the various hypotheses can be found in Lai To Lee, ed., *Early Chinese Immigrant Societies: Case Studies from North America and British Southeast Asia*, Asian Studies Series (Hong Kong) (Singapore: Heinemann Asia, 1988), 18.

⁴ Judy Yung, Gordon H. Chang, and H. Mark Lai, eds., *Chinese American Voices: From the Gold Rush to the Present* (Berkeley: University of California Press, 2006), 2.

⁵ Lucie Cheng Hirata, “Free, Indentured, Enslaved: Chinese Prostitutes in Nineteenth-Century America,” *Women in Latin America* Vol. 5., no. No. 1 (Autumn 1979): 3–29.

common practice for the emigrant male to marry before his departure to ensure that a wife be at home to fulfill his filial duties ... The relatives were charged with the duty of keeping the women of the emigrants ‘pure’, and in return the emigrants were obliged to send their earnings to support their families. Chinese prostitution was an integral part of that arrangement. While [the patriarchal society] prohibited the emigration of ‘decent’ women, it did not forbid the emigration of prostitutes. The emigration of Chinese prostitutes helped to stabilize and preserve the family because Chinese emigrant males could thereby avoid liaisons which might lead to permanent relationships with foreign women.”⁶

Table 1. Sex Ratio of Chinese and Total Population in California, 1850-1900

	Chinese*	Total*
1850	39,450**	1,228.6
1860	1,858.1	255.1
1870	1,172.3	165.4
1880	1,832.4	149.3
1890	2,245.4	137.6
1900	1,223.9	123.5

Sources. Ratio for Chinese from 1860 to 1900 based on California Department of Industrial Relations, Californians of Japanese, Chinese, and Filipino Ancestry (San Francisco: State Office), 1995.
 * Males per 100 females
 ** There were only two Chinese women in California in 1850

As can be seen in Table 1,⁷ the lack of female Chinese immigration caused the Chinese immigrant community’s sex ratio, the number of men per a hundred women, to remain extraordinarily imbalanced for a far longer period of time than for the population of California as a whole. By 1870, the sex ratio for the population of California as a whole was relatively balanced, at 165 men per 100 women. In contrast, the Chinese immigrant population remained grossly imbalanced, with 1,172 Chinese men per 100 Chinese women. Therefore, most Chinese immigrants were not able to establish families in the United States and produce a sizable second generation of immigrants that could successfully integrate into American society.

Another important factor contributing to the lack of women in Chinese immigrant communities in the United States was America’s racial hostility towards Chinese immigrants. The authorities sought to make the lack of Chinese women permanent so as to weaken and ultimately extinguish the Chinese population in the United States. In 1875, the United States Congress passed the Page Act, designed to prohibit the entry of “undesirable” immigrants, which included Chinese prostitutes. According to Judge Lorenzo Sawyer of the U.S. Circuit Court, “if [the Chinese] don’t bring their women here, [they] would never multiply ... When the Chinaman comes here and [doesn’t] bring his wife here, sooner or later he dies like a worn-out steam engine; he is simply a machine, and don’t leave two, or

⁶ Ibid., 6–7.

⁷ Ibid.

three or a half-dozen children to fill his place.”⁸ Far from being unaware of the impact the lack of women had on the Chinese immigrant community, prominent Americans deliberately sought to use the community’s lack of women to weaken and separate the Chinese community from mainstream white America.

The 1875 Page Act greatly increased scrutiny of Chinese women in San Francisco and Hong Kong in a complex and convoluted process. Before Chinese women were allowed to immigrate to the United States, the American consulate in Hong Kong first had to determine whether they were ‘virtuous’ or likely to be prostitutes.⁹ If the consul’s office found a woman to be of good moral character, she would have to undergo further examination by the harbor master; only if both were convinced would she then be allowed to purchase a ticket. Her certificate issued by the consul’s office and the harbor master would be mailed to San Francisco. The women would only be allowed to enter once the proper documentation had arrived and was cleared.

The system was naturally prone to corruption. In 1879, it was discovered that the US consul’s office had received \$10-\$15 from Chinese smugglers for every woman the consulate approved.¹⁰ The complexity of the system meant that only wealthy Chinese men were able to afford the luxury of having their wives join them in the United States.

According to the testimony of a poor Chinese immigrant in the 1870s, “it is impossible to get a Chinese woman out here [to the United States] unless one goes to China and marries her there, and then he must collect affidavits to prove that she is really his wife. That is in the case of a wealthy merchant. For a poor laundryman can’t bring his wife here under any circumstances.”¹¹

While Chinese did marry outside their ethnic group when they could, such as in New York where Chinese immigrants married Irish, German, or Italian poor working girls,¹² this was not possible in California, where the majority of Chinese immigrants lived, as anti-miscegenation laws prevented interracial marriage. These laws remained on the books until they were deemed unconstitutional by the Californian Supreme Court in 1948 in the case of *Perez v. Sharp*.¹³ Consequently, the vast majority of Chinese men in the United States were unable to marry or to bring their families to the United States. The effects of this lack of women and a second-generation cohort would decisively shape Chinese immigrant behavior, resulting in split households that contributed to the rise of prostitution, drug addiction, and gambling in Chinese communities and effectively led to passage of the Chinese Exclusion Act of 1882.

Chinese Immigrants Focus on Life at Home

The first consequence of the lack of women was that for a generation the Chinese immigrants continually remained focused on returning to China, viewing their lives in the

⁸ George Anthony Pepper, *If They Don’t Bring Their Women Here: Chinese Female Immigration before Exclusion* (University of Illinois Press, 1999), 108–9, <http://www.press.uillinois.edu/books/catalog/85tcb3hp9780252024696.html>.

⁹ Hirata, “Free, Indentured, Enslaved: Chinese Prostitutes in Nineteenth-Century America,” 10.

¹⁰ *Ibid.*, 11.

¹¹ Ronald Takaki, *Strangers from a Different Shore* (Ward & Balkin Agency, Inc., 2012), 125.

¹² *Ibid.*, 126.

¹³ *Andrea D. Perez et al., Petitioners, v. W.G. Sharp, as County Clerk, etc., Respondent* (California Supreme Court 1948).

United States as merely temporary. As previously mentioned, most Chinese immigrants came from southern China, which was dominated by powerful, patriarchal, extended kinship groupings called clans. For the most part, the clans were the main source of authority. The ruling philosophy of the clans was based on Confucian beliefs of ancestor worship that emphasized family relationships as the core of one's identity and as basic unit of Chinese society. It was necessary to give daily sacrifice and pay respect to the spirits of the ancestors in the ancestral temple, with an altar containing "wooden spirit tablets, each of them representing a dead ancestor." Along with an "ever-burning lamp", offerings of incense and candles, and the religious ceremonies made to them, deceased ancestors appeared to be part of the lives of the living.¹⁴ Every clan had an ancestral temple, which was usually the most prominent building located in the center of the village.¹⁵

Traditional Chinese belief was that the souls of the dead could not rest until they were returned to the ancestral shrine, so that surviving relatives could pay proper respects to the dead. If one died outside of his or her home town, it was necessary for his or her bones to be returned to the ancestral shrine so that future generations could keep the spirits from intruding on the lives of the living. Thus, one of the first tasks that a new Chinese immigrant would complete upon arrival in his new country was to arrange for the return of his bones to the family shrine should he die in the United States.¹⁶ Many immigrants formed special organizations that were responsible for returning the remains of the deceased immigrants to China. The excerpt below provides an account of associates from one such organization and depicts the basic function of their association:

"Who knows how many of [the Chinese] died with their ambitions unattained, their dreams unfulfilled. Instead, their spirits could not return to their homeland since their bones were buried in foreign soil. They could only gaze longingly toward home, and their anguish deepened with each passing of Ching Ming. One cannot but feel regret at such stories. Therefore, in 1858...those of us formerly from Panyu assembled our fellow expatriates and proposed the establishment of the Cheung Hau Association. We drew up a special rule: successful returnees to China would each donate ten American dollars... In this way we collected funds and hired workers to locate and exhume the remains of our fellow countrymen to be sent back to their native village so that, properly interred, they could forever enjoy the offerings due to the dead."¹⁷

The main ancestral shrine was always located where the majority of the clan members resided. Because Chinese immigrants did not bring their families or clans with them, the ancestral shrines remained in China. If their families had been brought to the United States, the shrines would have been brought there as well, and the center of Chinese family and spiritual life would have been in the United States. An example of this occurred when Chinese clans migrated to Taiwan in the 18th century. At first, immigrants only sent remittances home to the main ancestral clan, but in time, the main ancestral shrine was

¹⁴ Ching Kun Yang, *Religion in Chinese Society: A Study of Contemporary Social Functions of Religion and Some of Their Historical Factors* (Prospect Heights, IL: Waveland Pr Inc, 1991), 29.

¹⁵ *Ibid.*, 40–41.

¹⁶ Yung, Chang, and Lai, *Chinese American Voices*, 26.

¹⁷ *Ibid.*

moved to Taiwan.¹⁸ Had this occurred in the United States, the Chinese immigrants would no longer have been sojourners seeking to return, but rather would have been dedicated to creating a life in their new home. As their children assimilated into American society, Chinese immigrants would have been no different from the Irish, German, and Italian immigrants that successfully integrated into American society in the 19th century.

The Rise of Vices among the Chinese

The general lack of women and families in Chinatowns, as Chinese communities began to be called, led to a number of vices among the Chinese, such as prostitution, gambling, and increased opium smoking, that encouraged Americans to view the Chinese negatively and seek to exclude them from their own communities. These practices greatly contributed to the negative public opinion of white America toward Chinese immigrants. The second consequence of the lack of families and marriageable Chinese women meant that prostitutes were in high demand. The majority of the prostitutes were young women who had been kidnapped, tricked, or outright purchased from poor families in southern China who could not afford to feed their families.¹⁹ The United States Census of 1870 labeled 61% of all Chinese immigrant women as prostitutes.²⁰ Most of the Chinese prostitutes were in a condition of debt peonage; the contract presented below, signed by a prostitute Xin Jin, is an example of one such agreement:

“The contractee Xin Jin became indebted to her master/mistress for food and passage from China to San Francisco. Since she is without funds, she will voluntarily work as a prostitute at Tan Fu’s place for four and one-half years for an advance of 1,205 yuan (U.S. \$525) to pay this debt. There shall be no interest on the money and Xin Jin shall receive no wages. At the expiration of the contract, Xin Jin shall be free to do as she pleases...If she has the four loathsome diseases she shall be returned within 100 days; beyond that time the procurer has no responsibility. Menstruation disorder is limited to one month’s rest only. If Xin Jin becomes sick at any time for more than 15 days she shall work one month extra; if she becomes pregnant, she shall work one year extra. Should Xin Jin run away before her term is out, she shall pay whatever expense is incurred in finding and returning her to the brothel.”²¹

Despite the fact that the women’s contracts usually ended after four or five years, Chinese prostitutes lived in virtual slavery. Just like African-American slaves, they were brought to the United States on ships and sold to slave-owners. The following passage describes how brothels obtained prostitutes:

“In Canton, during 1849 to 1860, many Chinese girls could be purchased for five dollars apiece. Sometimes sons and daughters were not sold but kidnapped from better homes, taken from the families of gamblers whose losses frequently compelled them to surrender their children to pay their debts, or mortgaged to secure a loan of money and because of nonpayment were taken by the

¹⁸ George A. Devos and Takao Sofue, *Religion and the Family in East Asia* (San Francisco, CA: University of California Press, 1986), 177.

¹⁹ Yung, Chang, and Lai, *Chinese American Voices*, 15.

²⁰ Takaki, *Strangers from a Different Shore*, 121.

²¹ *Ibid.*, 122.

creditor²²...On Dupont Street in San Francisco... known as the Queen's Room... Chinese slave girls were brought from the ships and exposed for examination to the prospective buyers, who rated them according to their various standards of physical beauty. Sometimes the girls were treated in a horrible manner which was supposed to render them more valuable for the purpose for which they were purchased. This maltreatment of the slave girls was suggestive of the cunningness of the Chinese slave dealer...to add to their sale value."²³

Despite claiming to be only temporary, the contract was actually far more insidious. The temporary nature of the contract discouraged prostitutes from fleeing brothels as there was hope for a free life following its termination. However, very few women survived much longer than the five years specified in the contracts; most died of disease or malnutrition prior to their release. Thus, the Chinese brothels were able to fully exploit the Chinese prostitutes for the duration of their working lives, while giving the Chinese prostitutes an illusion of a life free from their debts.

For Americans, who had just fought a civil war to end slavery in the United States, the mistreatment and prominence of Chinese prostitutes was one of the strongest arguments to exclude the Chinese. A review of San Francisco newspapers between 1872 and 1882 shows no single favorable entry about Chinese American women.²⁴ Beginning in the 1870s, the most powerful and influential group crusading against Chinese prostitution were white, middle-class women, seeking to "clean up" California and end its reputation as a lawless state in the Wild West. According to Dr. Lucie Hirata:

"With the advent of the Victorian ladies from the East Coast concerned with the preservation of the family, [their] Puritan morality led them to crusade against prostitution in general and Chinese prostitution in particular. In 1873, the interests of the Victorian ladies in San Francisco found expression in the Women's Occidental Board. Reportedly alarmed by the immorality of the traffic ... and the sinfulness of the prostitute's sexual activity, Margaret Culbertson and her successor Ms. Cameron set out to rescue the Chinese slaves [alongside] clergymen."²⁵

For many white Americans, Chinese prostitution was a moral outrage. According to the official investigation of the impact of Chinese immigration on California by the 1885 Canadian Royal High Commission, "there can be no doubt that one of the causes of the strong feelings [of Americans] against the Chinese is that their immigration consists of unmarried men and prostitutes and it is said that the Chinese prostitutes are [most] injurious to the community."²⁶

²² Alexander McLeod, *Pigtails and Gold Dust* (Caldwell, Idaho: Caxton Printers, 1947), 174.

²³ *Ibid.*, 178.

²⁴ Peffer, *If They Don't Bring Their Women Here: Chinese Female Immigration before Exclusion*, 76.

²⁵ Hirata, "Free, Indentured, Enslaved: Chinese Prostitutes in Nineteenth-Century America," 28.

²⁶ J. A. Chapleau, Canadian Libraries, and Canadian Publications From 2013-, eds., *Report of the Royal Commission on Chinese Immigration: Report and Evidence* (S.l.: Ottawa: Printed by order of the Commission, 1885), LXXVIII.

Prostitution was not the only practice that made Chinese immigrants notorious; gambling and opium smoking were widespread as well. These behaviors were not limited to the mostly male Chinese immigrants but were also prevalent in China at the time, but the lack of wives and families in the United States, as well as the stressful lives of most Chinese immigrants working in gold mines or on railroads, made gambling and opium smoking especially popular in Chinese communities. The following account illustrates how Americans viewed Chinese gambling:

“Most gambling dens in California were operated behind strong doors, each guarded by a swarthy Chinaman who communicated any suspicious movements on the outside to the players in the den. These lynx-eyed sentinels on the least suspicion pulled a hidden cord, and in a twinkle, one or more heavy plank doors with sturdy bars closed before the invaders. Before the police could force their way into the dens, the occupants had disappeared through openings in the floor and walls.”²⁷

Despite the best efforts of the Californian police, gambling continued to be an essential part of Chinese life in Chinatowns. Traditionally, the Chinese immigrants were famous for their frugality, good work ethic, and business sense. According to Mr. Crocker, one of the five proprietors of the Central Pacific Railroad, “today if I had a big job of work ... I should take Chinese labor to do it with, because of its great reliability, steadiness and aptitude and capacity for hard labor...The Chinese men showed power of endurance equal to the best white men.”²⁸ Yet, when gambling, their famous business sense seemed to disappear as they risked a large portion of their meager earnings. Opium smoking became popular in China when the ruling Qing Dynasty lost the Opium Wars to the United Kingdom and was required to permit the importation of opium. When the Chinese immigrated to the United States, they brought the practice of opium smoking with them. According to one account of the Chinese opium dens:

“In San Francisco in the 1870s, ‘nearly every house, store, and shop in Chinatown was provided with the drug, together with the implements for using the article. While this was practiced among the Chinese alone, where forty out of every hundred Chinese smoked the drug, no particular attention was given to the subject. But soon eight places [were] started, furnished with opium pipes, and beds for sleeping off the fumes, run by Chinamen, and patronized by white men and women, who visited the dens at all hours of the night and day. The habit and its deadly results became so extensive as to attract the attention of the municipal authorities, and an ordinance was passed which had some effect in abolishing such places for a while. Frequently police raided these dens to enforce the law, and it was not unusual to find white women and Chinamen side by side under the effects of the drug...nothing could compare with the shock he received when he made his first trip to a well-patronized Chinese opium den. It was the most disgusting characteristic of Chinese life.”²⁹

²⁷ McLeod, *Pigtails and Gold Dust*, 164.

²⁸ Chapleau, Canadian Libraries, and Canadian Publications From 2013-, *Report of the Royal Commission on Chinese Immigration*, XVII.

²⁹ McLeod, *Pigtails and Gold Dust*, 155–56.

Unlike the Irish or German immigrants or the native-born American population, who frequently consumed alcohol, the Chinese preferred opium. However, while alcohol tended to stimulate anger, tavern brawls, and the beating of women and children by alcoholic fathers, opium tended to make its users lethargic; opium dens were notorious for their silence and tomb-like atmosphere, whereas taverns tended to be exceptionally rowdy. American prejudice towards opium can thus be thought of largely as racial and cultural. The lack of women and families in the Chinese immigrant community meant that Chinese men were frequently associated with prostitution, gambling, and opium smoking. Along with an underlying racial prejudice, these vices strengthened the argument of Americans seeking to exclude Chinese from American society.

No Chinese Second-Generation Cohort

The third major consequence of a lack of women and families meant that no sizeable Chinese second generation, the American-born children of Chinese immigrants, emerged in the United States in the 19th century. By 1882, the year of the passage of the Chinese Exclusion Act, there were perhaps a thousand children in an immigrant population of over 100,000.³⁰ The lack of a second generation had a wide-ranging detrimental effect on the behavior of the Chinese immigrant community. According to Alejandro Portes and Ruben G. Rumbaut, two scholars of immigration to the United States and authors of *Immigrant America*:

“To a large extent nativist fears and the feverish pitch reached by campaigns based on them are due to the peculiar position of immigrant communities that are ‘in the society, but not yet of it.’ Their very foreignness provides fertile ground for all sorts of speculations about their traits and intentions. At the same time, immigrants often lack sufficient knowledge of the new language and culture to realize what is happening and explain themselves effectively. For the most part, the first foreign-born generation lacks ‘voice.’ It is on this enforced passivity that the nativist fears of many... have flourished.”³¹

The first-generation Chinese immigrants had limited English language proficiency and did not have enough knowledge of the laws and customs to be able to defend themselves effectively in American society. Coupled with their non-European and thus foreign culture and their propensity for prostitution, gambling, and opium smoking, they were an easy target for politicians and activist groups seeking to gain support. Furthermore, the Chinese first-generation cohort continued to focus on life in China. They generally remained preoccupied with life in China and in their home villages, where their families remained. If they did care about politics, they focused on Chinese rather than American politics, which they were not able to participate in as they were not American citizens. Furthermore, the Chinese immigrants were also unfamiliar with the workings of American government since they were born into a patriarchal, clan-structured system rather than a democratic one.

While the first generation of immigrants usually retain most of the culture and habits of their home countries, their children, growing up in America and having only limited contact with their parents’ home country, usually integrate into American society and become Americans. Unlike their parents, the second generation tends to be fluent in

³⁰ *Ibid.*, 155.

³¹ Yung, Chang, and Lai, *Chinese American Voices*, 20–21.

English, focus on life in the country they have been living in, America, and have American citizenship, which grants them the ability to participate in all aspects of American life. As a generation, they are more attuned to American culture, successful in defending themselves and their ethnic groups, and able to reaffirm ethnic identities that were previously easy targets for nativists.³² For most immigrant families, it is only the second generation that begins to acculturate and assimilate into American society.

Unlike other immigrant groups coming into the United States, the lack of women in the Chinese immigration cohort meant that no second-generation immigrant group was formed. This meant that most Chinese immigrants would remain sojourners rather than establish families in the United States. This situation had several consequences that eventually led to the effective prohibition of Chinese immigration to the United States. First, without a second generation, there was effectively no assimilation of Chinese into American society. The vast majority of Chinese in the United States were born in China, learnt Chinese, spoke broken English, and, because their culture was primarily Chinese, and they immigrated to the US as adults, they mostly failed to assimilate into American society. While this was the case for most immigrant groups, the Chinese immigrant group continually replenished itself with new arrivals from China. Thus, from 1848-1882, the Chinese immigrant community remained dominated by first-generation Chinese. It appeared that unlike other European immigrant groups, the Chinese were simply unable to integrate into American society.

Second, as the first-generation Chinese immigrants did not have children in the United States, they did not have the impetus to position themselves in a way that would bring the greatest advantages possible to their children. Had the first-generation immigrants had children, they would have encouraged their children to learn English and adapt to American culture so that they would be able to enjoy a better life in the United States. Without children in the United States, the first-generation Chinese immigrant cohort would remain focused on life in China and would focus on earning money for remittances, rather than seeking to adapt to American society.

The consequences of the lack of a second-generation of Chinese immigrants were that there was no voting coalition of second-generation immigrants that could mobilize against the Chinese Exclusion Act and defend other Chinese interests. Unlike the Irish, Italians, or Jewish immigrants in the 19th century, Chinese could not vote as a bloc and thus gain political power. Furthermore, the first-generation Chinese immigrant cohort remained an easy target for nativists and politicians seeking to gain political office. As a result, the first-generation Chinese immigrant communities presented to white America the image of a secluded society that was above American law and showed no signs of integrating into the American society, giving the American nativists a powerful argument for Chinese exclusion before the supposed millions of Chinese immigrants could overwhelm white America.

Conclusion

The lack of Chinese immigrant women in America ensured that the Chinese immigrant community would be dominated by men until the 1960s. This caused Chinese male immigrants in California to maintain strong relationships and connections with their home country, and most of them believed that they would eventually return to China, either alive or to be buried in their ancestral land. In addition, this lack of women led to the rise of behaviors considered sinful, such as prostitution, gambling, and opium smoking, which fanned anti-Chinese sentiment and were used as ammunition by white supremacists.

³² *Ibid.*, 119–20.

Furthermore, as few Chinese families formed, no sizable second-generation Chinese community emerged that could begin to integrate into American society and defend Chinese interests. These factors ultimately ensured that the Chinese would remain marginalized in the United States from the 1880s until the 1960s, when immigration laws were once again relaxed.

Further research comparing the immigration patterns of Chinese women with those of Italian, Irish, Japanese, Filipino, and Jewish immigrant women in the 19th century would be salutary. What factors caused other immigrant women in the 19th century to immigrate to the United States, whereas Chinese women did not do so? Why were certain immigrant groups excluded earlier than others? A greater comparative perspective would not only enrich our understanding of Chinese migration patterns to the US at the time, but also provide a deeper historical perspective.

Bibliography

- Chapleau, J.A., Canadian Libraries, and Canadian Publications from 2013-, eds., *Report of the Royal Commission on Chinese Immigration: Report and Evidence* (S.I.: Ottawa: Printed by Order of the Commission, 1885), LXXVIII.
- Devos, George A., and Sofue, Takao, *Religion and the Family in East Asia* (San Francisco, CA: University of California Press, 1986)
- Farkas, Lani Ah Tye, *Bury My Bones in America: The Saga of a Chinese Family in California, 1852-1996: From San Francisco to the Sierra Gold Mines*, 1st ed. (Nevada City, Calif: Carl Mautz Publishing, 1998)
- Hirata, Lucie Cheng, "Free, Indentured, Enslaved: Chinese Prostitutes in Nineteenth-Century America," *Women in Latin America*, Vol. 5., No. 1 (Autumn 1979)
- Lee, Lai To, ed., *Early Chinese Immigrant Societies: Case Studies from North America and British Southeast Asia*, Asian Studies Series (Hong Kong) (Singapore: Heinemann Asia, 1988)
- McLeod, Alexander McLeod, *Pigtails and Gold Dust* (Caldwell, Idaho: Caxton Printers, 1947)
- Peffer, George Anthony, *If They Don't Bring Their Women Here: Chinese Female Immigration before Exclusion* (University of Illinois Press, 1999)
<http://www.press.uillinois.edu/books/catalog/85tcb3hp9780252024696.html>.
- Perez, Andrea D. Perez, et al., Petitioners, v. W.G. Sharp, as County Clerk, etc., Respondent (California Supreme Court, 1948)
- Takati, Ronald, *Strangers from a Different Shore* (Ward & Balkin Agency, Inc., 2012)
- Yang, Ching Kun, *Religion in Chinese Society: A Study of Contemporary Social Functions of Religion and Some of Their Historical Factors* (Prospect Heights, IL: Waveland Press, Inc, 1991)
- Yung, Judy; Chang, Gordon H., and Lai, H. Mark, eds., *Chinese American Voices: From the Gold Rush to the Present* (Berkeley: University of California Press, 2006)



Identification of Reliable Predictor of Primary Spontaneous Pneumothorax Recurrence Risk

Ruibing Xu

Author Background: Ruibing Xu grew up in China and currently attends Shenzhen Foreign Languages School in Shenzhen, China. Her Pioneer seminar topic was in the field of biology and titled "Integrative Physiology."

Abstract

Primary spontaneous pneumothorax (PSP), defined as the accumulation of air or gas in the pleural cavity without underlying lung disease, is a common clinical problem with high recurrence risk. To optimize treatment, patients should be stratified to different management options according to recurrence risk. However, there are controversies and ambiguities about the established stratification system.

The aim of this study is to identify a reliable predictor of PSP recurrence risk. Since previous studies revealed that female sex, greater height, and lower weight are associated with PSP recurrence, we selected sex hormones estrogen, androgen, and progesterone, as well as proteins estrogen receptor (ER)-alpha, estrogen receptor (ER)-beta, aromatase, elastase, and α_1 -antitrypsin as candidates, each of which has an important role in relation to sex, height regulation, or weight regulation. We planned to determine the correlation between PSP recurrence rate and the expression level of the selected hormones and proteins, measured by mass spectrometry assays and immunoassays respectively in first episode PSP patients.

We hypothesize that the abnormal expression of estrogen, androgen, progesterone, ER-alpha, ER-beta, aromatase, elastase, and α_1 -antitrypsin have strong association with higher PSP recurrence risk.

If strong correlation is confirmed, these molecules can be used to provide an assessment of PSP recurrence risk, outcompeting the ambiguous approaches currently being used. In addition, the result of this study deepens the understanding of the effects of particular molecules on PSP recurrence, shedding light on the pathogenesis of PSP.

Significance

Generally, primary spontaneous pneumothorax is not as serious as other types of pneumothorax. For instance, compared with secondary spontaneous pneumothorax, primary spontaneous pneumothorax is associated with a lower morbidity and mortality.^{1,2} However, recurrence is frequently observed despite management (20%-60%,³ varying between different management methods). International guidelines suggest a role for conservative management of clinically stable patients with primary spontaneous pneumothorax,¹ while invasive treatments should be offered promptly to those at high risk of recurrence.⁴ Currently, patients are stratified to treatment options depending on the combination of symptoms and an assessment of the size of the pneumothorax. Since many believe that PSP results from spontaneous rupture of a subpleural bleb or bulla,⁵ CT scanning is used as a

routine procedure for patients with PSP. However, prior studies revealed that the existence and number of bullae/blebs on CT were not associated with the risk of recurrence.⁶ Therefore, a more reliable predictor for risk of recurrence would be of great value in early stratification of patients to the appropriate management strategy.

In this study, we will test the correlation between the expression level of estrogen, androgen, progesterone, estrogen receptor (ER)-alpha, estrogen receptor (ER)-beta, aromatase, elastase, and α_1 -antitrypsin and PSP recurrence rate in order to identify a reliable predictor of PSP recurrence risk. On its own, each biomarker has potential advantage in sensitivity and specificity to PSP recurrence over phenotypes such as the presence of bullae/blebs, because while the biomarker could play a significant role in the pathways closely involved with the disease, phenotypes are more likely to be influenced by other factors irrelevant to the disease. Moreover, as a single marker is rarely ideal for prediction, an individual patient's optimal prediction, in most cases, comes from a multivariable model.^{7,8} The identification of new predictors, combined with existing predictors, will help build a multivariable model to better predict PSP recurrence risk.

If successful, the identified biomarker/ predictor will aid the development of a more sophisticated risk stratification system at the first PSP episode based on more reliable predictors or a multivariable model, which could identify patients who will benefit from intervention to prevent recurrence at the first presentation, rather than, as has historically been the case, simply waiting for a recurrence to occur. In addition, by revealing the relationship between the expression level of these molecules and the recurrence of PSP, our study allows us to have a more comprehensive view of the underlying mechanism of PSP as well as associated diseases such as secondary spontaneous pneumothorax, which is essential to improving the diagnosis and management of patients.

Investigator

As a friend of a PSP patient, I am strongly interested in the pathophysiology of PSP and willing to devote myself to improving the stratification system of PSP in order to optimize management for the good of the patients. Knowledge from the AP Biology course, AP Chemistry course, and the Pioneer research course on Integrative Physiology that I took laid a solid foundation for me to understand the molecules I will work with in this project—hormones and proteins—as well as their effects on human physiological function so as to come up with a rigorous project framework. Additionally, from my previous experience in iGEM, a synthetic biology competition—leading my team in designing a test paper to detect HCV based on RNA detection as the director of project design—I gained research skills such as converting information extracted from literature into ideas to build my own project and communication skills such as presenting the project to potential sponsors. These skills will help me carry out a project successfully and scientifically.

Innovation

Previous studies determined that the recurrence of PSP was significantly more common in taller patients, thinner patients, and female patients. However, few studies went further to provide an explicit explanation of the mechanism underlying the relationship between height, weight, sex, and PSP recurrence. Factors related to height and weight such as body mass index⁹ ($\text{weight}/\text{height}^2$) have been proposed as the predictor of PSP recurrence, but the results of some other experiments revealed that PSP recurrence was not related to the body mass index of the patient.³ Therefore, further and more detailed studies on the relationship between height, weight, sex, and PSP recurrence rate are needed. Our group will attempt to look into the molecular mechanism underlying the relationship, which

remains understudied, by identifying one or several biomarkers whose expression levels are strongly correlated with PSP recurrence rate. Compared with height, weight, other body features, or female sex, whose relationship with PSP recurrence is indirect, molecules play a more direct role in the causation and incidence of PSP recurrence, and thus are better predictors. If the expression levels of our selected hormones and proteins are confirmed to be significantly correlated with PSP recurrence rate, we are able to pioneer in using biomarkers as predictors for risk of PSP recurrence and to shed light on the molecular mechanism of PSP recurrence. We hope that the results of this study will lead to a better understanding of the pathogenesis of PSP and will improve its management.

Approach

Increased height and decreased weight in both sexes are positively correlated with higher recurrence of PSP. When multiple risk factor analysis was performed, the most important characteristic associated with recurrence was increased height ($P < 0.0045$) followed by decreased weight ($P < 0.0051$).¹⁰ While males have a higher incidence rate, females have a higher recurrence rate. The reported incidence is 18-28/100000 cases per annum for men and 1.2-6/100 000 for women.¹¹ However, the recurrence rate can reach 71.4% for women, while only 46.2% for men.³

The overall goal of this study is to identify a reliable predictor of PSP recurrence risk. Our primary targets were molecules that have close association with sex, height regulation, and weight regulation of three different groups: sex steroids that affect growth, hormone receptor proteins corresponding to the sex steroids of interest, and enzymes which are regulated by sex steroids and influence height and weight by regulating metabolism. These three groups of molecules have significant influence on physiology through biochemical interactions. Sex steroids regulate cell functions by binding to specific proteins, resulting in the activation of a signal transduction pathway. Enzymes also play an important role in signal transduction and cell regulation, and their activities can be regulated by sex steroids. The close relationship between PSP recurrence and female sex, height, and weight indicates the possible existence of overlaps in their biochemical pathways. One example of such overlaps is the AMP-activated protein kinase (AMPK) pathway which plays a role in both the development of PSP and weight regulation. The suppression of AMPK activity is associated with decreased food intake and weight loss.¹² Meanwhile, tumor suppressor gene *Folliculin (FLCN)* deficiency, a cause of PSP, is also related to decreased activity of the AMPK pathway.¹³ In addition, the MAPK pathway, which is involved in PSP,¹⁴ and the Wnt pathway, which is involved in weight regulation,¹⁵ are activated by the same protein kinase, Casein kinase 2,^{16,17} so these two pathways probably share the same or similar activation mechanisms and partially overlap. The abnormal expression of other molecules involved in the regulation of sex, weight, and height might also lead to or be influenced by PSP recurrence. Thus, it is likely that abnormal expression of sex steroids, hormone receptor proteins, or enzymes that are found predominately in females, taller people, or thinner people can be detected in patients with higher PSP recurrence risk and therefore can serve as the predictor we are looking for.

We selected sex steroids estrogen, androgen, and progesterone, hormone receptors estrogen receptor (ER)-alpha, and estrogen receptor (ER)-beta, as well as protein enzymes aromatase, elastase, and a-antitrypsin as our candidates. Recent studies revealed that impaired estrogen action or synthesis leads to tall stature, especially in males. Estrogen actions are further complicated by the presence of two distinct but related receptors, ER-alpha and beta.¹⁸ Androgens also contribute to the pubertal growth spurt in part through

conversion to estrogens by the enzyme aromatase, and in part through direct stimulation of growth.¹⁹

Other studies concluded that elastic fiber degradation due to an imbalance between elastase and α_1 -antitrypsin, an anti-elastase, was thought to be the cause of emphysema, which is frequently observed in PSP patients.²⁰ The abnormal expression level of elastase and α_1 -antitrypsin is associated with Pancreatic Exocrine Insufficiency (PEI), which might lead to low weight.²¹ In addition, progesterone is found to have an inhibition effect on elastase.

Our specific aim is to determine the correlation between PSP recurrence rate and the expression level of the selected sex steroids and proteins. We hypothesize that higher PSP recurrence risk is associated with the under expression of estrogen, ER-alpha, ER-beta, androgen, and aromatase, each of which could give rise to taller stature; the under expression of progesterone, which leads to higher elastase level; and/or the under expression of α_1 -antitrypsin, and the over expression of elastase, each of which could be the cause of emphysema and lower weight.

Since we seek to identify a predictor of PSP recurrence risk for PSP patients, the optimal subjects are clinical samples diagnosed with PSP. In addition, PSP recurrence rate cannot be determined using cell culture, and it is unnecessary to conduct preclinical studies since no drug or new treatment will be used. We will recruit participants from patients admitted with their first episode of PSP to the Respiratory Diseases Department of the First Affiliated Hospital of Sun Yat-sen University of Medical Sciences, Guangzhou, China, for its convenient access to PSP patients and research facilities, and for the opportunity to collaborate with a leading figure in PSP recurrence, Dr. Yubiao Guo (see Environment below). The following data will be collected: (1) age and sex, (2) height and weight, (3) estrogen expression level, (4) androgen concentration, (5) progesterone concentration, (6) ER-alpha expression level, (7) ER-beta expression level, (8) aromatase expression level, (9) elastase expression level, and (10) α_1 -antitrypsin expression level. Sex steroids and proteins will be extracted from the serum sample. Sex steroids concentration will be measured by mass spectrometry assays,²² while expression levels of proteins will be measured by immunoassays.²³

A standard chest-tube drainage will be used as management for all patients. After patients are discharged from hospital, follow-up data will be collected over a 4-year period regarding whether recurrence, defined as a novel pneumothorax episode occurring in more than a 1-month period after the end of the management, takes place. Because this study focuses on primary spontaneous pneumothorax, the data of patients that develop into secondary, iatrogenic, and traumatic pneumothorax during the study will be excluded.

Afterwards, the correlation between each variable and PSP recurrence rate will be analyzed with statistical methods. If there is a significant correlation between an abnormal level of one or more selected biomarkers, it will qualify as a predictor of PSP recurrence risk. To further determine whether each identified biomarker has greater predictive power than established predictors, and whether each identified biomarker can contribute to developing an optimized multivariable model for clinical prediction of PSP recurrence risk, univariable analysis, multivariable analysis that includes the novel marker as well as other established markers,^{7,8} as well as decision curve analysis²⁴ will be performed.

Although the results of this study are sufficient to determine the correlation between selected molecules and PSP recurrence rate in order to identify a biomarker as predictor of PSP recurrence risk, more information (e.g. genetics) is needed to fully unveil the pathogenesis of PSP. Follow-up experiments can be conducted to further study the effect of the molecules on PSP, including using gene-knockout (e.g. CRISPR) or gene-knockdown

(e.g. RNAi) to better study the role of the molecule on animals, and studying altered physiological functions of individuals with mutation on the gene encoding or associated with the molecules. The information gained will be used to construct a model of the pathogenesis of PSP.

Ethics

The overall goal of this study is to identify a reliable predictor of PSP recurrence rate, and we aim to determine the correlation between PSP recurrence rate and the expression level of sex steroids estrogen, androgen, and progesterone, as well as proteins ER-alpha, ER-beta, aromatase, elastase, and α_1 -antitrypsin in first episode PSP patients. Subjects will be recruited from the First Affiliated Hospital of Sun Yat-sen University of Medical Sciences. Serum samples will be collected from subjects for mass spectrometry assays of sex steroids concentration and immunoassays of proteins expression level. Subjects will be treated with a standard chest-tube drainage, and follow-up data regarding PSP recurrence will be collected for a 4-year period.

Since human subjects are involved, our group has ensured that we follow good clinical practice (GCP)²⁵ and meet all the requirements of the Ethics Committee of the First Affiliated Hospital of Sun Yat-sen University of Medical Sciences in order to gain approval.

Prior to a subject's participation in the trial, the written informed consent form will be signed and personally dated by the subject or by the subject's legally acceptable representative, and by the person who conducted the informed consent discussion.²⁵

Comprehensive information about the study will be included in the informed consent form. Only subjects who submit the form may participate in the study. We will not coerce or unduly influence a subject to participate or to continue to participate in a trial.

We will safeguard the rights, safety, and well-being of all trial subjects. The subject's primary physician will be informed about the subject's participation in the trial if the subject agrees to the primary physician being informed. Protocols agreed to by the ethics committee will be strictly followed. During and following a subject's participation in the study, we will ensure that adequate medical care is provided to a subject for any adverse events. The confidentiality of records that could identify subjects will be protected, respecting privacy and confidentiality rules.

Our group will submit the protocols, written Informed Consent Forms, subject recruitment procedures, and written information to be provided to subjects to the Ethics Committee of the First Affiliated Hospital of Sun Yat-sen University of Medical Sciences.

Environment

The study will be conducted in the First Affiliated Hospital of Sun Yat-sen University of Medical Sciences, one of the leading bases for research and medical service in South China. Over the past 10 years, the hospital has won 90 awards for achievements in research and development. One of the biggest hospitals in China, the hospital serves more than 4.80 million outpatients every year. Subjects will be recruited from the Respiratory Diseases Department, and collected serum samples will be analyzed in the medical laboratory. In addition to access to large number of PSP patients from South China and even other regions of the country attracted by its reputation, the Respiratory Diseases Department has a solid foundation in working with patients to conduct clinical research on diseases closely related to PSP such as bronchial asthma²⁶ and COPD.²⁷ Since the study focuses on PSP patients, we will exclude the data of patients that develop into secondary, iatrogenic, and traumatic pneumothorax during the study. Therefore, we need to recruit a sufficient number of patients with their first episode of PSP. With access to large numbers of PSP

patients receiving treatment in the First Affiliated Hospital of Sun Yat-sen University of Medical Sciences, as well as state-of-the-art facilities such as Abbott AxSYM Immunology Analyzer to carry out mass spectrometry assays and immunoassays in the medical laboratory, the group will be able to collect and analyze data crucial to the study.

The group will collaborate with Dr. Yubiao Guo from the First Affiliated Hospital of Sun Yat-sen University, who led a study in factors related to recurrence of spontaneous pneumothorax which determined that PSP patients who were taller or weighed less were more likely to have recurrences.¹⁰ Dr. Guo's findings inspired us to study the relationship between height, weight, and PSP recurrence. In addition, the methods used in his study, such as treating every subject with a standard chest-tube drainage as management, are important bases of our experimental design. We hope that Dr. Guo's experience and perspective on PSP research will be of great value to our study.

Summary

The lack of a reliable predictor of PSP recurrence risk hinders patients from receiving optimal management. Collaborating with an authoritative hospital and a leading figure in the field of PSP, our group aims to determine the correlation between PSP recurrence rate and the expression level of sex steroids estrogen, androgen, and progesterone and proteins ER-alpha, ER-beta, aromatase, elastase, and α_1 -antitrypsin in order to achieve our goal of identifying a biomarker as a reliable predictor of PSP recurrence risk, as well as to gain insight into the pathogenesis of PSP.

The successful identification of a biomarker of PSP recurrence risk will be not only an innovation and improvement of PSP assessment, but also will shed light on and draw attention to the molecular biology aspect of PSP, which has remained relatively understudied.

Methods employed include mass spectrometry assay and immunoassay to measure the concentration of target molecules in first episode PSP patients, and recurrence-related data will be collected for a 4-year-period after patients receive a standard chest-tube drainage. Patients with under expression of estrogen, androgen, progesterone, ER-alpha, ER-beta, aromatase, and α_1 -antitrypsin, and over expression of elastase are expected to have higher PSP recurrence rates. Strong correlation between the expression level of the molecule and recurrence rate is the qualification for a reliable predictor.

It is hoped that this study will be significant in better understanding PSP and improving its management.

For more comprehensive and detailed study of PSP pathogenesis, follow-up experiments on the role of successfully identified predictors in the pathophysiology of PSP will be conducted.

Reflection

Writing a comprehensive proposal is somewhat similar to the process of constructing a house. In order to build a solid foundation for my own study, I need to have an in-depth understanding of my field of interest, including its history, development, and existing problems. The data and information of previous studies are the steel and concrete, the building blocks to solidify my building. Once I have obtained the crucial materials, the next step is to organize them wisely—backing each key point with credible resources in order to be logical, and connecting different parts of the study to each other in order to be coherent. Lastly, when presenting the study to reviewers, like presenting the design of the house, it's important to keep the general outline concise and straightforward, and lay sufficient emphasis on details and distinctive features.

As I hope to pursue a career as a researcher in the future, the art of writing grant proposals will definitely help me think critically, present ideas, and most important, conduct studies.

References

1. Macduff, A., Arnold, A., and Harvey, J. (2010) Management of spontaneous pneumothorax: British Thoracic Society pleural disease guideline 2010. *Thorax* 65, ii18–ii31
2. Onuki, T., Ueda, S., Yamaoka, M., Sekiya, Y., Yamada, H., Kawakami, N., Araki, Y., Wakai, Y., Saito, K., Inagaki, M., and Matsumiya, N. (2017) Primary and Secondary Spontaneous Pneumothorax: Prevalence, Clinical Features, and In-Hospital Mortality. *Canadian Respiratory Journal* 2017, 1–8
3. Sadikot, R. T., Greene, T., Meadows, K., and Arnold, A. G. (1997) Recurrence of primary spontaneous pneumothorax. *Thorax*
4. Bintcliffe, O. J., Hallifax, R. J., Edey, A., Feller-Kopman, D., Lee, Y. C. G., Marquette, C. H., Tschopp, J. M., West, D., Rahman, N. M., and Maskell, N. A. (2016) Spontaneous pneumothorax: Time to rethink management? *The Lancet Respiratory Medicine*
5. Schil, P. V., Hendriks, J., Maeseener, M. D., and Lauwers, P. (2005) Current management of spontaneous pneumothorax. *Monaldi Archives for Chest Disease* 63
6. Ouanes-Besbes, L., Golli, M., Knani, J., Dachraoui, F., Nciri, N., Atrous, S. E., Gannouni, A., and Abroug, F. (2007) Prediction of recurrent spontaneous pneumothorax: CT scan findings versus management features. *Respiratory Medicine* 101, 230–236
7. Kattan, M. W. (2003) Judging New Markers by Their Ability to Improve Predictive Accuracy. *JNCI Journal of the National Cancer Institute* 95, 634–635
8. Nguyen, C. T. and Kattan, M. W. (2011) How to Tell If a New Marker Improves Prediction. *European Urology* 60, 226–228
9. Choi, S. Y., Park, C. B., Song, S. W., Kim, Y. H., Jeong, S. C., Kim, K. S., and Jo, K. H. (2014) What Factors Predict Recurrence after an Initial Episode of Primary Spontaneous Pneumothorax in Children? *Annals of Thoracic and Cardiovascular Surgery* 20, 961–967
10. Guo, Y., Xie, C., Rodriguez, R. M., and Light, R. W. (2005) Factors related to recurrence of spontaneous pneumothorax. *Respirology* 10, 378–384
11. Melton, L. J., Hepper, N. C. G., and Offord, K. P. (1987) Incidence of spontaneous pneumothorax in Olmsted County, Minnesota: 1950–1974. *Am Rev Respir Dis* 29, 1379–1382
12. Kahn, B. B., Alquier, T., Carling, D., and Hardie, D. G. (2005) AMP-activated protein kinase: Ancient energy gauge provides clues to modern understanding of metabolism. *Cell Metabolism* 1, 15–25
13. Goncharova, E. C. A. A., Goncharov, D. A., James, M. L., Atochina-Vasserman, E. N., Stepanova, V., Hong, S.-B., Li, H., Gonzales, L., Baba, M., Linehan, W. M., Gow, A. J., Margulies, S., Guttentag, S., Schmidt, L. S., and Krymskaya, V. P. (2014) Folliculin Controls Lung Alveolar Enlargement and Epithelial Cell Survival through E-Cadherin, LKB1, and AMPK. *Cell Reports* 7, 412–423
14. Dabbagh, K., Laurent, G. J., Shock, A., Leoni, P., Papakrivopoulou, J., and Chambers, R. C. (2000) Alpha-1-antitrypsin stimulates fibroblast proliferation and procollagen production and activates classical MAP kinase signalling pathways. *Journal of Cellular Physiology* 186, 73–81

15. Wells, J. M., Esni, F., Boivin, G. P., Aronow, B. J., Stuart, W., Combs, C., ... & Lowy, A. M. (2007). Wnt/ β -catenin signaling is required for development of the exocrine pancreas. *BMC developmental biology*, 7(1), 4.
16. Gao, Y. and Wang, H.-Y. (2006) Casein Kinase 2 Is Activated and Essential for Wnt/ β -Catenin Signaling. *Journal of Biological Chemistry* 281, 18394–18400
17. Castelli, M., Camps, M., Gillieron, C., Leroy, D., Arkinstall, S., Rommel, C., and Nichols, A. (2004) MAP Kinase Phosphatase 3 (MKP3) Interacts with and Is Phosphorylated by Protein Kinase CK2 α . *Journal of Biological Chemistry* 279, 44731–44739
18. Smith, E. P., Specker, B., Bachrach, B. E., Kimbro, K. S., Li, X. J., Young, M. F., Fedarko, N. S., Abuzzahab, M. J., Frank, G. R., Cohen, R. M., Lubahn, D. B., and Korach, K. S. (2008) Impact on Bone of an Estrogen Receptor- β Gene Loss of Function Mutation. *The Journal of Clinical Endocrinology & Metabolism* 93, 3088–3096
19. Jee, Y. H. and Baron, J. (2016) The Biology of Stature. *The Journal of pediatrics*.
20. Haraguchi, S. and Fukuda, Y. (2008) Histogenesis of abnormal elastic fibers in blebs and bullae of patients with spontaneous pneumothorax: Ultrastructural and immunohistochemical studies. *Pathology International*
21. Domínguez-Muñoz, J. E., D, P., Lerch, M. M., and Löhr, M. J. (2017) Potential for Screening for Pancreatic Exocrine Insufficiency Using the Fecal Elastase-1 Test. *Digestive diseases and sciences*.
22. Stanczyk, F. Z. (2006) Measurement of Androgens in Women. *PubMed*
23. Kuzniar, J. (2007) Elastase deposits in the kidney and urinary elastase excretion in patients with glomerulonephritis: evidence for neutrophil involvement in renal injury. *Scandinavian journal of urology and nephrology*
24. Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. (2010) Assessing the Performance of Prediction Models. *Epidemiology* 21, 128–138
25. Clinical Procedures. *Good Clinical Practice* 150–153
26. Chen, F.J., Liao, H., Huang, X.-Y., and Xie, C.-M. (2016) Importance of fractional exhaled nitric oxide in diagnosis of bronchiectasis accompanied with bronchial asthma. *Journal of Thoracic Disease* 8, 992–999
27. Xie, C.-M., Chen, F.-J., Huang, X.-Y., Liu, Y.-L., and Lin, G.-P. (2016) Importance of fractional exhaled nitric oxide in the differentiation of asthma–COPD overlap syndrome, asthma, and COPD. *International Journal of Chronic Obstructive Pulmonary Disease* Volume 11, 2385–2390



Investigating the Characteristics of the Optimal Point for Magnetic Nanoparticle Hyperthermia

Ege Özgüroğlu

Author Background: Ege Özgüroğlu grew up in Turkey and currently attends Robert College in Istanbul, Turkey. His Pioneer seminar topic was in the fields of engineering and was titled "Introduction to Nanoscience & Nanotechnology."

Abstract

This paper is an interdisciplinary study relating nanomedicine and nanothermodynamics, in particular the use of hyperthermia with magnetic nanoparticles (MNPs) as a cancer treatment. Hyperthermia, which is the method of using heat generation of MNPs for tumor diagnosis, is performed by suspending magnetic iron oxide particles (Fe_3O_4) with diameters 18-25 nanometers and 98.45% purity. Providing an alternating current to the magnetic nanoparticles makes them oscillate and generate heat. The point at which MNPs generate heat most efficiently is called the optimal point of hyperthermia. Previous studies have not been able to accurately determine the optimum point itself: Learning from the challenges faced in the previous studies, this paper develops ways to eliminate these additional variables and finds ways to optimize hyperthermia by conducting further experiments in which only those additional variables are observed. While the duration in which the AC field is applied is kept constant, the frequency and concentration of Fe_3O_4 of the suspensions are varied. The influences of concentration and frequency on the heat generation are analyzed alongside a detailed connection to the effects of sedimentation that occurs because Fe_3O_4 MNPs with bare surfaces tend to form aggregates.

1. Introduction

Hyperthermia, the activation of magnetic nanoparticles in an alternating magnetic field, is a technique for killing lymphatic metastases that has flourished recently. Historically, hyperthermia was known as a treatment that had failed to become one of the mainstream cancer treatments because scientists lacked the understanding of the hyperthermia cytotoxicity mechanism and were unable to target it directly to tumor cells: Tumor cells, in comparison to normal cells, were not more sensitive to this treatment. In 1957, Gilchrist et al. conducted the first set of *in vitro* demonstrations in which Fe_2O_3 particles were administered to lymph nodes. After exposing them to an alternating magnetic field, a temperature rise of 14°C was observed; however, the possibility of the electric field generating the heat undermined the importance of the proposed treatment. Over the last 20 years, scientists have started revisiting the hyperthermia treatment and realized that in order for it to be effective, hyperthermia had to be localized precisely to malignant cells. Currently, hyperthermia with magnetic nanoparticles (MNPs) can locate tumors virtually anywhere in the human body. In this *in vitro* experiment, an AC magnetic field and a container that isolates the Fe_3O_4 magnetic nanoparticle suspensions in ionized distilled water

modeled hyperthermia cancer treatment with MNPs. The process of exciting the magnetic nanoparticles was repeated with various frequencies and Fe_3O_4 concentrations, and the change in temperature levels were measured.

2. Method

Iron oxide (Fe_3O_4) nanopowder with 98.45% purity was used as the bare magnetic nanoparticle of the experimental setup. These MNPs were suspended in ionized distilled water. The average diameter of the iron oxide particles was 18 to 28 nanometers. One of the most important variables of the study was the effect of the percentages by mass of the MNPs in the ligand and their interaction with ionized distilled water, which was simulating the extracellular fluid. In order to measure the influence of the concentrations of the MNPs, five different solutions were prepared in 50 ml beakers. We initially sterilized each beaker with both ethylene oxide and autoclave steam, removing any possible remainders of chemical substances from them. The mass of 40 milliliters of ionized distilled water was measured with *Precisa XB 220A* electronic balance to be 39.2 grams. The five 50 ml beakers were then filled with 40 milliliters (39.2 grams) of ionized distilled water. The first 50 ml beaker, the control beaker, did not contain MNPs: Its Fe_3O_4 concentration remained 0%. For the rest of the beakers, 3.92, 5.88, 7.84, and 9.8-gram samples of Fe_3O_4 were measured with *Precisa XB 220A* electronic balance and poured into the beakers to produce 10%, 15%, 20%, and 25% percent suspensions of Fe_3O_4 respectively. The solutions in each test tube were poured into *Fisher Scientific* glass centrifuge tubes and rotated at 6,000 rpm for one-hour intervals in a 12. Hettich Universal 320R Benchtop Centrifuge. Then, magnetic stirring bars (fish) were immersed into each testing beaker, and the beakers were placed on four *Heidolph MR Hei-Standard* magnetic stirrers, shown in Figure 0. The final step of the suspension preparation was to stir the suspensions at 300 rpm until they were used in the experiment. Magnetic stirring distributed the Fe_3O_4 MNPs evenly to the beaker and held the solution as homogeneously as possible. This prevented the suspensions from sedimenting before testing started.



Figure 1. 20% and 25% Fe_3O_4 suspensions on *Heidolph MR Hei-Standard* magnetic stirrer

2.1 Experimental Setup

As per the aim of the experiment, to investigate challenges and ways to eliminate them in experiments that investigate the optimal point of hyperthermia, a complex but reasonable experimental design was prepared. The materials used for the setup are as follows:

1. TES 1307 K/J thermocouple;
2. BK Precision 10MHz sweep/function generator;
3. Kenwood CS-4125 20MHz oscilloscope;
4. Fluke 45 dual-display multimeter;
5. Copper conducting wires;
6. Styrofoam container;
7. 3M TC-2810 thermally conductive epoxy;
8. Insulation tape;
9. One-thousand-turn, hand-wound coil on amorphous oxide core;
10. DynaPro NanoStar dynamic light scattering (DLS) instrument;
11. Four Heidolph MR Hei-Standard magnetic stirrers;
12. Hettich Universal 320R benchtop centrifuge

The coil used was hand-winded on a magnetic amorphous alloy core with copper wire with a diameter of 0.9 mm. Due to the limited current range of the *BK Precision* 10MHz Sweep/Function Generator, the coil was prepared with 1,000 turns and with an amorphous core aiming to maximize the magnetic flux. The magnetic field due to the coil can be calculated using the following equation:

$$B = \frac{\mu_0 I}{2R} N, \quad (1.1)$$

where $\mu_0 = 4\pi \times 10^{-7} \frac{Tm}{A}$, I is the current, N is the number of turns (1,000), and R is the radius of the coil ring (0.9 mm).

Isolating the beakers containing the suspensions was an essential part of the experimental setup. Although the fluctuations in the room temperature were being reduced as much as possible it was imperative to conduct the experiment in a thermally isolated container. The container, made of Styrofoam had an inner vessel. A hole was punched between the inner and the outer vessels for the wire of the probe of the thermocouple. After plugging in the thermocouple wire, the hole was filled with *3M thermally conductive epoxy* and the wire was fixed to the walls of the inner vessel. After conducting the experiment with each of the four concentrations of Fe_3O_4 (10%,15%,20%,25%), the container was opened and a new beaker with a different suspension was placed in the inner vessel. Then, the container was closed with the aforementioned isolation procedure. The same isolation procedure was repeated five times, for each beaker, in the duration of the experiment. The probe of the thermocouple was fixed on the outside of each beaker with *3M thermally conductive epoxy*.

The styrofoam container was placed at the center of the coil both in the x and y axes in order to have the MNPs oscillating at a specific point where the magnetic field created by the coil was centered. The diagram of the electrical circuit that was used to create an alternating current (AC) magnetic field is shown in Figure 2.

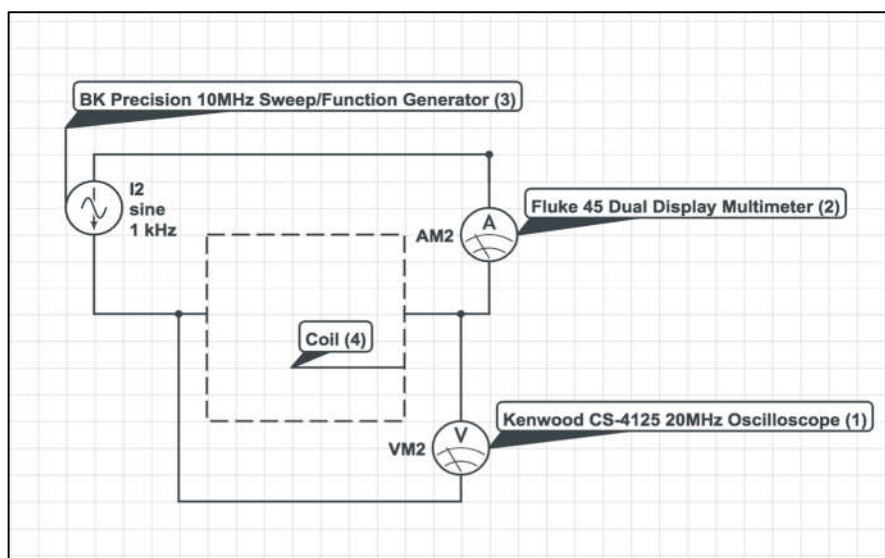


Figure 2. The electric circuit used in the experimental procedure.

One of the two copper wires coming out of the coil was connected to the *Fluke 45 Dual Display Multimeter (2)*, which was then connected to the *BK Precision 10MHz Sweep/Function Generator (3)*. Thus, the function generator and the dual display multimeter were connected to the coil in series. The *Kenwood CS-4125 20MHz oscilloscope (1)*, on the other hand, was connected to the coil in parallel: It was used to measure the peak to peak (pp) amplitude of the generated sinusoidal waves. The thermocouple was only connected to the inner vessel of the styrofoam container and was not a part of the electrical circuit.

2.2 Calibration

Before any measurement was taken, a beaker with 40 milliliters of water was placed in the inner vessel of the styrofoam container, and the laboratory was kept at the constant temperature with fans and an air conditioner. When the room temperature was stabilized at 23°C, that is, no fluctuations were detected by a thermometer, the AC field was turned on. The field was applied to the beaker for 30 minutes and the thermocouple screen was monitored. No temperature fluctuations were measured during that time. The same procedure was repeated with a beaker without water in the styrofoam container, and again no temperature fluctuation in the inner vessel was measured. Confirmation of the isolation was completed alongside the calibration of the thermocouple.

2.3 Measurement

Each of the four beakers containing its own concentrated amount of Fe_3O_4 was kept on the magnetic stirrers with the magnetic stirring fish inside of it. The beakers were stirred until they were tested. For example, during the testing of the 10% Fe_3O_4 suspension, the 15%, 20%, and 25% Fe_3O_4 suspensions were still on the *Heidolph MR Hei-Standard* magnetic stirrers. When the time came for a suspension, its magnetic stirrer was turned off, and it was placed in the styrofoam container. The thermocouple probe was fixed on the exterior of the glass beaker with a thermally conductive adhesive. Once the probe was fixed, the container was closed and isolated. The beaker was then left to rest for 10 minutes while the system was turned off to ensure constant room temperature, after which the function

generator was adjusted to the first value of frequency and turned on. At the instant the system started, a chronometer was started, and the initial temperature was recorded. At the end of five minutes, the final temperature was recorded, and the system was turned off. Then, the frequency was adjusted to the second value, and both the system and the chronometer were started. The same procedure was repeated for the remaining two frequency values. Also, this procedure was repeated for each of the four suspensions.

3. Results

3.1 The Extremes: Suspensions of 10% and 25% Fe₃O₄

The results of the experimental process do not indicate the optimum point of hyperthermia in general; however, several significant trends can be found with data analysis. While these trends demonstrate the influence of frequency, magnetic field strength, and concentration on the heat generation, they are also indicators of challenges and signs that other variables might be involved in the process of optimizing the heat generated by magnetic nanoparticles.

The analysis of the least and the most concentrated suspensions (10% and 25%) indicates an unexpected and counterintuitive notion. One might expect to see an increase in the amount of heat iron oxide nanoparticles generate as the concentrations of the suspensions increase from the least to the highest, but this is not the case. When the frequency is 900 Hz, the MNPs in the 10% Fe₃O₄ concentration change the temperature of the solution by 1°C. This is the same as the change in temperature of the 20% Fe₃O₄ suspension at the same frequency. Similarly, the change in temperature caused by 10% Fe₃O₄ concentration at 8,064 Hz is again 1°C, which is identical to the temperature change caused by 20% Fe₃O₄ at 8,064 Hz. These data show that increasing the concentration from the least concentrated suspension to the most does not directly correlate to higher heat generation. In fact, both 10% and 25% concentrated suspensions caused no change in temperature at the highest frequency the *BK Precision* 10MHz Sweep/Function Generator could achieve: At 11,232 Hz, the 10% concentrated suspension did not generate any measurable amount of heat; its temperature remained 26.5°C while the AC field was applied to the MNPs for five minutes. Similarly, the 25% concentrated suspension remained at 26.7 °C.

Table 1. The change in temperature levels of the 10%, 15%, 20%, 25% Fe_3O_4 suspensions depending on the tested frequency values

	Frequency (Hz)	Electric Current (mA)	Initial Temperature (°C)	Final Temperature (°C)
10% Fe_3O_4 A (pp) (V): 6.8 Duration: 5 minutes	900	9.66	26.1	26.2
	3,020	8.53	26.2	26.4
	8,064	7.39	26.4	26.5
	11,232	7.41	26.5	26.5
15% Fe_3O_4 , A (pp) (V):6.8 Duration: 5 minutes	900	9.66	25.8	26.1
	3,020	8.53	26.1	26.2
	8,064	7.39	26.2	26.6
	11,232	7.41	26.6	26.7
20% Fe_3O_4 , A (pp) (V):6.8 Duration: 5 minutes	900	9.66	25.7	25.8
	3,020	8.53	25.8	26.0
	8,064	7.39	26.0	26.2
	11,232	7.41	26.2	26.6
25% Fe_3O_4 , A (pp) (V):6.8 Duration: 5 minutes	900	9.66	26.5	26.6
	3,020	8.53	26.6	26.6
	8,064	7.39	26.6	26.7
	11,232	7.41	26.7	26.7

Similar to how increasing the concentration did not have a significant effect on heat generation, increasing the frequency from 900 Hz to 11,232 Hz did not increase the heat generated as one might have expected. When compared to the change in temperature data from the suspension with 10% Fe_3O_4 , the overall changes in temperature were very minor and less likely to be due to the heat generated by the MNPs in the 25% concentrated Fe_3O_4 suspension. Therefore, the data from the least concentrated suspension is a better source for the analysis of the effect of frequency on heat generation. As seen from Figure 3, for the 10% concentrated suspension, the temperature increased from 1 to 0.2°C as the frequency increased from 900 Hz to 3,020 Hz; however, the change in temperature was once more 0.1°C at 8,064 Hz. Remaining faithful to the assumption that the exterior temperature fluctuations were not affecting the measurements due to effective styrofoam isolation, one must question the rationale of such a decrease. It was expected that frequency would cause an increase in the change in temperature, heat generation so to speak; on the other hand, this experiment showed a decrease in the change of temperature for 8,064 and 11,232 Hz (higher frequencies). Also, as mentioned earlier, the overall increase in temperature of the 10% concentrated suspension was relatively higher than the 25% concentrated suspension. The

counterintuitive nature of these results may be due to a new variable the experimental procedure has not controlled: sedimentation.

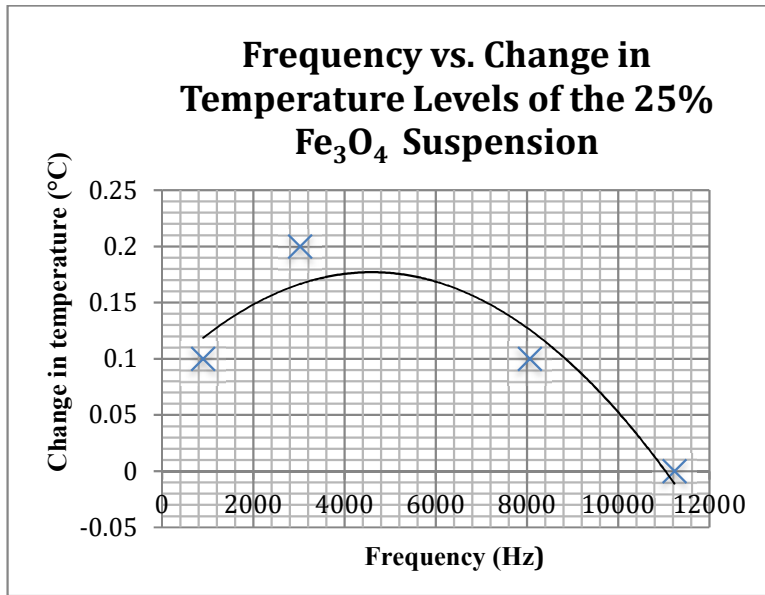


Figure 3. The effect of frequency on the change in temperature levels of the 20% Fe₃O₄ suspension.

3.2 Sedimentation and Aggregation of Suspended Fe₃O₄ Magnetic Nanoparticles

Sedimentation is a process in which particles sink and settle in a suspension under the effect of gravity. Aggregation, the precursor of sedimentation, brings iron oxide particles together, making them denser. As a result, these “homoaggregates” of nanoparticles start to sediment.

3.2.1 Mathematical Models for Aggregation and Sedimentation

Aggregation between Fe₃O₄ nanoparticles mainly occurs due to the Van der Waals forces that increase in magnitude when the distance between molecules shortens, and magnetic attraction, which occurs when electron spins are aligned. The following is the Lennard-Jones potential equation, a good approximation of the interaction energy, E , between the same type of atoms:

$$E(x) = \left[\left(\frac{2r_{vdw}}{x} \right)^{12} - 2 \left(\frac{2r_{vdw}}{x} \right)^6 \right]. \quad (2.0)$$

Taking the negative derivative of Equation 2.0 gives the equation for the Van der Waals forces since the force curve is the derivative of the energy curve:

$$F(x) = -\frac{dE}{dx} = 12\epsilon \left[\left(\frac{2r_{vdw}}{x^{13}} \right)^{12} - \left(\frac{2r_{vdw}}{x^7} \right)^6 \right]. \quad (2.1)$$

In Equations 1.0 and 1.1, x represents the separation between two individual atoms, and ε (well depth) and r_{vdw} (Van der Waals radius) are parameters that vary from one atom to another. A quick analysis of how the attractive Van der Waals forces change when atoms get closer together explains the aggregation of the Fe_3O_4 particles in suspensions with higher concentrations. Some mathematical intuition shows that as x , the separation distance, decreases, the attractive force increases. Similarly, looking at Equation 1.0, it can also be said that as the separation distance decreases, individual atoms of the MNPs will have a greater interaction energy; that is, they will be more likely to form aggregates.

Furthermore, the probability of attachment for similar magnetic nanoparticles is defined as the attachment efficiency (α). When the AC field is provided, the iron oxide magnetic nanoparticles form large, but less dense, dendritic aggregates. ($\alpha = 1$). On the other hand, as the nanoparticles undergo collisions during their oscillation, the attachment efficiency decreases ($\alpha < 1$) and denser but less dendritic aggregates start to form.

Both the mathematical model for the Van der Waals forces and the assumption of how the density of the aggregates changes as nanoparticles undergo more and more collisions is the rationale of the suspension with 25% Fe_3O_4 : As percentage by mass in a suspension increases, the separation distance between the nanoparticles and the atoms they form decreases. This increases the attraction between the iron oxide MNPs, causing them to form aggregates and then suspend due to their higher densities.

3.2.2 Experimental Validation of the Effect of Suspension Concentrations on Sedimentation with Dynamic Light Scattering (DLS)

The mathematical models in Section 3.2.1 are based on the assumptions that the Lennard-Jones potential equation is a good approximation of the attractive Van der Waals interactions, and that the attachment efficiency, that is the probability of aggregation, decreases as more collisions occur, increasing the density of the newly formed aggregates. Regardless of how credible these assumptions are, they are not epitomes of the behaviour of the four Fe_3O_4 suspensions used in our experimental process. In order to best validate the sedimentation theory and relate it to the lower changes in temperature at higher concentrations and, often, frequencies, dynamic light scattering (DLS) technology must be used. DLS is a technology often used to determine particle sizes or their changes by measuring the changes in the intensity of light as it is scattered by shining it through a solution. In this case, data on the sedimentation rates of the suspensions were obtained with the use of a DynaPro NanoStar DLS Instrument. The following data were obtained while the suspensions were in the AC field, and the results are shown in Figure 4. During the DLS measurement, the styrofoam container was not used due to its opaque nature. The effect of concentration on the rates of aggregations was analyzed.

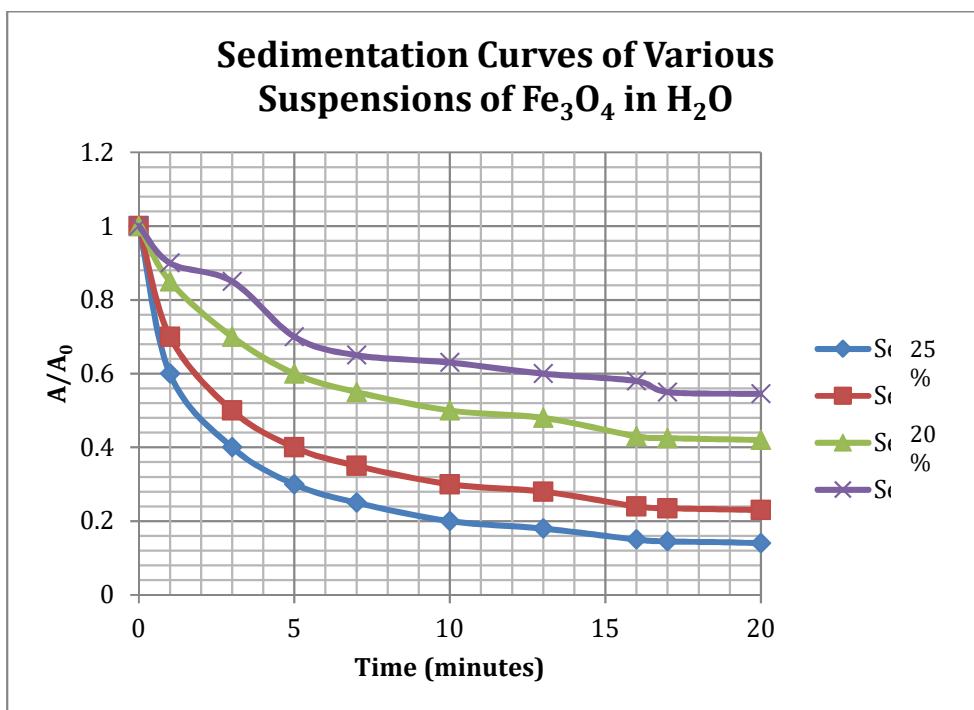


Figure 4: The sedimentation curves of the 10%, 15%, 20%, 25% Fe_3O_4 suspensions in ionized distilled water and under the effect of an AC Field.

The y-axis values represent A/A_0 , where A is the absorbance and A_0 is the absorbance at $t = 0$ minutes. The more aggregates form in a particular suspension at a particular time, the less light the suspension will absorb. Therefore, as aggregation and sedimentation rates increase, the ratio of A/A_0 decreases: The more aggregated MNPs are, the less light will be absorbed. The results of this experimental check imply that the mathematical models based on the Lennard-Jones potential equation are applicable to our set of Fe_3O_4 suspensions. More importantly, it can be concluded that as concentration increases, sedimentation takes place more rapidly. In other words, at a specific time, the 25% concentrated suspension will be the most sedimented one. On the other hand, the 10% concentrated suspension will encounter the lowest rate of sedimentation at the same time. As the rate of sedimentation of nanoparticles increases, they cannot oscillate easily in the AC field. To understand whether the amount of concentration is also inversely proportional to the amount of heat generated by the Fe_3O_4 magnetic nanoparticles (which is directly proportional to the change in temperature) still requires more analysis. If this inverse relationship is in accordance with our further analysis, we can treat changes in temperature values at higher concentrations more significant than temperature changes at lower concentrations. It will be harder for the MNPs to change the temperature of the suspension by the same value if the sedimentation rate is higher. Further analysis will be presented in the following section.

3.2.3 Sedimentation Data Analysis

Looking at the change in temperature values when the frequency is 8,064 Hz, which happens to be the optimal spot for our set of experimental data, the effect of

sedimentation can be seen more clearly. Since 8,064 Hz is the third value of the tested frequencies, the corresponding time interval is between $t = 10$ minutes and $t = 15$ minutes. The changes in temperature at the frequency the DLS measurements were taken were 0.1 °C, 0.4 °C, 0.2 °C, and 0 °C for 10%, 15%, 20%, and 25% Fe_3O_4 suspensions, respectively (see Figure 5).

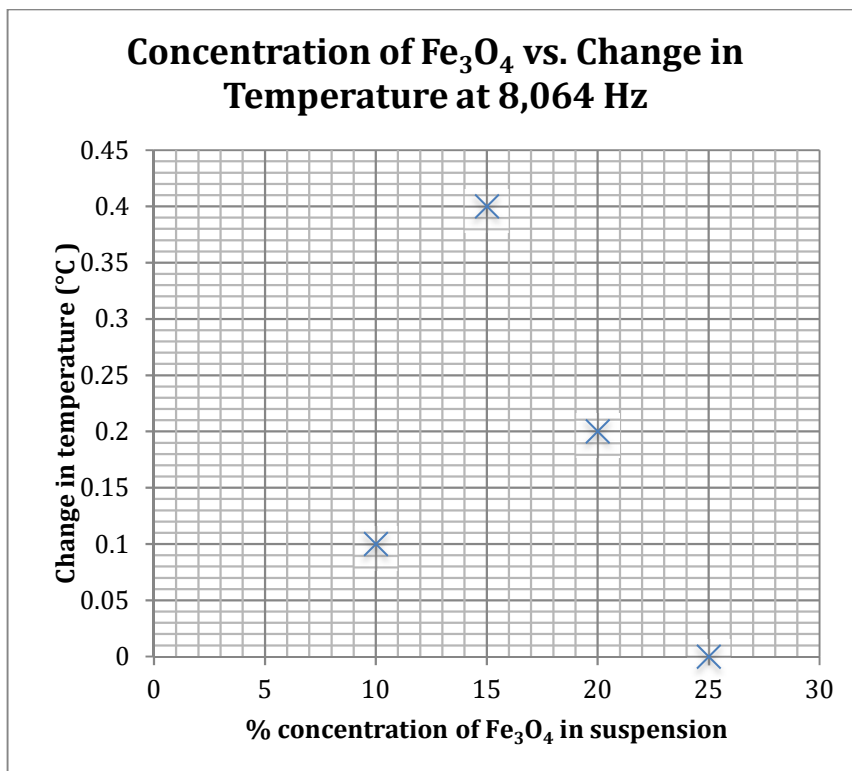


Figure 5. The effect of Fe_3O_4 concentration on the change in temperature values at 8064 Hz.

According to the results of the DLS data collected between $t = 10$ and $t = 15$, the most sedimentation occurred in the suspension with the least concentration of iron oxide (10%); however, it was not the 10% suspension that had the greatest change in temperature at 8,064 Hz. The greatest change in temperature, that is heat generation, was obtained in the 15% Fe_3O_4 suspension. It is also worth noting that after 15%, every increase in concentration coincided with a decrease in the change of temperature. Unlike the linear trend observed in the case of the 15% suspension concentration, the change in temperature in the 10% Fe_3O_4 solution (0.1°C) was not in accordance with what we expected from the DLS data: Since it had a lower rate of sedimentation, it was expected to generate more heat. A possible comment on this discrepancy would be intuitively that the concentration was too low, causing a decrease in the amount of generated heat, which in turn slowed down the process of transferring energy to the glass beaker (which is connected to the thermocouple probe); however, the reason behind this unexpected decrease in temperature value cannot be identified precisely from the data this experiment yields, since it is inconsistent with all the variables tested in this experiment.

3.2.4 The Optimal Point of the Experimental Data

Although the data obtained from this experiment is quite limiting, primarily due to limited resources (limited and deviating electric current values), the usage of bare nanoparticles, etc.) and time constraints, it is still possible to identify an optimum point.

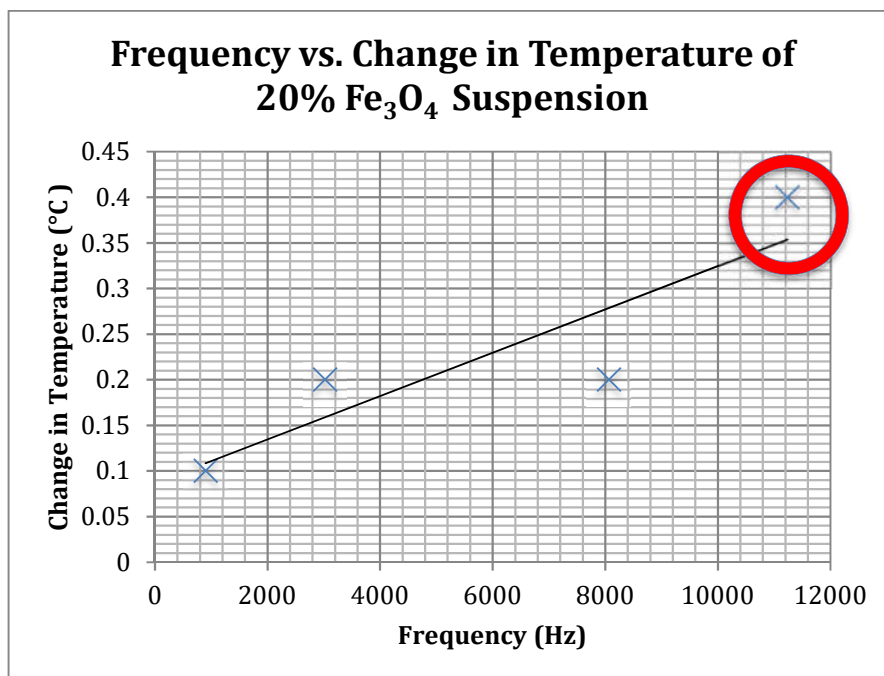


Figure 6. The effect of frequency on the change in temperature levels of the 20% Fe₃O₄ suspension.

From the data in Table 1, the optimal point of the experimental data is determined as the 20% suspension at a frequency of 11,232 Hz. Figure 6 shows the change in temperature for each frequency value in the 20% Fe₃O₄ suspension, and our optimal point is circled in red. Even though the sedimentation rate at this point was higher than at any other time interval (Figure 3) for the 20% Fe₃O₄ suspension itself and also the same time interval for the 10% and 20% Fe₃O₄ suspensions, the change in temperature had one of the two highest values: 0.4 °C. The other 0.4 °C change in temperature occurred at 8,064 Hz in the 15% Fe₃O₄ suspension, causing its particles to aggregate at a lower rate, thereby generating heat more easily. As indicated by Figure 3, since it occurred at a lower frequency, less time had passed until that point from the start of sedimentation, so the sedimentation rate was lower. Also, the 5% difference in the Fe₃O₄ concentration helped it aggregate less. Our optimal point (11,232 Hz and 20% Fe₃O₄) had the highest change in temperature in the most challenging circumstances where particles cannot move easily (more sedimentation), so to speak.

4. Suggestions for Further Experimentation

Sedimentation, the deviating electric current values, the long duration of heat transfer (between the glass beaker and the iron oxide suspensions), and a low magnetic field strength were the major limiting factors of this experiment. The usage of dress coated nanoparticles, a water-resistant thermocouple probe, and a current amplifier are possible solutions that eliminate the limiting factors as explored in depth in Sections 4.1, 4.2, and 4.3. Implementing these solutions, though, would require endless time and no budgetary constraints.

4.1 Dress Coated Nanoparticles

If this experiment were conducted with enough time, the magnetic nanoparticles would have been dress coated. In this experiment, bare Fe_3O_4 particles were suspended, and this created a significant setback: aggregation and sedimentation. The higher the sedimentation rate, the harder it is for the MNPs to oscillate and generate heat. It is possible to extrapolate the influence of concentration on heat generation using mathematical models, or to use a dynamic light scattering device to obtain data on the sedimentation rates. Nevertheless, with dress coated nanoparticles covered with surfactants composed of silver, gold, or polymers that lower the surface tension of nanoparticles and increase solubility, we would have one less variable to worry about. We could then measure higher changes in temperature or spot the same changes in shorter amounts of time. If surfactants will not be available in future experiments, the densities of the aggregates formed by the particles in the suspensions can be measured; this will act as a validation for the assumption we made on the attachment efficiency. Whether or not the attachment efficiency becomes less than one as particles undergo many collisions would then be answered with data.

4.2 Thermocouple Probe Replacement

The thermocouple probe was not in the suspension itself; rather, it was fixed to the exterior of the glass beaker. The limited laboratory conditions only enabled the usage of a probe that cannot be put in liquid. If, in a future experiment with more resources and time, a fiber coupled probe that can take measurements inside of a liquid were used, higher changes of temperature would be measured in the same time intervals. Since the heat generated by the nanoparticles is absorbed by the ligand, the heat will also need to be transmitted from the ligand to the glass. Even though the probe on the glass instantaneously measures changes of temperature of the glass, heat transmission from the water to the glass will take a long time. This is why the actual temperature changes could have occurred after total measurement duration of 20 minutes for each substance. Using a fiber coupled probe that is very sensitive will also decrease the uncertainty in data while helping the iron oxide nanoparticles show the amount of the heat they generate more quickly. Furthermore, using a thermocouple with a datalogger would be beneficial: In that case, not only would the initial and final values of temperature within an interval be known, but also all the fluctuations within a certain interval will be measured. This way, we would also prevent the possibility of contaminating the results of different trials by considering the increases in temperature less than 1°C . For example, the final temperature at 900 Hz may be 11.40°C , and we will measure it as 11°C . At the end of the next frequency interval, 3,020 Hz, let us assume that the final temperature is 12°C ; in this case, we will note that the change in temperature for the 3,020 Hz interval is 1°C , as opposed to the actual change in temperature of 0.6°C . This discrepancy is due to the measurement scale of the thermocouple, causing readings to be greater or lower than the actual temperature changes. Using a thermocouple with a datalogger would allow us to read the temperature changes accurately.

4.3 Higher Currents without Deviation

The maximum current *BK Precision* 10MHz Sweep/Function Generator could generate was 13.44 milliamperes. By using Equation 1.1 the maximum magnetic field strength, B , at the center of the coil in this experiment can be extrapolated using the following equations:

$$B = \frac{\mu_0 I}{2R} N$$
$$B = \frac{4\pi \times 10^{-7} (1344 \times 10^{-3})}{2(09 \times 10^{-3})} 1000,$$
$$B = 9.38 \text{ mT}$$

By using a current amplifier, which was not available for this experiment, this magnetic field strength value can be greatly increased. Also, the deviations in the order of milliamperes in the current value that also affect the magnetic field strength can be eliminated by directly measuring the magnetic field at the center of the coil with a Gauss Probe instead of using the *Fluke 45* Dual Display Multimeter. Also, an electronic voltmeter with more than three decimal places can be used instead of an oscilloscope in order to vary the voltage instead of keeping it constant. This would more or less stabilize the value of the current, eliminating it as a variable. By having a constant electric current value, the magnetic field strength will also be kept constant. This will help us correlate the reason for the changes in temperature levels directly to changes in frequency and iron oxide concentration.

5. Conclusion

In this paper, hyperthermia, the use of heat generation to eliminate secondary malignant growths after the administration of MNPs, was modelled with iron oxide nanopowder suspensions at concentrations of 10%, 15%, 20%, and 25%. Due to the limiting amount of current generated by the function generator, the strengths of the alternating magnetic fields were in the order of milliteslas. This, alongside the lack of surfactants on the iron oxide particles caused the temperature changes to be relatively small. We expected highest concentrations to have the greatest amounts of heat generated; however extreme cases of frequency and concentration were measured to have a lower affinity to generate heat considering the effect of aggregation and sedimentation. As the experiment showed, these conditions significantly hindered the oscillation of magnetic nanoparticles. Sedimentation, the most significant obstacle of oscillation, was explained either by the Van der Waals forces or the magnetic interactions that result from the aligning of electron spins. The yields of the Van der Waals mathematical equations for force and interaction energy were also experimentally validated, showing that as the concentrations of iron oxide in suspensions increased, the sedimentation rates were measured to be higher. The optimal point from the experimental data was found to be 11,232 Hz and in the suspension of 20% Fe₃O₄.

Works Cited

Derlecki, Stanislaw. "Magnetic Properties of Amorphous Materials Used as Corps of Electric Machines." *Instytut Mechatroniki I Systemów Informatycznych (2) Magnetic Properties of Amorphous Materials Used as Corps of Electric Machines* (2012): n. pag. Web.

- Giustini, Andrew J., Alicia A. Petryk, Shiraz M. Cassim, Jennifer A. Tate, Ian Baker, and P. Jack Hoopes. "Magnetic Nanoparticle Hyperthermia in Cancer Treatment." *Nano LIFE*. U.S. National Library of Medicine, 2010. Web.
- Hotze, Ernest M. "Nanoparticle Aggregation: Challenges to Understanding Transport and Reactivity in the Environment." *Journal of Environmental Quality* 39 (2010): n. pag. Carnegie Mellon University, Dec. 2010. Web.
- Hotze, Ernest M., Tanapon Phenrat, and Gregory V. Lowry. *Nanoparticle Aggregation: Challenges to Understanding Transport and Reactivity in the Environment* 39 (2010): n. pag. *Journal of Environmental Quality*. Carnegie Mellon University, Dec. 2010. Web.
- Keller, Arturo A., Hongtao Wang, Dongxu Zhou, Hunter S. Lenihan, Gary Cherr, Bradley J. Cardinale, Robert Miller, and Zhaoxia Ji. "Stability and Aggregation of Metal Oxide Nanoparticles in Natural Aqueous Matrices." *Environmental Science & Technology* 44.6 (2010): 1962-967. *Environmental Science & Technology*. Web.
- "Magnetic Field of Current Loop." *Magnetic Field of a Current Loop*. Hyper Physics, n.d. Web.
- Markus, A. A. "Modeling Aggregation and Sedimentation of Nanoparticles in the Aquatic Environment." *Science of The Total Environment*. Elsevier, 24 Nov. 2014. Web.
- Salager, Jean Louis. "Uses of Phosphates: Industrial Demand." *Surfactants Types and Uses* 2012.10 (2012): 4. *Laboratory of Formulation, Interfaces Rheology and Processes*. Universidad De Los Andes Facultad De Ingenieria Escuela De Ingenieria Quimica, 2002. Web.
- Wiesner, Mark, and Zhahohan Zhang. "Transport of Cerium Oxide Nanoparticles in Saturated Silica Media: Influences of Operational Parameters and Aqueous Chemical Conditions." *Nature News*. Nature Publishing Group, n.d. Web.
- Woo, S.h., Min Ku Lee, and Chang Kyu Rhee. "Sedimentation Properties of TiO₂ Nanoparticles in Organic Solvents." *Solid State Phenomena Nanocomposites and Nanoporous Materials VII* (2007): 267-70. *University of Bergen-Research Cluster*. Haukeland University Hospital. Web.



Kiezdeutsch: A Youth Dialect in the Face of Linguistic Conservatism

Zehan Zhou

Author Background: Zehan Zhou grew up in Canada and currently attends St. George's School in Vancouver, Canada. His Pioneer seminar topic was in the field of anthropology and titled "Communication and Culture."

Abstract

This article illustrates how a burgeoning youth register¹, Kiezdeutsch, serves as a mechanism for both fringe nationalist groups and modern German society to project their negative evaluations of that register. The register itself was originally construed as a form of self-empowerment for second and third-generation immigrant Germans through using a unique grammatical and lexical repertoire, conspecific neither to Standard High German nor to their native cultures. To show the grammatical and lexical repertoire, this article grammatically analyzes Kiezdeutsch grammar and vocabulary by regarding Kiezdeutsch utterances found in a variety of different sources, such as rap and interview. To demonstrate the negative evaluations of Kiezdeutsch and how these negative evaluators take advantage of the youth register to vent their sentiments, I included examples of scholarly research projects and media analyses.

Exclusionary Influences

The coming of age of second-generation immigrant children in Germany has given rise to a new youth register, *Kiezdeutsch*. This register takes some influence from *Gastarbeiterdeutsch*, the fixed "learner" German register used by the first-generation immigrants to Germany. Characterized by linguistic simplification and interferences from home languages, *Gastarbeiterdeutsch* used both grammatical constructions and loanwords from a variety of non-Germanic sources in a pattern of language crossing (Androutsopoulos 2001), and this characteristic has been readily adopted by *Kiezdeutsch*. However, while *Kiezdeutsch* was developed as an avenue to escape exclusionary influences by mainstream German society and their own native societies, the register itself has been made an acceptable avenue for the airing of xenophobic sentiments by not only linguistic conservatives, but also by the mainstream German society.

The exclusionary influences in German society that prompted the genesis of *Kiezdeutsch* are best seen in the terms *Kanak* and *Kümmel*. *Kanak* is equivalent to the English term "wog," used as a xenophobic epithet to designate the immigrant children themselves (Wiese 2015). The word *Kanak*, originating from the Hawaiian *kanaka*, or "man", gained a derogatory meaning in the German language from the immigration process

¹ A register here is defined as the linguistic repertoire that is internally associated with particular social practices and with persons who engage in such practices (Agha 2007).

(Matthes 2007, Wiese 2015). While it was first used to refer to those immigrants of southern European descent, it grew to include, and primarily pinpoint, the later-arrived Turkish and Arab immigrants. *Kanak* grew to represent the gap between “foreigners” and the German mainstream. Formerly, *Kiezdeutsch* was designated as *Kanak Sprach*, with the word *Sprach* a truncation of the Standard High German word *Sprache*, which means “language” or “speech” (Wiese 2015), and this usage implicitly expresses a certain prejudice, by linguistic conservatives, on the supposedly lower intelligence of *Kiezdeutsch* speakers. The coinage of the term *Kanak Sprach* can be attributed to Feridun Zaimoğlu, who wished to change *Kanak* in an empowering word for second and third-generation immigrants via his 1995 book *Kanak Sprach: 24 Mißtöne² vom Rande der Gesellschaft* (lit. *Kanak Sprach: 24 Off-Tones from the Margins of Society*). The usage of *Kanak* as an empowering word is much like the use of the word *perker*³ in Danish (Madsen 2013). Originally a similarity derogative term pointing to Danish migrants, the usage of *perker* has grown to invoke values of toughness and street credibility, and much like *Kanak*, has been used to characterize the Danish youth immigrant register, *perkersprog* (Madsen 2013). Unlike the successful attempt in Danish in converting *perker* into a positive term, however, the word *Kanak* continues to retain its negative connotation, so much so that the youth language was renamed *Kiezdeutsch*, or “neighbourhood German”, to infuse it with positive meaning.

The word *Kümmel* denotes a similarly pejorative connotation: originating from the expression *Kümmeltürke*, it referred to the high consumption of cumin by the German Turkish immigrants, and thus, further labels the immigrants as an “outsider” group, separate from the communal German identity (Matthes 2007). While there have been past efforts to turn the derogatory appellations into words of self-empowerment, such as Feridun Zaimoğlu’s coinage of *Kanak Sprach*, most of these efforts have met with mixed results, and these exclusionary words still bear much of their derogatory significance (Wiese 2015). Perhaps it is no wonder, as Heike Wiese states, that ethnic Turkish immigrants in Germany are known as *Deutschtürke*, or “German Turks,” while ethnic Germans in Russia are called *Russlanddeutsche*, or “Russian Germans,” no matter how great their degree of removal from German culture. (Wiese, 2014). Much of what defines the German identity comes from German ethnicity and thus, Turkish immigrants fall outside of that definition, being perpetually characterized as exotic and foreign.

Exclusion also comes from the native countries of the second and third generation immigrants. An outstanding example can be observed in the term *almançı*, a Turkish pejorative to refer to those Turkish immigrants who live in Germany and who have become “Germanized,” and thus no longer belong to Turkey (Matthes, 2007). As a result, many of the second and third generation immigrants feel an internal division between the exclusionary proclivities of both cultures, posing the question of to which culture they belong. Zaimoğlu shows this very well in his record of Memet, a 29-year-old poet in Germany, who states that “he feels unwell, because two souls, or rather, two cultures live in him . . . The *Kanak* is something like a synthetic product that hates both himself and the factory, by which he was made” (Zaimoğlu, 1995). Another one of Feridun Zaimoğlu’s interviewees, Hasan, a 13-year-old tramp and student, states that “you also stay a bastard with the family and with the name, you have frizzy hair and you don’t behave like the

² In Standard German, the letter “ß”, or the *Eszett*, is used as the double “s” in fixed contexts. Sometimes, words double “s” do not require the *Eszett* (i.e. *nass*, meaning “wet”), while in other contexts, the *Eszett* is required as a part of the word’s spelling (i.e. *Maßnahme*, IPA [ˈma:s,na:mə], meaning “measure”)

³ Related to the English term “Paki”

Germans . . . you have a plan, but a lot of assholes would like to have you outside of your element, and when you don't protect yourself, they bring you to the management and make you a dark candy-ass" (Zaimoğlu 1995). As a social response to the emphases on their outsider identity and to the assimilationist forces on either side, many second and third generation immigrants balance both their German-ness and their immigrant roots through the means of *Kiezdeutsch*, crafting a unique identity balance between themselves and the wider world.

Kiezdeutsch Grammar and Vocabulary

The *Kiezdeutsch* grammatical systems and loanword usage play a large role in helping *Kiezdeutsch* speakers achieve a balance. The unique morphosyntax used in *Kiezdeutsch* shows the language crossing phenomena that allow the speakers of *Kiezdeutsch* to emphasize their German-immigrant duality as integral to their community identity (Androutsopoulos 2001). An example of the grammatical changes in *Kiezdeutsch* is in an utterance by Leyla to her peer Hatice (Pohle et al. 2014):

Gestern isch war Ku'damm, bei Veromoda.
 Yesterday I was Ku'damm⁴ at Vero-Moda⁵

Isch guck so, alles Rabatt, ischwöre.
 I see PTCL⁶ everything discount, i-swear.

In Standard High German, Leyla's utterances would be

Gestern, war ich am Ku'damm, bei Vero - Moda.
 Yesterday was I at+the⁷ Ku'damm, at Vero - Moda.

Ich guckte, dass alles im Angebot waren, wirklich.
 I saw, that everything in+the⁸ offer was really

⁴ An abbreviation of the *Kurfürstendamm*, a street in Berlin

⁵ A shopping mall

⁶ PTCL is the abbreviated form used in this paper to mean a speech particle, a word with a grammatical function but which cannot fit into any of the major speech classifications (i.e. noun, verb, adjective, adverb, pronoun, conjunction, etc.)

⁷ *Am* is a contraction between the two-way preposition *an* and the dative form of the masculine definite article *der, dem*. In Standard High German, articles and adjectives modifying nouns, and occasionally the nouns themselves, will change in form based on the position of the noun, or its grammatical case, and its grammatical gender (in Standard High German, either masculine, feminine, or neuter.) In the dative case, the noun is an indirect object, and the masculine definite article would change accordingly.

In English, Leyla's speech could be translated as

Yesterday, I was at Ku'damm, at Vero-Moda.

I saw, like, everything on discount, really.

Leyla's speech displays numerous characteristics of *Kiezdeutsch*. At the beginning of her preliminary utterance, instead of using the Verb-Subject-Object (V-S-O) word order after the adverbial "*Gestern*", as would be the case in Standard High German, she maintains the Subject-Verb-Object (S-V-O) word order. This rigidity of word order is not simply unique to *Kiezdeutsch* in relation to the Standard Register: in Sweden and Denmark, the development of *Rinkeby Svenska* and *Københavensk Multiethnolekt*, both registers used by second and third-generation immigrant Swedes and Danes (Paul, Freywald, Wittenberg, 2008) show a similar word order rigidity through keeping the S-V-O word order after elements in the face of Standard Swedish and Danish, languages that employ a V-S-O word order after elements.

Here is an example of a *Rinkeby Svenska* utterance, as provided by the *Kiezdeutsch Infoportal*, compiled by Professor Heike Wiese and supported by the Federal Ministry for Education and Research of Germany, showing the rigid S-V-O word order when the clause is preceded by an adverbial.

Igår jag var sjuk
Yesterday I was sick

In Standard Swedish, this would be

Igår var jag sjuk
Yesterday was I sick

Such grammatical simplification employed by these migrant youth languages parallels development in Standard English, a Germanic language with similar word rigidity.

In her speech, Leyla also employs a pronunciation of German words unique to *Kiezdeutsch*. Instead of pronouncing the first-person pronoun *ich* as [ɪç], Leyla palatalized it into [ɪʃ], converting it into an *isch*. This palatalization is what enables her to use contractions of the first person singular pronoun with words such as *schwöre*, the singular first-person form of the verb *schwören*, "to swear". The end result is the phatic⁹ *Kiezdeutsch* expression "*ischwöre*", literally translating as "I swear", found at the end of Leyla's second utterance.

⁸ *Im* is a contraction between the two-way preposition *in* and the dative form of the masculine definite article. In German grammar, there are specific prepositions that can be followed by articles, nouns, and adjectives both in the accusative case and the dative case, with the accusative typically used in cases where A to B motion is implied and the dative case used in all other cases.

⁹ The phatic function focuses on the medium or channel of communication, serving either to establish, prolong, or discontinue communication. It is often the first verbal function acquired by infants. Here, *ischwöre* ensures that Hatice continues the discussion by questioning the verity of Leyla's claim and whether or not she can "swear" to it. (Jakobson 1960)

This phenomenon of “*ischwöre*” shows a *Kiezdeutsch* characteristic to assimilate verb expressions, so as to make them more fluid for colloquial use.

Similar shortening can easily be observed in other contexts, such as this utterance from the *Kiezdeutsch Infoportal*.

Lassma Moritzplatz aussteigen!

Let's Moritzplatz outclimb!

In Standard High German, the same utterance would be

Lass uns mal aus Moritzplatz aussteigen!

Let¹⁰ us¹¹ PTCL out Moritzplatz out+climb!

In Standard English, this would translate as,

Let's get out on Moritzplatz!

Here, *lassma* is used as a contracted expression of the Standard High German *lass (uns) mal*, or “let’s”, showing an easy route to sending an inclusive, conative¹² message (cf. Jakobson 1960) to the addressees of this communication. The speaker, in this situation on the bus, is proposing to those around him to all get off on Moritzplatz, and the *lassma* allows for a quick and easily-produced message through accordance of the *Kiezdeutsch* expression. A similar example of such a contracted, conative *Kiezdeutsch* expression is with the use of *musstu*, in a sentence such as,

Musstu Doppelstunde fahren!

Must-you Double-hour drive!

The same expression can be transposed into Standard High German as

Du musst für eine Doppelstunde fahren!

You must¹³ for a¹⁴ Double-hour drive!

¹⁰ Imperative mood form of the verb *lassen*.

¹¹ Dative first-person plural pronoun, indirect object of the dative verb *lassen*. In Standard High German, as is the case in English (cf. he, him, his), pronouns also change based on their positioning, and this positioning can be affected by the verb. As *lassen* always takes an indirect object, and thus an object in the dative case, the German first-person plural pronoun *wir* is transposed, or declined, into its dative form *uns*.

¹² In the speech theory of Roman Jakobson, the conative function indicates a speaker orientation towards the addressee.

¹³ Second person singular form of the modal verb *müssen*

¹⁴ Feminine indefinite article, direct object of the accusative preposition *für* and modifying the feminine noun *Stunde*. The accusative case is used when the noun is a direct object of the verb.

And this, in Standard English, would correspond to,

You must drive for two hours!

Here, *musstu*, a contraction of *musst du*, or “must you” shows a V-S-O word order used in a situation where it is unpreceded by an element, typical of *Kiezdeutsch* usage. The opposite is the case in Standard High German: in a clause without any preceding elements, the word order stays S-V-O, unless the clause bears an interrogative function. The fixed usage of this expression, like *lassma*, is conative, but unlike *lassma*, it is exclusionary: here, the addresser is speaking to a specific addressee, with the usage of *du*, as implied in *musstu*, exemplifying this.

Interestingly, while *Kiezdeutsch* expressions such as *musstu* often fix the familiar second-person singular pronoun *du* into daily expression, no such fixation has been observed with the honorific *Sie*, which encompasses the second-person singular, the third-personal feminine singular, and the third-person feminine plural pronouns. The reason behind this fixation is to further establish the audience of *Kiezdeutsch*. As the youth register is primarily used between adolescents, typically friends, speakers specify the addressed recipient in their youthful discourse use familiar pronouns, in order to isolate the specific audience of their utterance from the adult over-hearers of the conversation (cf. Goffman 1981).

Both *lassma* and *musstu* follow a similar pronunciation structure, as both are composed of one stressed syllable in the form of one unstressed syllable, in a trochaic format characteristic of Standard High German, viz. *lássma* and *músstu*. This regularity, combined with the ease of integrating both *lassma* and *musstu* in *Kiezdeutsch* sentences patterning German grammar, shows how well both of these expressions fit into everyday German, in a pattern widely observed in youth slang, in which slang expressions tend to follow grammatical regularities in usage (Eble, 1996). However, the distinctiveness of *lassma*, *musstu*, and *ischwöre* constructions from Standard High German patterns designates these features as stereotypical of *Kiezdeutsch* for the non-*Kiezdeutsch* speaker.

Kiezdeutsch also has a conspicuous tendency to drop prepositions and articles in sentences, with notable examples such as,

Wir sind jetzt anderes Thema

We are now other¹⁵ Theme

The same sentence, rendered in Standard High German, would be

Wir sind jetzt bei einem anderen Thema

We are now by a¹⁶ other¹⁷ Theme

¹⁵ Predicative nominative neuter form of the unpreceded adjective *ander*, modifying the neuter noun *Thema*. The nominative case is used when the noun is the subject of the verb and a predicative nominative usually happens when the noun completes a linking verb (a copula) of a sentence, in which the word would be declined as if it also were a subject.

And this would be translated into Standard English as

We are now on another theme.

The *Kiezdeutsch* speaker is dropping out the preposition “*bei*” and the indefinite article “*einem*” as both are implied within the context of the sentence, and thus, the adjective “*anderer*” is declined and modified as “*anderes*” rather than “*anderen*.” This omission of articles and prepositions is attested to in *Gastarbeiterdeutsch* and highlights the influence of the German register spoken by their parents on the register shared among second and third-generation immigrant youth (Androutsopoulos, 2001). The omission of the preposition and the article here is also a stereotypical element of *Kiezdeutsch* and is indicative of the diverse roots that came to form the register.

Despite the unique grammatical constructions in *Kiezdeutsch*, the crossover-phenomenon between immigrant and German cultures in *Kiezdeutsch* is most salient in the *Kiezdeutsch* vocabulary itself. As speakers of *Kiezdeutsch* hail from multiethnic areas, with most of them speaking both their ancestral language at home and Standard High German in formal situations, *Kiezdeutsch* tends to integrate these different linguistic influences by crafting a new vocabulary. Examples include the usage of *Çüş*,¹⁸ a Turkish word which originally meant “whoa”, used when stopping a donkey. However, in modern Turkish, it can mean “go!” or “play up!”, but also “you fool!” or “you ass!” (Paul et al. 2008). A song text, taken from the Fler album *Trendsetter*, exemplifies *Çüş* usage in stating,

Çüş Junge isch komm gar nicht mehr klar
PTCL boy I come PTCL not more clear

The same lyrics, in Standard High German, would be

Hey Junge, komme ich nicht mehr aus
Hey boy, come I not more out

And the utterance would be translated into Standard English as

Hey boy, I don't get along any more.

Another example of the multiethnic nature of *Kiezdeutsch* is the usage of the Arabic expression *wallah*, meaning “I swear by God”, as an emphasizer, such as in the sentence from the *Kiezdeutsch Infoportal*,

¹⁶ Dative neuter form of the indefinite article *ein*, object of the dative preposition *bei* and modifying the neuter noun *Thema*. German prepositions have an influence on the declination of articles, pronouns, adjectives, and occasionally nouns, and here, the preposition *bei* renders *Thema* an indirect object, necessitating the indefinite article to be thus declined.

¹⁷ Dative neuter form of the adjective *ander*, preceded by the dative neuter indefinite article *einem*

¹⁸ The IPA of this being [tʃyʃ]

Und da stand und hat mir seine Hand gegeben. Wallah.
 And there stood and has to+me¹⁹ his²⁰ hand given. By+God
 In Standard High German, this would correspond to

Und da, stand er und er hat mir seine Hand gegeben. Echt!
 And there, stood he and he has to+me his hand given. Really!

In English, the sentence would be rendered as
And there he stood, and he gave me his hand. Really!

Interestingly, *wallah* has begun to appear not only in *Kiezdeutsch* contexts, but also in mainstream German pop culture. For example, in Firas Alshater's sixteenth video in the *ZUKARStückchen* Series, "Für'n Arsch", he sings, in otherwise consistent Standard High German,

Und sauberer wallah sowieso
 And cleaner by+God anyway

Sometimes, *Kiezdeutsch* may borrow entire phrases from other languages, and sometimes, it may coin its own fixed expressions, in a characteristic of lexeme usage common to slang (Agha 2015), as in the following example,

Siktir lan, isch masch disch Messer
 Fuck off man, I make you Knife

In this utterance, the first half, "*siktir lan*" derives wholly from Turkish, showing a fixed insult expression used in the *Kiezdeutsch* register. The usage of "*Isch masch disch Messer*", a unique coinage in *Kiezdeutsch*, roughly approximates to the slang expression, "*I'll end you*" in English and shows the unique takes on German that the youths themselves make: the expression consists wholly of Germanic origin words, but with a new twist in meaning.

One of the more interesting loanword phenomena in *Kiezdeutsch* is the usage of *moruk* and *Alder*. *Moruk* directly derives from Turkish, meaning "geezer" while *Alder* is an assimilation of the word *moruk* into German,²¹ showing strong similarities to Bavarian German, which uses the word *Oider* in a similar context. Examples of such usage are easily observable in the sentence in the Bushido's album *Vendetta*,

Moruk, guck, wo isch jetzt bin, heißer wie Lava

¹⁹ Dative form of the first-person singular personal pronoun

²⁰ Accusative feminine form of the possessive adjective *sein*, modifying the feminine noun *Hand*

²¹ From the German noun *Alter*, "age", and related to the German adjective *alt*, meaning "old".

Geezer, look, where I now am, hotter like Lav

And also in,

Ischwöre, Alder, war so

I-Swear, Old man, was so.

The usage of *moruk* and *Alder* as implicit conative forms of address (cf. Jakobson 1960) may seem puzzling at first, noting the register's value boundary (cf. Agha 2015) as a register positively valorized only by the youth population. However, this ironic usage parallels youth habits of playfulness in language: many youth slang terms depend on making fun of adult language, and *Kiezdeutsch* is no exception. In fact, *Kiezdeutsch* is used by adolescents to create a sense of community by drawing a line between them and other groups (Paul et al. 2008). Intent on emphasizing the otherness of the *Kiezdeutsch* youth, parodying the adult world in *Kiezdeutsch* has the stronger role of "rebell[ing]" against the exclusionary mainstream societies they inhabit. Thus, the double lexemic crossover of both *moruk* and *Alder* embodies the internal dualities of the *Kiezdeutsch* identity in the period between adolescence and adulthood, between their migrant and their German identities.

Social Evaluations of *Kiezdeutsch*

As is the case with the majority of slang registers, *Kiezdeutsch* is subject to different value judgments and exists at a "value boundary", positively valorized by second and third-generation immigrant youth as well as some native German youth, and negatively valorized by authority figures in mainstream German society (cf. Agha 2015). This value boundary can be best observed in a study done by Maria Pohle and Kathleen Schumann from the University of Potsdam in 2012, in the Kreuzberg and Neukölln neighbourhoods of Berlin. Thirty-nine research subjects were selected from the study, all between the ages of 15 and 20 and all belonging to six different ethnic backgrounds: Arabic, Turkish, Kurdish, Albanian, Bosnian, and Serbian, as well as monolingual native German youth. The youth were placed in two informal, and then, two formal situations of communication: the participants first were asked to send a text to a friend, and then were asked to call the same friend. In the second stage, the participants were required to write a testimony and to give an oral report to a police officer. Obtaining a total of 155 oral and written productions, and after a quantitative analysis of the registers the different speakers used, the researchers found that while a little below 60% of the informal spoken and written situations exhibited features of *Kiezdeutsch*, that number dropped down to around 10% for the formal spoken situation and to nearly 0% for the formal written situation.

The intermittent usage of *Kiezdeutsch* usage in informal speech parallels a study done by anthropologist Michael Mouffatt from 1977 to 1987, in which he tape-recorded two hours of conversation between undergraduates. Throughout the pieces of recorded conversation, slang vocabulary appeared rarely, showing that slang vocabulary is not a characteristic of the intimate register, even among people who share knowledge of a similar slang repertoire (Eble 1996). The similar nature of *Kiezdeutsch* shows that like many forms of youth slang, *Kiezdeutsch* is primarily used to construe a group identity and to negotiate the individual identity within that group (Eble 2005) rather than to characterize relationships between its users. Thus, the positive speaker evaluation by *Kiezdeutsch* speakers primarily derives from the ability of *Kiezdeutsch* to create a community rather than from any expressive advantages over Standard High German.

The absence of *Kiezdeutsch* usage in formal situations also pinpoints the negative evaluations that *Kiezdeutsch* faces in mainstream German society. A part of the negative evaluation is the worry that *Kiezdeutsch* is holding back youth from integration into mainstream German culture, or the *Leitkultur*. According to the data released in 2009 and 2010 by the German Federal Statistical Office, one in five Germans has a foreign background²² approximately 31% of minors in Germany live in a family with a migrant background and in cities with over 500,000 inhabitants; 46% of children grow up in families with a migrant background (Wiese 2015). Adolescents with a migrant background are much more likely to fail at school compared to native German pupils, reaching significantly lower results in achievement tests (i.e. the PISA²²) compared to children without an immigrant background, especially on the literacy portion. Migrant youths also have a higher unemployment rate (cf. Greve and Nur Orhan 2008). Thus, much of mainstream German society views *Kiezdeutsch* as an indicator of both educational failure and an obstacle to social mobility, leading to unemployment and welfare dependence (Wiese 2014). A 2012 article on *Kiezdeutsch* in *The Economist* stated

At least a third of children with non-German roots and a tenth of those growing up in German-speaking homes do not speak standard German properly. Unless they learn they will be hard to employ and may end up on the dole. To babble in *Kiezdeutsch* instead of proper Hochdeutsch can suggest acceptance of a parasitical future. *The Economist*, 2/11/2012

A comment in a *Kiezdeutsch* article on *Bild*, a German news agency reads,

Oh, if they knew only how they mark themselves, through language, both art, and clothing, as belonging to the lowest caste. A life style at the level of minimal wage, Hartz IV [social benefits] is predetermined this way. *Bild* 2/18/2012

As a result, youth speaking *Kiezdeutsch* become associated with the cliché of the masculine Turkish youth, most likely in an aggressive posture (Wiese 2012). These evaluations parallel some of the evaluations of immigrant youth languages of other European countries: *perkersprog* in Denmark is similarly associated with masculinity and toughness, with a pan-ethnic minority “street culture” and a lack of academic prestige (Madsen 2013).

Another part of the negative evaluations targets facets of the *Kiezdeutsch* language as “bad grammar” and exemplary of “linguistic degeneration”, culminating in the evaluation that *Kiezdeutsch* is a product of the negative influence of the heritage languages of second and third-generation German immigrants, and thus, an impediment to integration in German society (Pohle et al. 2014). This belief is exemplified in an article by the linguist Helmut Glück, of the University of Bamberg, when he states, “*Kiezdeutsch* is neither a dialect nor a sociolect, but an unusual transitional language, which depends on the influences of other languages and on the faults in German...in no case is it for dialectology, but for linguistic psychology and for fault analysis.” A teacher from Berlin, regarding *Kiezdeutsch*, stated that, “They speak in shreds; the syntax is simplified. That

²² The PISA, or the Programme for International Student Assessment, is an international survey organized every three years by the Organization for Economic Cooperation and Development (OECD), to evaluate the educational systems of different countries through testing the skills and knowledge of 15-year-old students.

contributes to the decay of the German language. (cited in Wiese 2008); and Dietmar Krug, from *Die Presse*, stated that

Through the media, the worry that the German language is in danger wanders in again like a ghost. This time, the blame goes to the youth jargon in the immigration scenario. “End with Turk-speak!”, titles approximately the “Berliner Kurier” and in this thread, follows an aggressive example of this “Kanak Sprak”, as the author Feridun Zaimoğlu named this phenomenon. Something like, “I’ll hospital you!” or “Do you need tough?”

Even a high-profile German politician gave a direct negative evaluation on *Kiezdeutsch*:

I say to you, I would always interrupt them, when the two speak like this on the street, I say, you, assholes, this is the start of the end! This is how they begin to speak to one another! (Heide Simonis, former German regional prime minister of Schleswig-Holstein, in the talk show “III nach neun”, Norddeutscher Rundfunk, May 8, 1998, courtesy of Paul et al. 2008)

The phenomenon in *Kiezdeutsch* of shortening words, changing traditional German word-order patterns, palatalizing consonant sounds, crafting unconventional expressions and words, and, most frustratingly, dropping out prepositions and declining adjectives and articles differently from Standard High German reinforces the negative metadiscourse that *Kiezdeutsch* is a “Neanderthal language”, as many critics of the register state. This evaluation stands opposed to “High German”, the language of “evolution” and the heritage of the German *Dichter und Denker*, the “poets and thinkers.”

This imagination of “High German” is a cultural motif that aids in the creation of a positive German self-image, as a land of culture, with its own “high language” (Wiese 2014). This perception of “High German” as the embodiment of the German *Leitkultur* has very much to do with its origins and its usage, both historical and current. Originally, the term “High German”, or *Hochdeutsch*, was used to describe the Irminonic²³ language used by the Germanic peoples of the more mountainous regions of Central and Southern Germany, as opposed to the Ingaevonic²⁴ language used by the Germanic peoples of the flatter regions of the Northern German seaboard, *Plattdeutsch*, or “Low German.” However, in the popular mindset, the metacommunicative descriptor conveys a sense of a “higher form” of the German language, establishing a powerful case of a standard language ideology (Wiese 2014, Wiese 2015).

Hochdeutsch was initially formulated as solely a written language, used by German authors to communicate to the largest number of people possible, all speaking different Irminonic languages and dialects, with Martin Luther’s translation of the Bible and the

²³ Traditionally referring to the Irminones, the West Germanic peoples who inhabited the highlands of central Europe. Their language evolved into the various German dialects of modern day Germany, Switzerland, Liechtenstein, Luxembourg, Belgium, and Austria, as well as the indigenous German-speaking groups in Alsace, South Tyrol, Denmark, and Silesia, along with the languages used by some religious groups in North America (the Amish, the Hutterites, etc.). Yiddish is also a descendant of the Irminonic language.

²⁴ Traditionally referring to the Ingaevones, the West Germanic people who inhabited the northern European seaboard. The languages of these peoples later evolved into Low German, Low Saxon, Friesian, and English.

works of Goethe and Schiller being notable examples of the usage of this High German register. Thus, Standard High German incorporated a variety of different elements from the different Irminonic dialects and began to acquire an identity not centralized on any specific German region, but rather, on the abstract concept on a united German nation. Gradually, this standard Irminonic language grew to become a spoken language, and with German unification, Standard High German was employed as the standard register of the German language. It soon developed a mythology of being the symbol of the unified German nation and of the high literary culture associated with the nation (cf. Wiese 2015). In current times, Standard High German continues to enjoy elements of this prestigious position as the “high language.”

Many fringe right-wing linguistic protection organizations take advantage of the association of Standard High German with “German high culture.” They argue for the discontinuation of the use of “Neanderthal” or “Stone Age speech” (cf. Wiese 2015) such as *Kiezdeutsch*. Such action is exemplified by *Deutsche Sprachwelt*, a right wing “language guardian” organization participating in a “complaint tradition” (cf. Milroy et al. 1999). Quoting from page 46 from the *Alternativ für Deutschland*, a right-wing populist and Eurosceptic party in Germany, *Deutsche Sprachwelt* posts this on its Facebook page:

“The national language is the heart of a cultural nation. As a central element of the German identity, the *AfD (Alternativ für Deutschland)* desires to codify the German language as a national language in the constitution: ‘In German schools, there should be no retreat of German in the face of immigrant languages.’” *Deutsche Sprachwelt* 8/5/2017. And to one of its posts propagating the rumor that the word “Alter” is used as an all-purpose implicit emotive²⁵ form, a commentator stated that: “One can understand also that [usage] as a symptom...of decadence. The fact is, it’s not a vocabulary reduced through dementia (age), but through stupidity.” *Deutsche Sprachwelt* 5/21/2017

In 2015, linguist Heike Wiese, from the University of Potsdam for German Studies and Centre for Language, Variation, and Migration, documented all 25 of her emails after her talk on *Kiezdeutsch*, the “first wave” of emails, and all 51 emails after her publication of a book on *Kiezdeutsch* as a German dialect, the “second wave” of emails. (Wiese 2015). Many of these emails were triggered by a report on a German website from the extreme right, *pi-news* “politically incorrect news”, and most of them were either sent anonymously or were posted under nicknames, used expressively to suggest certain social roles (cf. Lindholm 2009) but not revealing the poster’s identity. As a result, commenters had the social liberty to speak their minds without fear of social judgement, allowing for many negative postings regarding *Kiezdeutsch*. In total, only 8.7% of all emails were positive, or contained no devaluations of *Kiezdeutsch* or even took a stance against *Kiezdeutsch* devaluation.

The bulk of the data she collected consisted of participants reporting negatively about *Kiezdeutsch*, with postings in four categories. The first type of negative evaluation illustrated *Kiezdeutsch* as a “broken language”, a deficient version of German. The second type saw that *Kiezdeutsch* was a sign of “language decay”, threatening the integrity of German. The third type believed that the youth register was a display of the speakers “opting out”, refusing to integrate with German society. The last negative evaluation thought *Kiezdeutsch* to be a contributor to “social demolition”, with the speakers threatening the national cohesion of Germany. Throughout her investigation, the strong negative evaluations indicated a mindset that *Kiezdeutsch* speakers, through their language,

²⁵ The emotive function is sender-focal, and it aims to directly express the speaker’s attitude towards the subject. (Jakobson, 1960)

demarcated themselves as members of an alien out-group, without mastery of *Hochdeutsch*, and thus, contributing to its decay through their “broken language.”

However, the vast majority of the German population do not frame their negative evaluations of *Kiezdeutsch* in such strong terms. Rather, they express their perceptions of the register through mass-media propagated comedy, with actors parodying *Kiezdeutsch* speakers to convey a similar sense of criminality, aggression, masculinity, educational deficiency, and a certain “outsider-ness”.

These depictions are perhaps best exemplified by *Mundstuhl*, a Frankfurt-based native German comedy duo playing the “Turkish immigrant youth” Dragan and Alder (Androutsopoulos 2001). In an episode from *Comedy Tower* from December 12, 2014, the characters Dragan and Alder star in *Dragan und Alder machen freiwilliges asoziales Jahr*, which, translated literally, means “Dragan and Alder make voluntary antisocial year.” The usage of “*asozial*” implies the outsider perception that *Kiezdeutsch* speakers face from the “*sozial*”, the German world. The absence of a preposition here, as well as the unique grammar, mark this as a title parodying speech patterns among *Kiezdeutsch* youth. This grammatical parody also indirectly implies their inability to use German correctly, a sentiment further echoed in a part of the video in which Dragan has difficulty articulating.

While Alder tells about his experiences being sentenced to community service for throwing a hand-grenade into a bar because it charged him 26 Euros for vodka, he stumbles on one of the long German words denoting a government work office, emphasizing again his foreign identity in “not being German enough” to correctly pronounce the names of the German institutions. Interestingly, as Dragan described throwing the hand-grenade into the bar, Alder responded by stating that it was a “normal reaction”, accentuating the stereotype of the *Kiezdeutsch* youth not only being oriented to violence and crime, but also being victims of endemic poverty. The association between *Kiezdeutsch* and petty crime is furthered later in the video, when Alder describes how he steals gas from cars: he would go to the parking lot of a movie theater and siphon off the gas while the moviegoers were in the theater, “sucking the entire car dry” during the ninety minutes of time he has. When asked by Dragan whether or not he drinks down the gasoline and barfs it into a container, Alder responds that while he did that before, today, he is a “professional”: he would use a long hose and a large enough gasoline container, suck the gasoline until it almost came up to his mouth, and immediately drop the hose into the gasoline container. Afterward, Alder went into a rapid-fire discourse about the physics and the principles of collecting the gasoline, garnering the cheers of the audience.

The description of Alder committing a petty crime serves only to reinforce the popular mentality about the criminality associated with *Kiezdeutsch*, but also a certain streetwise-ness accompanying life “in the ghetto.” The “Neanderthal” stupidity stereotypically attributed to *Kiezdeutsch* speakers is shown through Alder drinking in the gasoline and barfing it into a gasoline container, as well as Alder dropping the hose into a container as a mark of his “professionalism.” The audience’s applause when Alder later goes into a seemingly intelligent discourse only further shows the depth of the popular association of *Kiezdeutsch* with stupidity. Throughout the entirety of the video, the phrase “*Alder was geht*” is repeated multiple times, fixing the discourse between the two actors as a unique *Kiezdeutsch* one, and a simple look at the body language of both *Mundstuhl* characters also shows their stereotypical depiction of *Kiezdeutsch* youth. Both of them have their hands in their pockets and lean slightly backwards, sending a certain metacommunicative signal displaying their societal indifference and a masculinity associated with their “ghetto lifestyle.” The stereotype about the hypermasculinity displayed by the *Kiezdeutsch* youth can be best seen near the beginning of the video: after Alder tells

Dragan about his new favorite drink, Dragan mocks him by asking whether or not he has “turned homo”, sarcastically questioning whether or not he has a *Biene Maya*²⁶ bedspread and a Hello Kitty lunchbox. Dragan then tells Alder that the “real man’s drink” is vodka, which he frequently consumes. This questioning shows the perception of hypermasculinity of *Kiezdeutsch* youth: while they are children in reality, they act like adults, questioning the traditionally “German” children’s activities and drinking hard alcohol as proof of their masculinity.

The capitulation of the comedy industry on *Kiezdeutsch* stereotypes highlights how negative evaluations are propagated among *Kiezdeutsch* speakers. Much of these negative evaluations are without base and play on racist undertones, otherwise unacceptable in mainstream society. With discrimination based on color, race, religion, or gender being considered socially unacceptable, language has become one of the final avenues in which racism is still considered socially acceptable (Wiese 2015, Paul et al. 2008). The acceptability of racism in modern linguistic discourse stems from the pedagogic tradition of using linguistically conservative principles to teach children how to communicate in the Standard variety of the language, the variety that would allow the speaker to have the most opportunities. Through acts based on linguistically conservative principles, such as insisting on native vocabulary as opposed to loanword vocabulary, teachers integrate students into the grammatical and lexical fabric of the language. Using linguistic conservatism can be pedagogically useful. Thus, the widespread usage of linguistic conservatism allows it to inevitably be an acceptable avenue for the expression of uglier sentiments, such as the belief that those speaking a certain register are of “inferior intelligence.”

Currently, there are many measures against the well-established negative evaluations of *Kiezdeutsch*. German linguists, such as Maria Pohle and Heike Wiese, are leading efforts to increase *Kiezdeutsch* awareness through extensive documentation of *Kiezdeutsch* phenomena and speaking about *Kiezdeutsch* in popular media channels. Most German news media are beginning to report on *Kiezdeutsch* as a new register in the process of growth rather than a degenerate variety of Standard High German. German rappers often rap in *Kiezdeutsch*, promoting an increase in positive evaluations of *Kiezdeutsch* by youth. However, *Kiezdeutsch* is still quite far from the prestigious position that regional German dialects, such as Bavarian and Swabian, enjoy. Currently, mainstream Germany still views *Kiezdeutsch* with a suspicious eye, and only time may wash the populace of their negative perceptions.

Bibliography

Agha, Asif. "Registers of Language." *A Companion to Linguistic Anthropology*: 23-45. doi:10.1002/9780470996522.ch2.

Agha, Asif. "Tropes of Slang." *Signs and Society* 3, no. 2 (2015): 306-30. Accessed September 5, 2017. doi:10.1086/683179.

Androutsopoulos, Jannis. "From the Streets to the Screens and Back Again: On the Mediated Diffusion of Ethnolectal Patterns in Contemporary German." *LAUD Linguistic Agency*, June 29, 2001, 1-24. Accessed September 1, 2017. https://jannisandroutsopoulos.files.wordpress.com/2009/09/iclave_2001_laud.pdf.

²⁶ In German popular culture, *Biene Maja*, or “Maya the Bee” is a popular German children’s cultural icon. First featured in Waldemar Bonsels’ *Die Biene Maja und ihre Abenteuer* (Maya the Bee and her Adventures), *Biene Maja* grew to be incorporated in film, theater, comic strips, and most famously, a TV series. For the Germans, *Biene Maja* grew to represent childhood memories.

- Beatlefield. "Bushido (Ft. Chakuza & Eko Fresh) – Vendetta." Genius. December 1, 2006. Accessed September 5, 2017. <https://genius.com/Bushido-vendetta-lyrics>.
- Botica, Melania. „Machst Du Rote Ampel?" FOCUS Online. September 09, 2015. Accessed September 6, 2017. http://www.focus.de/familie/schule/unterricht/machst-du-rote-ampel-kiezdeutsch_id_2444004.html.
- "Deutsche Sprachwelt." Deutsche Sprachwelt. Accessed September 5, 2017. <http://deutschesprachwelt.de/>.
- Eble, Connie. *Slang & Sociability: In-Group Language among College Students*. Chapel Hill, NC: Univ. of North Carolina Press, 2003.
- Fiedler, Gregor. "Deutsche Sprachwelt." Facebook - Log In or Sign Up. May 21, 2015. Accessed September 14, 2017. <https://www.facebook.com/deutschesprachwelt/>.
- Fler. "Fler (Ft. Muhabbet) – Çüs Junge." Genius. June 23, 2006. Accessed September 8, 2017. <https://genius.com/Fler-cus-junge-lyrics>.
- Goffman, Erving. "Footing." *Forms of Talk* 59, no. 2 (1981): 124-59.
- Gun, Alpa. "Alpa Gun – Intro (Almanci)." Genius. Accessed September 4, 2017. <https://genius.com/Alpa-gun-intro-almanci-lyrics>.
- Heine, Matthias. "Kiezdeutsch Ist in Wahrheit Rassistisch - WELT." DIE WELT. July 04, 2014. Accessed August 31, 2017. <https://www.welt.de/kultur/article129622721/In-Wahrheit-ist-Kiezdeutsch-rassistisch.html>.
- Jakobson, Roman. *'The Speech Event & the Functions of Language'*. 1960.
- "Kiezdeutsch | Alltagsdeutsch – Lektionen | DW | 14.02.2017." DW.COM. Accessed August 14, 2017. <http://www.dw.com/de/kiezdeutsch/1-19516144>.
- 'Kiezdeutsch Und Wann Man Es Benutzt' Maria Pohle Beim #40 Science Slam Berlin. Performed by Maria Pohle. Science Slam Deutschland. April 17, 2015. Accessed September 14, 2017. <https://www.youtube.com/watch?v=9emFFPr4sMs>.
- Krug, Dietmar. "Kiezdeutsch." Die Presse. February 09, 2013. Accessed September 10, 2017. <http://diepresse.com/home/meinung/diesedeutschen/1342946/Kiezdeutsch>.
- Madsen, Lian Malai. "'High" and "low" in Urban Danish Speech Styles." *Language in Society* 42 (September 28, 2012): 115-38. Accessed August 17, 2017. doi:10.1017/S0047404513000018.
- Matthes, Frauke. "Was Deutsch Ist, Bestimmen Wir": Definitions of (Turkish-) Germanness in Feridun Zaimoglu's Kanak Sprak and Koppstoff." *Focus on German Studies* 14 (2007): 19-31. Accessed September 1, 2017. <https://drc.libraries.uc.edu/bitstream/handle/2374.UC/1993/Matthes%2c%20Frauke%3b%20Definitions%20of%20%28Turkish-%29%20Germanness%20in%20Feridun%20Zaimoglu%27s%20Kanak%20Sprak%20and%20Koppstoff.pdf?sequence=1>.
- Miersch, Michael. "Kommentar: Kiezdeutsch Ist Keine Katastrophe – Sischer! - WELT." DIE WELT. October 16, 2012. Accessed August 31, 2017. <https://www.welt.de/debatte/kommentare/article6074696/Kiezdeutsch-ist-keine-Katastrophe-sischer.html>.
- Mundstuhl: Dragan Und Alder Machen Freiwilliges Asoziales Jahr - Comedy Tower*. Performed by Ande Werner and Lars Niedreichholz. Hrfernsehen. November 3, 2014. Accessed September 14, 2017. <https://www.youtube.com/watch?v=5zM-WIEfL1w>.
- OECD. *About - PISA*. Accessed November 20, 2017. <http://www.oecd.org/pisa/aboutpisa/>.
- Paul, Kerstin, Ulrike Freywald, and Evan Wittenberg. "'Kiezdeutsch Goes School"-A Multi-Ethnic Variety of German From an Educational Perspective." *Proceedings of the 1st International Conference on Linguistic and Intercultural Education*, 2008, 636-82.

Accessed August 31, 2017.

http://evawittenberg.com/i/publications_files/B6_Paul_Freywald_Wittenberg_draft_Dec-2008.pdf.

"Particle." Teaching English | British Council | BBC. Accessed November 20, 2017. <https://www.teachingenglish.org.uk/article/particle>.

Pohle, Maria, and Kathleen Schumann. "Keine Angst Vor Kiezdeutsch! Zum Neuen Dialekt Der Mutlikulti-Generation." *Zeitschrift Für Jugendkriminalrecht Und Jugendhilfe*, 2014, 216-24. Accessed September 1, 2017. https://www.uni-potsdam.de/fileadmin01/projects/dspdg/Publikationen/ZJJ_Publikation_Pohle_Schumann.pdf.

Wiese, Heike. *Kiezdeutsch Ein Neuer Dialekt Entsteht*. München: C.H. Beck, 2012. Accessed August 31, 2017. http://www.beck-shop.de/fachbuch/zusatzinfos/Leseprobe_Kiezdeutsch.pdf.

Wiese, Heike. "Nicht Nur „Kanak Sprach“." Kiezdeutsch - Ein Infoportal Zu Jugendsprache in Wohngebieten Mit Hohem Migrantenanteil: Informationen Für Interessierte Und Handreichungen Für Schule. Accessed August 23, 2017. <http://www.kiezdeutsch.de/nichtnurkanak.html>.

Wiese, Heike. "Sprachliche Innovation in Kiezdeutsch." Kiezdeutsch - Ein Infoportal Zu Jugendsprache in Wohngebieten Mit Hohem Migrantenanteil: Informationen Für Interessierte Und Handreichungen Für Schule. Accessed August 23, 2017. <http://www.kiezdeutsch.de/sprachlicheneuerungen.html#so>.

Wiese, Heike. "'This Migrants' Babble Is Not a German Dialect!": The Interaction of Standard Language Ideology and 'us'/'them' Dichotomies in the Public Discourse on a Multiethnicity." *Language in Society* 44 (2015): 341-68. Accessed August 31, 2017. doi:10.1017/S0047404515000226.

Wiese, Heike. "Voices of Linguistic Outrage: Standard Language Constructs and the Discourse on New Urban Dialects." *Working Papers in Urban Language and Literacies*, 2014, 2-20. Accessed August 31, 2014.

<https://www.kcl.ac.uk/sspp/departments/education/research/Research-Centres/ldc/publications/workingpapers/the-papers/WP120-Wiese-2014-Voices-of-Linguistic-Outrage.pdf>.

ZUKAR 16 - Für'n Arsch. Directed by Firas Alshater. Performed by Firas Alshater. Zukar. April 11, 2017. Accessed August 31, 2017.

https://www.youtube.com/watch?v=rP9aF4shGoE&index=1&list=PLi6_IV6u95P81Tv-fi8IgX03IKNJISzEG.



Mangrove Forests Mitigate Tsunami Hazard and Require Conservation

Qiqing Li

Author Background: Qiqing Li grew up in China and currently attends Shenzhen Middle School in Shenzhen, China. Her Pioneer seminar topic was in the field of environmental science and titled "Understanding Earthquakes and the Problem of Earthquake Prediction."

Abstract

Mangroves have many ecological values that are often overlooked, among them, wave attenuation in coastal cities. This paper emphasizes mangroves' capacity for tsunami mitigation, analyzes determinants of this function, offers plantation suggestions to city planners, and underlines the necessity of mangrove conservation. Research is based on existing books, journals, articles, and other online materials. Studies conducted by different scholars are compared, analyzed, and synthesized. It is revealed that forest band width of mangroves plays a critical part in tsunami mitigation. Other determinants of mangroves' tsunami attenuation function include forest density, tree age, tree species, to name some of the most important. In coastal mangrove plantation design, planners should take the above factors into account, while modifications based on specific geographical and climatic elements are also required. Mangroves are ideal coastal vegetation even in cities without tsunami threats. They offer a wide range of ecological services, but they are currently dying out rapidly due to artificial and natural reasons. It is urgent and necessary for governments, experts, and civilians to conserve mangrove forests.

Keywords: mangrove forests, tsunami, wave reduction, conservation

1. Introduction

Mangroves are highly productive forests consisting of a small group of trees and shrubs that have adapted to the harsh intertidal zones in tropical and sub-tropical areas. The name "mangrove" probably comes from the combination of Portuguese "mangue", Spanish "mangle", and English "grove" (Webster). "Mangrove" refers to both the swamp ecosystem and the individual trees, and the latter are rare species throughout the world, due to their great reliance on tropics and a few warm temperate regions. Along wetter coastlines, deltaic and estuarine areas, mangroves reach their greatest abundance and diversity. They are found in 123 countries and territories, and cover a total of 152,000 square kilometers (equivalent to half the land area of Philippines) globally (Spalding *et al.*, 2010).

Mangroves have developed a series of physiological structures and functions in order to survive in intertidal areas. They typically exclude salt from their xylem to handle changing salinities, and employ various roots, including stilt roots, pneumatophores, knee roots, and buttress roots, to transport oxygen in waterlogged and anaerobic soils. To prevent their offspring from getting flushed away by tides, major families of mangroves have

developed vivipary, in which “parent trees release growing plants rather than simple seeds of fruits” (Spalding *et al.*, 2010). Seedlings can thus thrust into the soil and stand the waves.

Although mangroves are rare species, their ecological services are non-negligible. Apart from providing direct forest products, mangroves also support fisheries, reduce carbon emission, purify water, and create enormous economic value (Priya *et al.*, 2010; Spalding *et al.*, 2010). Their ability to protect coastal land from tsunamis and wave disturbances is discussed exhaustively in section 2.

Unfortunately, global mangrove areas are declining rapidly at about 1% every year (Priya *et al.*, 2010). Conservation measurements ought to be undertaken to prevent further reduction of this valuable species.

2. Mangrove Forests’ Ability to Mitigate Tsunamis

2.1. Functions and Effects of Tsunami Mitigation

Mangrove forests play a pivotal role in coastline protection. In regions that suffer from tsunamis, typhoons, and storms, they protect shorelines from erosion and ensure human safety.

In their study conducted in 2005, Harada and Imamura further divided mangroves’ function of tsunami mitigation and showed that mangrove forests can: (1) act as a natural barrier, (2) stop drifts, (3) reduce the energy of a tsunami, and (4) save lives. Mangrove forests collect wind-blown sand and form dunes, which dissipate tsunami energy as the water spends energy to pass through the dunes. Therefore, waves will not rush directly onto the shore. When waves rush through mangrove forests, the vegetation can stop drifts from flooding inhabited areas and reduce the inundation area and flow current, mainly through friction, thus reducing the possibility and intensity of property damage. Mangroves also help prevent residents from being washed away by tsunamis, for people can cling to trees for survival. These four functions are shown in Figure 1.

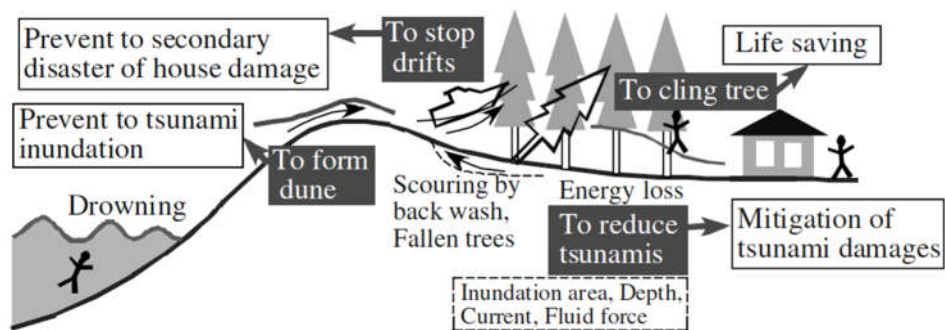


Figure 1. Functions and effects of coastal control forest to prevent tsunami disaster. <taken from Harada and Imamura (2005)>

While mangrove forests’ functions and effects are roughly depicted in Figure 1, their effect during the 2004 Indian Ocean tsunami is revealed in Figure 2 below.

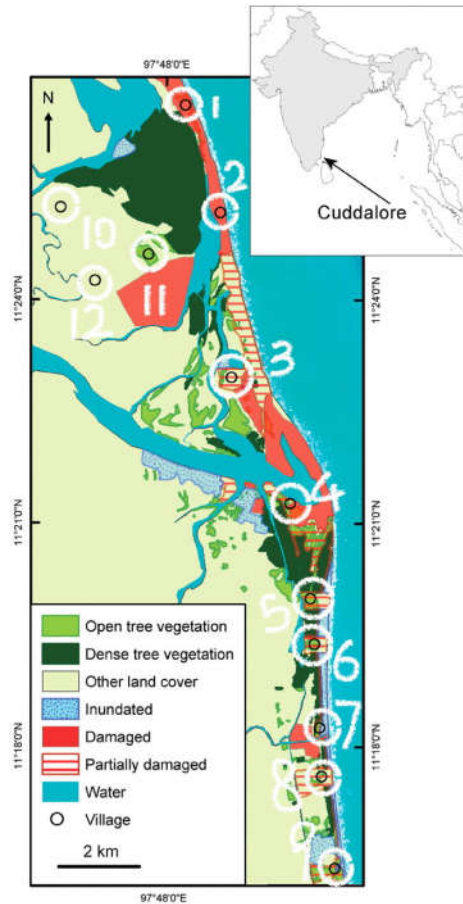


Figure 2. Pre-tsunami tree vegetation cover and post tsunami damages in Cuddalore District, Tamil Nadu, India. Villages are numbered 1 to 12 for convenient description. <taken and modified from Danielsen *et al.* (2005)>

Figure 2 depicts how coastal tree vegetation shielded villages in Cuddalore, a southern Indian district, from severe damage during the 2004 Indian Ocean tsunami, and Danielsen *et al.* (2005) point out that mangrove forests are the most important coastal tree vegetation in this area. According to Figure 2, villages marked 1 and 2, located on the northern coast, suffered severe damage because there was no protection by vegetation. Villages marked 3 to 9 were located nearby or behind either open (sparse) tree vegetation or small areas of dense tree vegetation. As a result, they only suffered partial damage. Meanwhile, villages marked 10 to 12 were the most fortunate, as they were located behind a large area of dense tree vegetation and therefore remained intact, despite the tremendous impact of the tsunami. The large damaged area adjacent to the village marked 11 sets a great contrast with the three intact villages, clearly illustrating the protective function of coastal tree vegetation, in this case provided by mangrove trees. Additionally, this damaged area negates the possibility that villages marked 10 to 12 survived intact due to long distances from the coast, and reinforces the powerful function of mangrove forests.

2.2. Determinants of Successful Hazard Mitigation by the Mangrove Forests

The mangroves' ability to mitigate tsunamis varies along several dimensions, such as forest band width, forest density, mangrove species, etc. Each factor contributes to damage mitigation on a different scale. Blankespoor, Dasgupta, and Lange (2016) argue that forest width, which is the distance between the front and back of the forest, is one of the main factors that affect the extent of wave height decline. Forest width will be examined closely in the following section.

2.2.1. Forest Width

In his study conducted in coastal Vietnam, Bao (2011) discovered that wave height decreases exponentially with the increase of cross-shore distance (i.e. mangrove band width) as waves rush through coastal forests, and the relationship is significant, as shown in Figure 3.

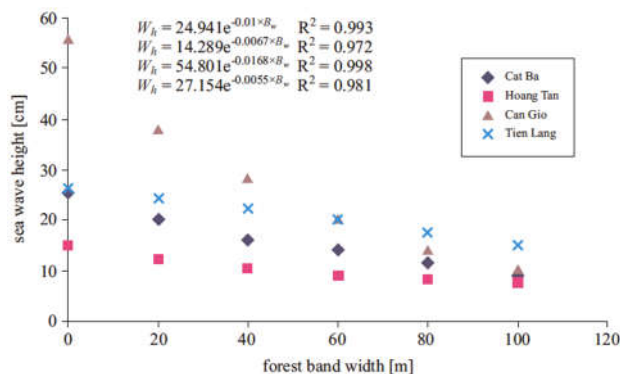


Figure 3. The reduction of wave height by cross-shore distances. B_w is the forest band width (m), W_h is the sea wave height behind the forest band (cm), and the four types of plots refer to four locations where the data were collected. <taken from Bao (2011)>

The influence of forest band width (B_w) is most obvious in Can Gio, where the initial wave height of around 55 cm decreased to around 10 cm after passing through the mangrove forest with a 100-meter width. Although the effects are not as obvious in the other three locations, the wave height (W_h) still decreases exponentially with the increase of forest band width (B_w).

Bao (2011) derived an equation to depict the relationship between wave height (W_h) and forest band width (B_w) in a more direct way, based on all the data he collected on the Vietnamese coast:

$$W_h = a \times e^{b \times B_w} \quad (1)$$

In Equation (1), B_w and W_h have the same definitions as mentioned above, while the coefficient a is directly proportional to the initial wave height, and coefficient b is always a negative, which decreases with increasing value of canopy closure (%), tree height (m), and forest density ($\text{tree} \cdot \text{ha}^{-1}$) as a function of forest structure. When the forest band width (B_w) and the forest structure (represented by b) are fixed, the final wave height (W_h) is directly proportional to the initial wave height (represented by a). Otherwise, the absolute value of negative coefficient b increases along with increases in tree height, forest density, and percentage of canopy closure, and the absolute value of $b \times B_w$ rises more quickly with the growing forest band width (B_w), which is a more important determinant

than forest structure. As a result, the wave height drops exponentially in response to the increase in forest band width, and improvement of forest structure plays a less significant part as well.

In another study carried out by Harada and Imamura (2005), the effect of forest width on coastal forests' ability to mitigate tsunami damage is examined using a numerical simulation. To target the influence posed by forest width, all the other variables are fixed, based on simulation evaluations and data collected from coastal forests in Japan, with forest density equal to 30 trees/100 m², trunk diameter equal to 0.15 m, initial tsunami height equal to 3.0 m, and tsunami period equal to 10 min. The initial tsunami height is selected to be 3.0 m because tsunamis with heights exceeding 4.0 m can damage trees (Shuto, 1987; as cited in Harada and Imamura, 2005), therefore posing difficulty in conducting numerical simulations. Since mangrove forests are regarded as an important species of coastal vegetation, and they are actually found in Japan, mostly on "the long chain of the Nansei Islands up to the very southern tip of the larger island of Kyushu" (Spalding *et al.*, 2010), it is reasonable to assume that the result of this numerical simulation, which focuses on coastal vegetation, could be applied to mangrove forests in Japan.

Harada and Imamura's 2005 study adopts three standards to measure the intensity of tsunamis: inundation depth, current, and hydraulic force (a product of fluid density, the inundation depth and the square of current velocity, which is used for the estimation of house damage), all of which are represented by "value" in the equation below, based on the target of analysis. The authors define reduction rate for a convenient description of experiment results using the following equation:

$$r (\text{reduction rate of maxvalues}) = \frac{(\text{max. value with forest})}{(\text{max. value without forest})} \quad (2)$$

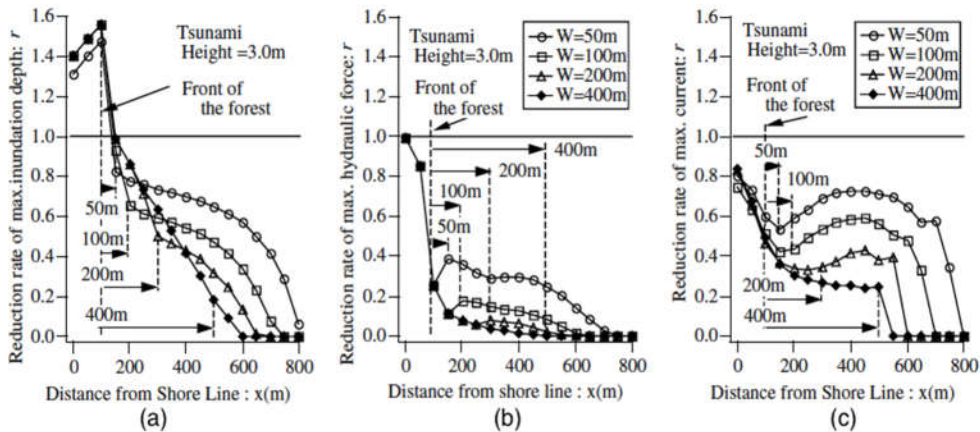


Figure 4: The effects of forest band width on the reduction of inundation depth, current, and hydraulic force. (Forest density: 30 trees/100 m², Trunk diameter: 0.15 m, Tsunami height: 3.0 m, Tsunami period: 10 min). <taken and modified from Harada and Imamura (2005)>

First, inundation depths decrease as a function of the forest width. It is recorded that the reduction rates of inundation depth just behind the forest are decreased from $r=0.86$ with forest width 50 m to 0.18 with forest width 400 m, which is shown in Figure 4 (a).

When the forest width is 50 m, the attenuation mechanism with or without forests does not show a lot of deviation, as indicated by $r=0.86$, a value close to 1, whereas a 400 m-wide forest remarkably reduces the inundation depth to 18% of water depth without forest. As for current, which is decreased due to the reflection at forest front, and reduced by reflection and energy loss due to passing through the forest, similar tendencies can be observed. The reduction rates of the current just behind the forest decreased from $r=0.54$ with forest width 50 m to 0.24 with forest width 400 m, as portrayed in Figure 4 (c). The effect of forest width on current mitigation is not as significant as that on inundation depth, while the difference between $r=0.54$ and $r=0.24$ is still large enough to highlight the function of coastal forests. Moreover, the tendencies of four lines underscore the fact that only forests with width of 400 meters and above can prevent all the damage from the current: other forests whose widths were 50 m, 100 m, and 200 m all saw increases in current after the tsunami passed through them in the simulation. Regarding hydraulic force, the reduction rates dropped from $r=0.48$ with forest width 50 m to almost zero with forest width 400 m, revealing the efficacy of mangrove forests once more. The force still increased after passing forests with widths of 50 m, 100 m, and 200 m, but the increments were insignificant for the latter two, and much less significant compared to the current scene.

As a result, both studies emphasize that forest band width plays an important part in mitigating potential tsunami damage, based on the data collected either from real-life observation or from numerical simulation.

2.2.2. Forest Density

Dasgupta, Shaw, and Abe (2014) claimed in their study that a mangrove density of 30 trees/100 m² in a 100-meter wide belt may reduce the maximum tsunami flow pressure by more than 90%. The forest width has been examined in section 2.2.1, and the forest density will be examined in this section. Following the analysis of forest band width, Harada and Imamura (2005) continued exploring the influence of forest density. They illustrated their findings in Figure 5.

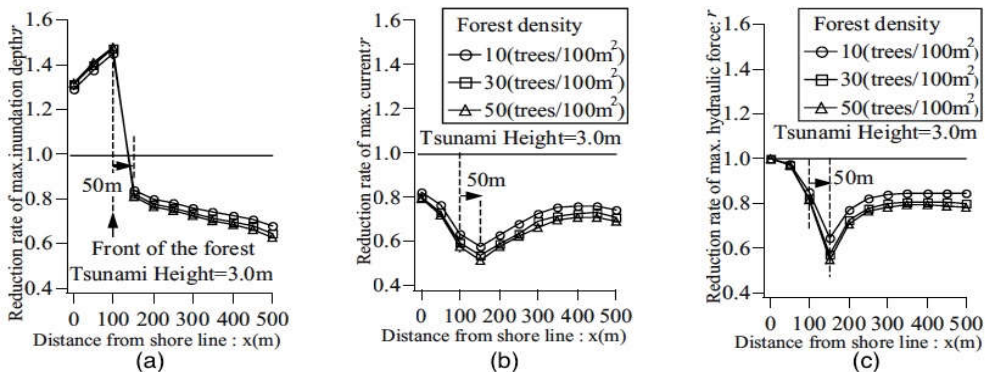


Figure 5. Spatial distribution on reduction rate of inundation depth, current and hydraulic force (Forest width: 50 m, Tsunami height: 3.0 m, Tsunami period: 10 min). <taken and modified from Harada and Imamura (2005)>

Differing from the significant influence of forest band width on damage mitigation, the lines in each of the three figures in Figure 5 almost overlap with each other, indicating only slight functional differences between forests of different densities. However, the effect of the forest density on tsunami damage mitigation was shown to be non-negligible in the

study by Danielsen *et al.* (2005), which focused on the 2004 Indian Ocean tsunami. The result of the study is presented in Table 1 below.

Table 1. Tsunami damage distribution on tree vegetation cover within 1000 m of the shore, expressed in area (a) and in proportion of the total area at every vegetation density (b).
<taken from Danielsen *et al.* (2005), supporting online material>

(a)	Damaged area (ha)	Partially damaged area (ha)	Undamaged area (ha)	Total (ha)
Dense tree vegetation	2.2	15.7	437.1	455.0
Open tree vegetation	30.9	84.4	86.0	201.3
No trees vegetation	502.9	384.2	547.0	1434.1
(b)	Damaged area (%)	Partially damaged area (%)	Undamaged area (%)	Total (%)
Dense tree vegetation	0.5	3.5	96.1	100.0
Open tree vegetation	15.4	41.9	42.7	100.0
No trees vegetation	35.1	26.8	38.1	100.0

Again, the authors point out that mangrove forests are the most important coastal forests in the study area. Therefore, the data above is applicable to mangroves. In the comparison between dense tree vegetation and open (sparse) tree vegetation, the dense one resulted in 96.1% of undamaged area, whereas only 42.7% of the area with open (sparse) tree vegetation remained intact. Clearly, denser tree vegetation was more than twice as effective as open tree vegetation, which means it provides much more protection than no vegetation at all.

The significant discrepancy between these two studies may be attributed to the following two factors:

1. Difference in wave heights. Harada and Imamura's 2005 study employed numerical simulation instead of direct observation, which was conducted by Danielsen *et al.* (2005). There were various limitations in the simulation process. For instance, the model was difficult to render if the wave height exceeded 4 m in the simulation. According to Gibbons and Gelfenbaum's study conducted in 2005, however, the actual wave height in the 2004 Indian Ocean tsunami managed to reach 30 m, and the heights may have ranged from 15 to 30 m along a 100-km stretch of the northwest coast in Sumatra. When the wave is as low as 3 m, forest density fails to display an obvious influence on tsunami damage mitigation. In contrast, the effect becomes much more pronounced when an actual powerful tsunami takes place. The discrepancy in wave heights could have potentially led to different evaluations of the impact of forest density.

2. Difference in vegetation species. Harada and Imamura (2005)'s study derived data used for simulation from coastal forests in Japan, whereas Danielsen *et al.* (2005) based their analysis on coastal forests in India. The geographical and climatic factors of these two places vary from each other, thus possibly leading to the divergence in the evaluation of the influence of forest density.

Regardless of the discrepancy, it is a generally accepted fact that the denser the coastal forests, especially mangrove forests, the stronger mitigating power they have.

2.2.3. Tree Age

After examining two macroscopic properties of mangrove forests—forest width and forest density—the focus will now shift to the microscopic attributes of mangrove vegetation: tree age. In 1997, Mazda, Magi, Kogo, and Hong derived the conclusion that a greater degree of mangrove growth leads to a greater ability to mitigate waves. They selected the Tong King delta in Vietnam as the experiment site, where they planted mangroves of different degrees of growth in an ordered layout, as shown in *Figure 6*.

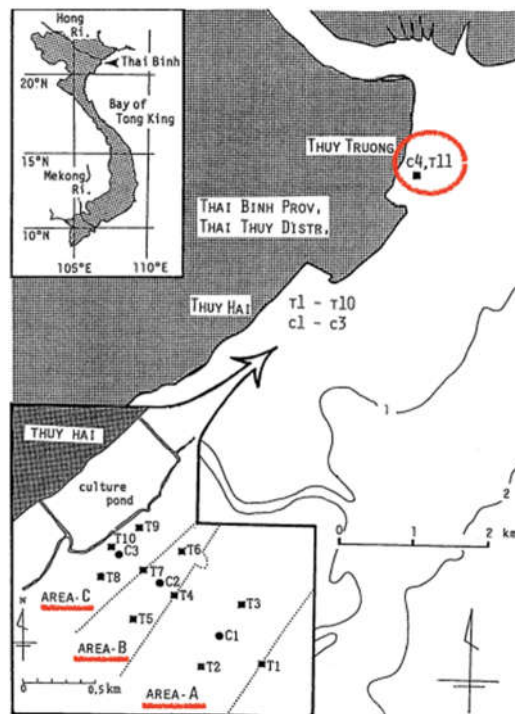


Figure 6. Location map and field sites of the Vietnam mangrove study. <taken and modified from Mazda *et al.* (1997)>

According to Mazda *et al.* (1997), half-year-old mangroves were planted in Area-A, depicted in Figure 6, trees between the ages of 2 and 3 were planted in Area-B, and trees between 5 and 6 years old were planted in Area-C. The trees planted in C₄ and T₁₁ on the coast of Thuy Truong were planted only two months before the study was conducted. “The three areas on coast Thuy Hai were continuous from offshore to an artificial dyke. Water levels were measured at 11 stations (Stns T₁ to T₁₁ as indicated in Figure 6) together with

current velocities 2 cm above the sea floor measured at 4 stations (Stns C₁ to C₄ as indicated in Figure 6)” (Mazda *et al.*, 1997).

Differing from Harada and Imamura (2005), who described stronger mitigation ability with a lower reduction rate, the definition of reduction rate given by Mazda *et al.* (1997) is given by Equation (3):

$$r = \frac{H_S - H_L}{H_S} \quad (3)$$

In the equation, H_S is the wave height at a seaside station, and H_L is the wave height at the station 100 m further inshore. When the attenuation effect is significant, the difference between H_S and H_L is large, giving rise to an r value close to 1. In other words, the stronger the mitigation effect, the larger r is. The relationship between wave height and reduction rate among plants with different ages is depicted in Figure 7.

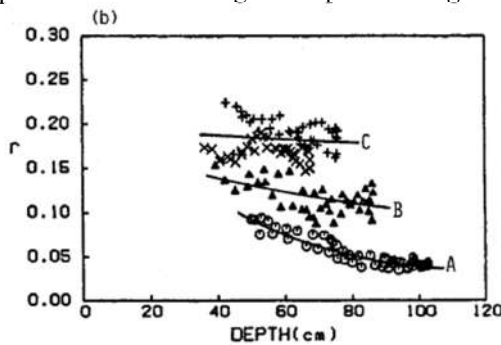


Figure 7. Variation of the wave reduction r with the water depth h . The smooth lines are suggested best-fits. A, B, and C refer to Area-A, Area-B, and Area-C, respectively, underlined in Figure 6. <taken from Mazda *et al.* (1997)>

For mature plants—those planted in Area-C—the reduction rate does not decrease with the increase of water depth, signifying these plants’ capacity for maintaining their wave-attenuation mechanism despite changes in water depth. Nevertheless, reduction rate decreases for young plants—those planted in Area-B and Area-A, showing that their wave-attenuation mechanism is not powerful enough to ignore the impact of water depth increase. The reduction rate of mangroves in Area-A becomes so insignificant that the r value falls to around 0.04 when the water depth is equal to or higher than 90 cm, indicating that younger plants do not have a significant mitigating effect. The reduction rate for trees in Area-C at 90-cm depth, however, amounts to around 0.18. Additionally, the reduction rates are different at the initial water depth for trees in each area, as plotted in Figure 7, when the water is relatively shallow at around 45 cm: around 0.19 for trees in Area-C, 0.14 for those in Area-B, and 0.10 for those in Area-A.

Different wave-attenuation ability between trees of different ages could be measured on the dimension of drag force as well. Mazda *et al.* (1997) used the resistant coefficient C_d to describe the drag force in their study. The larger the value of C_d , the greater the drag force is. The results are plotted in Figure 8.

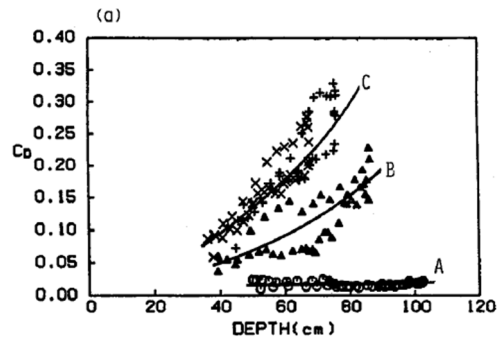


Figure 8. Variation of the resistance coefficient C_d with the water depth h . The smooth lines are suggested best-fits. <taken from Mazda *et al.* (1997)>

Pertaining to seedlings—mangrove trees planted in Area-A—the drag force indicated by C_d remains insignificant and stays below 0.05 as the water depth changes. For mature plants in Area-C and Area-B, the drag forces increase with the increase of water depth, highlighting the fact that older trees have a significant ability to mitigate waves. The C_d values for trees in Area-C soar from around 0.07 at 33 cm to around 0.33 at 83 cm. The increase for Area-B is not as significant, but it is still detectable as 0.05 at 40 cm to 0.20 at 90 cm.

In short, mangroves' wave-attenuation ability strengthens as they grow older. In fact, as claimed by Mazda *et al.* (1997), "six-year-old mangrove trees strips 1.5 km wide will reduce 1 m high waves at the open sea to 0.05 m at the coast, whereas without mangroves the wave will arrive on the coast at 0.75 m." The two scenarios are portrayed in Figure 9 below.

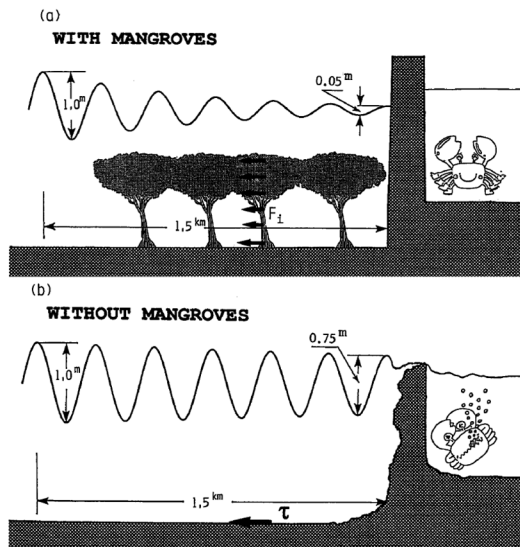


Figure 9. The effect of wave reduction with and without mangroves. <taken from Mazda *et al.* (1997)>

2.2.4. Other Factors

Based on Alongi's study in 2008, the extent to which mangroves offer significant protection of shorelines depends on the type of environmental setting, the intensity of disturbance, and the internal properties of mangroves. To wit, there are a lot more factors that influence mangrove forests' ability to mitigate tsunamis.

Alongi (2008) summarized various modelling and mathematical studies (Brinkman *et al.*, 1997; Mazda *et al.*, 1997, 2006; Massel *et al.*, 1999; Quartel *et al.*, 2007) which indicate that the magnitude of the energy absorbed by mangrove forests "strongly depends on forest density, diameter of stems and roots, forest floor slope, bathymetry, the spectral characteristics (height, period, etc.) of the incident waves, and the tidal stage at which the wave enters the forest." He also invoked Tanaka, Sasaki, Mowjood, Jinadasa, and Homchuen's 2007 studies that showed another important factor: vegetation type. For instance, when the percentage of forest floor area covered by either prop roots or pneumatophores varies, the drag coefficient of these structures varies as well, since the drag coefficient is related to the Reynolds number, which differs for each species depending on diameter and aboveground root architecture. Alongi (2008) then presented Tanaka *et al.* (2007)'s work, which modeled the relationship of species-specific differences in drag coefficient and in vegetation thickness with tsunami height; the result is shown in Figure 10. The study found that "species differed in their drag force in relation to tsunami height, with the palm, *Pandanus odoratissimus*, and *Rhizophora apiculata* being more effective than other common vegetation, including the mangrove *Avicennia alba*." Figure 10 clearly indicates that *Pandanus odoratissimus* always has larger values of drag coefficient—signifying larger drag force—regardless of the tsunami height. This data reveals the possible limitation of mangrove forests' functions, and they point to "the importance of preserving or selecting appropriate species to act as wave barriers to offer sufficient shoreline protection." Nevertheless, even if this study carried out by Tanaka *et al.* (2007) is credible, mangrove forests' ability to mitigate tsunamis is undoubtedly significant, and reasonable plantation strategies could produce ideal effects.

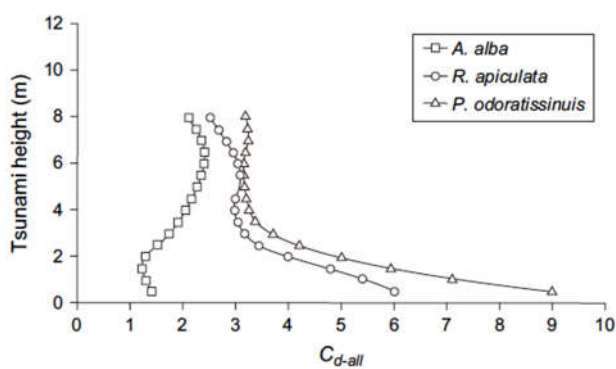


Figure 10. Changes in the relative drag coefficient for vertical vegetation structure with increased height of a tsunami for the mangroves, *Avicennia alba* and *Rhizophora apiculata*, and the palm, *Pandanus odoratissimus*, estimated from modelling data from Sri Lanka and Thailand. <taken from Alongi (2008)>

2.3. Plantation Strategy

The effects of various factors on mangrove forests' ability to mitigate tsunamis were analyzed in section 2.2. In this section, plantation strategies of mangrove forests for coastal regions will be discussed, based on the studies mentioned in section 2.2.

Harada and Imamura (2005) used numerical simulation to examine the influence of forest width on coastal forests' ability to mitigate tsunamis, and they synthesized the data in Table 2, which described how coastal forests of different widths behaved in mitigating waves under different tsunami heights.

Table 2. *Tsunami reduction rate by coastal control forest. <taken and modified from Harada and Imamura (2005)>*

		Tsunami height (m)			
		1	2	3	
Effect by coastal control forest (Shuto, 1987)		Mitigate damage, Stop drifts, Mitigate tsunami			
Run up distance	Forest width	50 m	0.98	0.86	0.81
		100 m	0.83	0.80	0.71
		200 m	0.79	0.71	0.64
		400 m	0.78	0.65	0.57
Inundation depth	Forest width	50 m	0.86	0.86	0.82
		100 m	0.76	0.74	0.66
		200 m	0.46	0.55	0.50
		400 m	---	0.11	0.18
Current	Forest width	50 m	0.71	0.58	0.54
		100 m	0.57	0.47	0.44
		200 m	0.56	0.39	0.34
		400 m	---	0.31	0.24
Hydraulic force	Forest width	50 m	0.53	0.48	0.39
		100 m	0.33	0.32	0.17
		200 m	0.01	0.13	0.08
		400 m	---	0.02	0.01

The run-up distance does not provide critical information, for Harada and Imamura (2005) did not specify the distances where tsunami heights were reduced to a safe range. For instance, if the tsunami rushed 200 m onto the shore, and was reduced to 0.3 m [a safe height according to Bao (2011)] in height halfway, then run-up distances of “200 m” and “100 m” could be regarded as the same, based on the tsunami's impact on residents and properties.

On the other hand, inundation depth, current, and hydraulic force all provided valuable information. The analysis in this section will focus on a tsunami height of 3 m, since resulting plantation plans under this circumstance will certainly ensure safety when tsunami heights are either 1 m or 2 m. Remember that reduction rate is defined by Equation (2) in Table 2.

$$r (\text{reduction rate of maxvalues}) = \frac{(\text{max .value with forest})}{(\text{max .value without forest})} \quad (2)$$

According to Table 2, the reduction rate of inundation depth is 0.18 for forests with 400 m width. Even if we assume that the maximum value of inundation height is the same as the initial tsunami height—3 m—the maximum value of inundation depth with the coastal forest is 0.54 m, which is already very close to the safe height of 0.3 m. Additionally, if residential areas are built hundreds of meters away from the mangrove forests, the inundation depth will certainly reduce to the safe range. As for forests with 200 m width, the inundation height is nevertheless still 1.5 m after penetrating the coastal forest, remaining dangerous for residents. The differences in efficacy between coastal forests of 200 m width and 400 m width are not as significant for current and hydraulic force mitigation, whereas

forests with 400 m width are still more powerful in mitigating tsunami waves. The forests' ability to dissipate hydraulic force is especially remarkable, illustrated by the data that confirms that the reduction rate of hydraulic force is only 0.01 for forests of 400 m width.

The analysis based on Harada and Imamura (2005)'s study suggests that coastal forests of width 400 m significantly contribute to reduce inundation height to a safe range and dissipate wave energy, therefore ensuring residents' physical and property safety. Among coastal forests, mangroves are certainly an excellent choice. In order to generalize the conclusion of his study, Bao (2011) developed an equation to calculate the required mangrove band width that could effectively mitigate tsunamis. The equation is shown below.

$$B_w = \frac{\ln(W_h) - \ln(a)}{b} \quad (4)$$

Remember that B_w is forest band width (m), W_h is safe wave height behind the forest (cm), coefficient a is directly proportional to the initial wave height, and b is a negative coefficient as a function of forest structure. Bao (2011) assumed that the average maximum initial wave height was 300 cm, based on his observation of wave heights along Vietnam coasts. He also interviewed 50 people who worked in aquaculture and agriculture in the research area to finally determine the safe wave height behind the forest band as 30 cm. After inserting all the known values, the required mangrove band width (B_w) is a function only of the forest structure index (b). Bao (2011) then defined V as the vegetation index, and replaced b with V in the equation below.

$$V = -b = [-0.048 + 0.016 \times H + 0.00178 \times \ln(N) + 0.0077 \times \ln(CC)] \quad (5)$$

In Equation (5), H is tree height (m), N is forest density (tree·ha⁻¹), and CC is canopy closure (%). The detailed calculation method of V is irrelevant and will not be discussed in this section. Bao (2011) discovered that required mangrove band width (B_w) decays exponentially with increase in vegetation index (V). That is, when the mangrove forest is tall and dense, and the percentage of canopy closure is high, a narrower forest band is required. The relationship between vegetation index (V) and required mangrove band width (B_w) is shown in Figure 11.

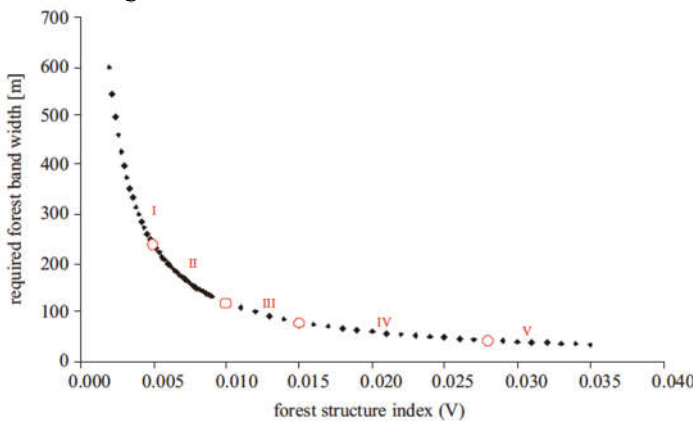


Figure 11. Theoretical curve showing the relationship between mangrove structure index (V) and mangrove band width (B_w) (m). <taken from Bao (2011)>

As an important factor influencing mangroves' ability to mitigate tsunamis, forest width is discussed further from the aspect of plantation strategies in this section. Although both Harada and Imamura (2005), and Bao (2011)'s study were not carried out in real-life tsunami scenarios, their conclusions were still valid. The two studies simultaneously examined tsunami heights of 3 m, which is claimed as the height of most tsunamis by National Geographic (2011). Although extremely devastating tsunamis could exceed that height—the 2004 Indian Ocean tsunami reached heights of 30 m in some places—those are not typical instances. The conclusion that coastal forests with a width of 400 m, a density of 30 trees/100 m², and a trunk diameter of 0.15 m could ensure safety under the tsunami with a 3m height and a 10-min period is a valuable reference for coastal city administrators in plantation strategy. Equation (4) for calculating the required mangrove forest band based on mangrove forest structure and the tsunami parameters is a valid tool as well.

However, planners should make modifications when drafting the plan, taking into consideration geographical and climatic differences between the target areas in studies and the real situation. They are supposed to take other factors into account as well, including forest density, tree species, forest floor slope, and characteristics of incident waves. For instance, planned mangrove forest width and density could be reduced by virtue of powerful wave-resistant species, and denser forests with wider widths are required for regions where forest floors are flat, and tsunamis take place frequently. It is noteworthy that even in coastal cities without frequent tsunami disasters, it is worthwhile to plan mangrove forests, as they provide a great number of ecological services, including fisheries, biofiltration, and carbon emission reduction, to name but some benefits.

3. Mangrove Forests Require Conservation

3.1. Necessity of Protection

3.1.1. Ecological Services of Mangrove Forests

According to Spalding *et al.* (2010), mangroves offer a wide range of goods and services to humans. These ecological services are listed below.

1. Coastal Protection. According to Priya *et al.* (2010), mangrove forests protect coastlines from erosion, storm damage, wave action, tsunami, cyclones, and typhoons. As a consequence, the forests protect coastal land and adjacent residents both during natural disasters and through their longer-term influence on coastal dynamics (Blankespoor *et al.*, 2016). This was already examined in section 2.

2. Forest products. Mangroves offer timber and fuel wood for tropical residents, and the tannin they secrete is used for preparation of leather.

3. Fisheries. Mangrove forests provide fish and other coastal wildlife with ideal shelters consisting of wide intertidal mudflats and complex systems of channels and pools. They also ensure a rich source of nutrients due to high rates of primary production. It is revealed that “mangrove-related species have been estimated to support 30% of fish catch and almost 100% of shrimp catch in South-East Asian countries, while mangroves and associated habitats in Queensland, Australia, support 75% of commercial fisheries' species” (Spalding *et al.*, 2010).

4. Economic Valuation. In Spalding *et al.* (2010)'s research conducted in 2010, global mangrove forests created a summary value of \$2000 to \$9000 per hectare per year, and the values were derived from all products and services mangrove forests provided, including timber, fisheries, coastal protection, etc.

5. Tourism and Recreation. Either taking a broad walk in well-preserved mangrove forests or boating around is enjoyable.

6. Biofiltration. Mangrove forests constrain water movement, trap sediments, and remove pollutants when water passes through.

7. Carbon Emission Reduction. Mangrove forests are comparable to higher canopy terrestrial forests, since they have a larger proportion of below-ground biomass, and sequester more CO₂ than other forests. Spalding *et al.* (2010) pointed out in their research that “the total above-ground biomass for the world’s mangrove forests may be over 3700 Tg of carbon, and further that sequestration of organic matter directly into mangrove sediments is likely to be in the range of 14 to 17 Tg of carbon per year” (Spalding *et al.*, 2010).

Apart from the major ecological services listed above, mangroves were also used as medicine for ancient indigenous residents, and some societies also ate their fruits. Because of the great number of ecological functions of mangrove forests, they are well-worth preserving.

3.1.2. Destruction of Mangrove Forests

According to Priya *et al.* (2010), “Mangrove forests are currently among the most threatened habitats in the world and are disappearing at an accelerated rate.” In 2010, Spalding *et al.* published their research data:

“Some 35,600 square kilometers of mangrove forests were lost between 1980 and 2005, and while we have no accurate estimates of the original cover, there is a general consensus that it would have been over 200,000 square kilometers, and that considerably more than 50,000 square kilometers, or one-quarter, of original mangrove cover have been lost as a result of human actions. While rates of loss decreased from 1.04% per year through 1980s to 0.66% per year in the five years to 2005, these rates are still three to five times greater than overall rates of global forest loss (Spalding *et al.*, 2010).”

When the evaluation of the extent of destruction is narrowed down to countries and regions, it was claimed in the study carried out by Danielsen *et al.* in 2005 that “human activities reduced the area of mangroves by 26% in the five countries most affected by the tsunami, from 5.7 to 4.2 million ha, between 1980 and 2000”, and Bao (2011) revealed a similar trend in Vietnam where he conducted his research: “In 2002, Vietnam had approximately 155,290 ha of mangrove forests. More than 200,000 ha of mangrove forests have been destroyed over the last two decades as a result of conversion to agriculture and aquaculture, as well as development for recreation.”

Both artificial and natural factors contribute to mangrove forests’ destruction. According to Spalding *et al.* (2010), the greatest artificial element goes to direct conversion of mangrove lands. On one hand, mangrove lands are converted into residential or industrial areas because of intense pressure for land in typical coastal zones. On the other hand, some countries, like the Philippines, affect mangrove forests adversely through high-level policy decision, such as encouragement of aquaculture. The drive for conversion of mangrove lands into agricultural lands is typically due to demand for cash crops. Likewise, demand for seafood, especially shrimp, results in conversion into aquaculture ponds. These ponds are built in intertidal areas, and they generally stock shrimp larvae from incoming tides. The out-flowing water from these ponds are rich in chemicals and nutrients, thus being highly polluting.

There are other artificial contributors to mangrove forest destruction. Overharvesting occurs when either mangrove ferns fill the gap, preventing large trees from growing, or mangrove foliage is eaten by goats and camels. Both oil pollution and chemical

pollution, which mainly come from fisheries, impose threats on mangrove forests. Although mangroves can deal with certain levels of sediments, they suffer mass mortality when there are rapid sediment build-ups of 50 to 100 cm.

In addition to artificial elements, natural conditions threaten mangroves as well. For example, climate change imposes a wide range of negative impacts on mangroves: rise in sea level, rise in atmospheric CO₂, rise in air and water temperature, and change in frequency and intensity of precipitation/storm patterns due to climate change (Blankespoor *et al.*, 2016). Mangroves are struggling to survive in deteriorating natural situations, and the process is typically outside of civilians' awareness.

As a species with plenty of ecological functions, mangroves ought to be protected. Leaving them subject to various means of destruction until they are endangered is the last thing people expect. In section 3.2., conservation strategies are discussed in detail.

3.2. Conservation Strategies of Mangrove Forests

Mangrove distribution by regions in 2000 is shown in *Figure 12*.

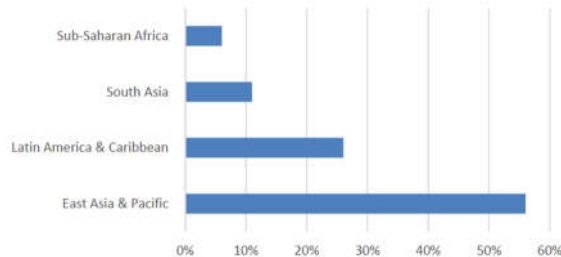


Figure 12. *Distribution of mangroves by region, 2000. <taken from Blankespoor et al. (2016)>*

Based on *Figure 12*, the majority of mangrove forests in the world were found in Latin America, Caribbean regions, East Asia, and Pacific regions. Although the investigation was undertaken 17 years ago, the distribution trend is still applicable to the current condition. A study conducted in 2010 claimed that “destruction rate of the mangrove habitat is now at about 1% every year” (Priya *et al.*, 2010), indicating that mangrove forests worldwide are declining at similar paces. Hence, a relative abundance of mangrove forests in different regions has been kept stable until now.

Conservation of mangrove forests requires efforts made by people from all walks of life. Government officials, for one, should enforce appropriate laws and regulations. They need to ensure that “mangroves are properly taken into account in the assessments that drive policy, economic decision-making and governance” (Spalding *et al.*, 2010). Apart from benefits of industrial and commercial exploitation, local and subsistence benefits given by mangrove forests need to be calculated as well. The ecological, social and financial costs of degradation or clearance of mangrove forests, likewise, need to be fully assessed and taken into account.

Spalding *et al.* pointed out in their 2010 research that some countries were already setting good examples in policy enforcement. It is presented that “Brazil, Mexico, Cambodia, El Salvador and Tanzania all have established legal frameworks for the protection of mangroves” (Spalding *et al.*, 2010). Specifically, Brazil set a federal law to protect all coastal vegetation, and restrict tourist and aquaculture developments; the

Philippines enforced strict rules over the establishment of new aquaculture ponds, requiring mangrove forests to be shelters for fish and other wildlife species; Australia and the US stuck to the “no net loss” principle, and both of them requested developers to compensate for areas proposed for conversion by investing in afforestation restoration projects elsewhere; Kenya and Malaysia, similarly, render all mangroves to fall under state ownership and manage them as forest reserves.

With the support of laws and regulations, government administrators and ecological workers could practice a series of conservation actions, as listed below:

1. Sustainable Silviculture. Spalding *et al.* (2010) showed that a number of large forests in Asia, especially in the Sundarbans in Bangladesh and India and Matang in Malaysia, were well-protected. In some areas, protectors conducted intensive management of mangrove forests with well-established structures of plantation. Thinning of mangrove forests were practiced as well, meaning that slower growing or defective trees were removed to provide more space for the remaining trees to grow, resulting in bigger, healthier trees in a shorter period of time (Oregon Forest).

2. Restoration and Afforestation. When mangroves are returned to areas where they previously lived, this is called restoration. However, when they are transported to places with no evidence of prior existence, it is named as afforestation. In the past, people restored mangrove forests for timber and fuelwood, or they relied on mangrove forests for protection from typhoons. Since the 1980s, however, the purposes of mangrove forest conservation have diversified, including “protection of inland resources and human life during storm surges, erosion reduction, biodiversity conservation, fisheries enhancement, aesthetic improvement”, to name a few. Restoration is conducted more often than afforestation, as proper locations and species should be selected for successful afforestation. Otherwise, failure occurs easily (Spalding *et al.*, 2010).

3. Flow Restoration. Maintenance of environmental flows upstream of mangrove areas is also critical (Priya *et al.*, 2010). Coastal engineering interventions, such as dams, could lead to downstream water cessation and wetland hyper-salinization. Protectors, therefore, should restore flows naturally or artificially to rivers and wetlands, allowing full or partial ecosystem recovery while maintaining the required services of coastal engineering interventions as well (Spalding *et al.*, 2010).

4. Alternative Fuel Development. Priya, *et al.* (2010) discovered this creative means for mangrove conservation. Alternate fuels could replace mangrove fuelwood, thereby protecting mangrove forests from large-scale felling (Priya *et al.*, 2010).

While the above actions are confined to expert protectors, civilians can also make different choices to protect mangrove forests. As groups, residents could form NGOs to help urge the issue of mangrove protection. As individuals, they could help plant mangroves for coastal protection, which was underlined as effective in the Philippines. In order to encourage citizens to take action, the values of mangroves should be widely spread through public bodies, the media and governments. Dwellers should also realize the benefits from their non-destructive direct uses of mangrove forests, including fisheries, tourism, and recreation.

For a quick review, conservation strategies employed by different countries are illustrated in Table 3 below. The five regions in the table are derived from Figure 12, with Oceania added. Based on their distribution of mangroves than their geographic locations, countries are classified into different regions. There're no underlines for conservation strategies deployed by governments. However, actions taken by ecological experts and ordinary citizens are denoted with wavy lines and straight lines, respectively.

Table 3. Mangrove conservation strategies employed by different countries.

Region	Country Name	Conservation Strategies
East Asia & Pacific	The Philippines	1. Strict rules over the establishment of new aquaculture ponds. 2. Individual residents' voluntary service to plan mangroves for coastal protection.
	Mexico	1. Legal frameworks for the protection of mangroves.
	El Salvador	1. Legal frameworks for the protection of mangroves.
Latin America & Caribbean	Brazil	1. Federal law to protect all coastal vegetation. 2. Restriction of aquaculture developments.
	The United States of America	1. "No net loss" principle. 2. Developers' compensation for areas proposed for conversion by investing in afforestation restoration projects elsewhere.
South Asia	Cambodia	1. Legal frameworks for the protection of mangroves.
	Malaysia	1. All mangroves under state ownership. 2. All mangroves managed as forest reserves. 3. <u>Sustainable silviculture.</u>
	Bangladesh	1. <u>Sustainable silviculture.</u>
	India	1. <u>Sustainable silviculture.</u>
Sub-Saharan Africa	Kenya	1. All mangroves under state ownership. 2. All mangroves managed as forest reserves.
	Tanzania	1. Legal frameworks for the protection of mangroves.
Oceania	Australia	1. "No net loss" principle. 2. Developers' compensation for areas proposed for conversion by investing in afforestation restoration projects elsewhere.

According to *Table 3*, almost all countries included issue laws to protect mangrove swamps. Additionally, three of the South Asian countries—Malaysia, India, and Bangladesh—implement sustainable ecological practices. These should be promoted in the East Asia & Pacific region, and the Latin America & Caribbean region, where more mangroves locate, according to *Figure 12*. Likewise, the civic engagement in mangrove conservation practiced in the Philippines should be introduced in large scale to more countries and regions, especially the East Asia & Pacific region, and the Latin America & Caribbean region. When more citizens undertake the conservation work, mangrove protection will surely be much more efficient.

Generally speaking, when government officials, expert ecological workers, and ordinary residents coordinate their efforts, the effect on mangrove forest conservation will undoubtedly be significant. In this case, mangroves could continue guarding the coastlines and providing people with a variety of benefits.

4. Conclusion

Mangroves living in intertidal areas are special species with a great variety of ecological functions. As examined in numerous studies, their capacity for tsunami mitigation is distinguished, and this ability is influenced by several factors including forest width, forest density, tree age, to name some of the most significant.

Forest width plays a prominent role in tsunami damage mitigation, especially in inundation depth reduction. Although the significance of forest density's effect on tsunami remission function is disputed, it is commonly allowed that the denser the coastal forests, especially mangrove forests, the stronger mitigating power they have. Mature mangroves aged around five to six years are proved to be effective in wave attenuation, and this mechanism is almost unaffected by increasing water depth.

When making plantation plans, the conclusion that "coastal forests with width of 400 m, density of 30 trees/100 m², and trunk diameter of 0.15 m could ensure safety under the tsunami with 3 m height and 10 min period" is a valuable standard city that administrators could depend on. Equation (4) for calculating the required mangrove forest band based on mangrove forest structure and the tsunami parameters consider more factors in the calculation, and it is worth referring to in plantation design. Additionally, city administrators are supposed to take other factors into consideration, including forest density, tree species, forest floor slope, and characteristics of incident waves. They should adopt the above conclusions flexibly and make necessary modifications taking into account varying geographical and climatic factors.

Despite the wide range of ecological services provided by mangroves, people tend to overlook their value. As a result, global mangrove forests are currently disappearing at a fast pace. Not only do human interventions impose threats to mangroves, but changing climate is also threatening their lives. In order to preserve this precious vegetation, governments ought to enforce conservation laws and regulations, experts should actively employ conservation actions—such as sustainable silviculture, afforestation and flow restoration—and civilians need to learn more about mangroves and participate in conservation activities.

Works Cited

Alongi, D.M., 2008, Mangrove forests: Resilience, protection from tsunamis, and responses to global climate change: *Estuarine, Coastal and Shelf Science*, v. 76, p. 1–13, doi: 10.1016/j.ecss.2007.08.024.

Bao, T.Q., 2011, Effect of mangrove forest structures on wave attenuation in coastal Vietnam: *Oceanologia*, v. 53, p. 807–818, doi: 10.5697/oc.53-3.807.

Blankespoor, B., Dasgupta, S., and Lange, G.-M., 2016, Mangroves as Protection from Storm Surges in a Changing Climate: Policy Research Working Papers, doi: 10.1596/1813-9450-7596.

Brinkman, R.M., Massel, S.R., Ridd, P.V., Furukawa, K., 1997, Surface wave attenuation in mangrove forests. *Proceedings of 13th Australasian Coastal and Ocean Engineering Conference 2*, p. 909-914, doi: search.informit.com.au/documentSummary; dn=036154632045063; res=IELENG.

Carylsue, 2014, Bangladesh Braces for Oil Spill Impact: Nat Geo Education Blog: <https://blog.education.nationalgeographic.org/2014/12/17/bangladesh-braces-for-oil-spill-impact/> (accessed November 2017).

Danielsen, F. et al., 2005, The Asian Tsunami: A Protective Role for Coastal Vegetation: *Science*, v. 310, p. 643, doi: 10.1126/science.1118387.

Danielsen, F. et al., 2005, Online Supporting Material for The Asian Tsunami: A Protective Role for Coastal Vegetation: *Science*, v. 310, p. 643, doi: sciencemag.org/cgi/content/full/310/5748/643/DC1.

Dasgupta, R., Shaw, R., and Abe, M., 2014, Environmental Recovery and Mangrove Conservation: Post Indian Ocean Tsunami Policy Responses in South and

Southeast Asia: Recovery from the Indian Ocean Tsunami Disaster Risk Reduction, p. 29–42, doi: 10.1007/978-4-431-55117-1_3.

Gibbons, H. and Gelfenbaum, G., 2005, Astonishing Wave Heights Among the Findings of an International Tsunami Survey Team on Sumatra: <https://soundwaves.usgs.gov/2005/03/> (accessed July 2017).

Harada, K., and Imamura, F., Effects of Coastal Forest on Tsunami Hazard Mitigation—A Preliminary Investigation: Tsunamis Advances in Natural and Technological Hazards Research, p. 279–292, doi: 10.1007/1-4020-3331-1_17.

Massel, S.R., Furukawa, K., and Brinkman, R.M., 1999, Surface wave propagation in mangrove forests: Fluid Dynamics Research, v. 24, p. 219–249, doi: 10.1016/s0169-5983(98)00024-0.

Mazda, Y., Magi, M., Ikeda, Y., Kurokawa, T., and Asano, T., 2006, Wave reduction in a mangrove forest dominated by *Sonneratia* sp.: Wetlands Ecology and Management 14, v. 14, p. 365–378, doi: 10.1007/s11273-005-5388-0.

Mazda, Y., Magi, M., Kogo, M., and Hong, P. N., 1997, Mangroves as a coastal protection from waves in the Tong King delta, Vietnam: Mangroves and Salt Marshes, v. 1, p. 127–135, doi: 10.1023/A:1009928003700.

Mazda, Y., Wolanski, E., King, B. et al., 1997, Drag force due to vegetation in mangrove swamps: Mangroves and Salt Marshes, v. 1, p. 193–199, doi: 10.1023/A:1009949411068.

National Geographic, 2011, Tsunami Facts in Wake of Japan Earthquake: <http://news.nationalgeographic.com/news/2011/03/110311-tsunami-facts-japan-earthquake-hawaii/> (accessed July 2017).

Oregon Forest, Thinning for Forest Health: <http://oregonforests.org/content/thinning> (accessed July 2017).

Priya, K.R., Kiran, K.S., and Pema, U., 2010, Role of Sand Dunes and Mangroves in the Mitigation of Coastal Hazards with Reference to 2004 Tsunami: Management and Sustainable Development of Coastal Zone Environments, p. 245–258, doi: 10.1007/978-90-481-3068-9_16.

Quartel, S., Kroon, A., Augustinus, P., Santen, P.V., and Tri, N., 2007, Wave attenuation in coastal mangroves in the Red River Delta, Vietnam: Journal of Asian Earth Sciences, v. 29, p. 576–584, doi: 10.1016/j.jseas.2006.05.008.

Shuto, N., 1987, The effectiveness and limit of tsunami control forests: Coastal Engineering in Japan, v. 30, p.143–153.

Spalding, M., Kainuma, M., and Collins, L., 2010, World Atlas of Mangroves: London, UK, Washington, DC., p. 1–43.

Tanaka, N., Sasaki, Y., Mowjood, M.I.M., Jinadasa, K.B.S.N., and Homchuen, S., 2007, Coastal vegetation structures and their functions in tsunami protection: experience of the recent Indian Ocean tsunami: Landscape Ecology and Engineering, v. 3, p. 33–45, doi: 10.1007/s11355-006-0013-9.

Webster, M. Mangrove: Merriam-Webster: <https://www.merriam-webster.com/dictionary/mangrove> (accessed November 2017).

Appendix

To view a clear picture of regions studied in this paper, please refer to *Figure A1* and *Table A1* below.



Figure A1. Locations of areas discussed in the paper. <modified from Carylsue (2014)>

Table A1. Detailed information on regions studied in the paper.

Region Number and the Corresponding Country Name in Figure 13	Section(s) Where Mangroves of This Region are Studied	Corresponding Author(s) of Materials Studied	More Specific Area Descriptions
Region 1: India	Section 2.1	Danielsen <i>et al.</i> (2005)	Cuddalore coastline
	Section 2.2.2	Danielsen <i>et al.</i> (2005)	Cuddalore coastline
Region 2: Vietnam	Section 2.2.1	Bao (2011)	
	Section 2.2.3	Mazda <i>et al.</i> (1997)	Tong King Delta
	Section 2.2.4	Mazda <i>et al.</i> (2006)	The Vinh Quang coast
Quartel <i>et al.</i> (2007)		The coast north of Do Son in the Red River Delta	
Region 3: Japan	Section 2.2.1	Harada and Imamura (2005)	The long chain of the Nansei Islands up to the very southern tip of the larger island of Kyushu

Region 3: Japan (cont.)	Section 2.2.2	Harada and Imamura (2005)	The long chain of the Nansei Islands up to the very southern tip of the larger island of Kyushu
	Section 2.2.4	Brinkman <i>et al.</i> (1997)	Iriomote Island
		Massel <i>et al.</i> (1999)	Iriomote Island
		Mazda <i>et al.</i> (1997)	Nakama-Gawa in Iriomote Island
Region 4: Australia	Section 2.2.4	Brinkman <i>et al.</i> (1997)	Townsville
		Massel <i>et al.</i> (1999)	Townsville
		Mazda <i>et al.</i> (1997)	Coral Creek in Hinchinbrook Island
Region 5: Sri Lanka	Section 2.2.4	Tanaka <i>et al.</i> (2007)	Southern coast
Region 6: Thailand	Section 2.2.4	Tanaka <i>et al.</i> (2007)	The Andaman coast



Microneedle Transdermal Drug Delivery for Traditional Chinese Medicated Plasters

Chendan Luo

Author Background: Chendan Luo grew up in China and currently attends WHBC of Wuhan Foreign Languages School in Wuhan, China. Her Pioneer seminar topic was in the field of engineering and titled “Mechanical Engineering.”

Abstract

As an outstanding member in the third generation of transdermal drug delivery systems, the field of microneedle is a new and intriguing area that combines medical, engineering, and material sciences. The concept of a microneedle, which was first proposed in 1976, is a more advanced painless transdermal method applicable for the delivery of both small and large drug molecules compared to the first two generations. There is a wide variety of different microneedles (solid, hollow, coated, and dissolvable) that allow for transdermal drug delivery using different approaches (“poke and patch”, “poke and coat”, “poke and release” and “poke and flow”) for different situations. Microneedle patches and microneedle rollers are two outstanding applications of the microneedle system. Traditional Chinese medicine and clinics, which applied pharmacology totally differently from Western medicine, are widely accepted around the globe today, and traditional Chinese medicated plaster is an important part of Chinese medicine. However, there are few research studies into the ineffective traditional method to meet the current market. With consideration from permeability, mechanics, kinetics of skin resealing, and other related areas, this paper discusses and evaluates the possible utilization of microneedle systems in the traditional Chinese plasters, as well as the potential for future development.

1. Introduction

The concept of the transport of drug across the skin has been first described by Ibn Sina (AD 980–1037), a Persian physician known as Avicenna within the Western World. In *the Canon of Medicine*, he divided topical drugs into two states: soft and hard. He suggested that, when applied to the skin, the drugs belonging to the ‘soft’ category could penetrate the skin, while the drugs in the ‘hard’ category could not. He further proposed that dermally applied drugs not only have local effects, but that they may also affect areas immediately beneath the skin including joints (regional effects), as well as remote areas of the body (systemic effects). Sina created a topical formulation that acted systemically in order to help treat conditions where drugs could not be taken orally [13]. His conclusion partially reveals the truth and advantages of transdermal drug delivery.

Transdermal drug delivery is a non-invasive delivery of medication from the surface of the skin, down through its layers. With specific merits and disadvantages, the oral, parenteral, ophthalmic and transdermal routes are available for drug delivery. Compared to conventional injection and oral methods, an ideal transdermal drug delivery system (TDDS) has multiple desirable advantages: it is painlessness; there is avoidance of

first-pass metabolism and gastrointestinal incompatibility; the drug can be delivered selectively to a specific target area; it is suitable for self-administration, and finally shows improved therapeutic efficacy and efficiency.

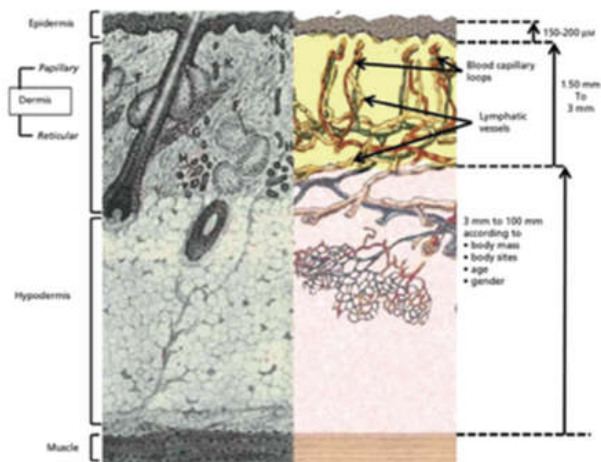


Figure 1. Skin structure showing three major regions: epidermis, dermis and hypodermis (with the thickness range). (Copyright 2008 Elsevier) [2]

As seen in Figure 1, the skin can be divided into three regions: the epidermis, dermis and hypodermis. The first skin layer, epidermis (150–200 μm), is made up of viable cells without a vascular network. This layer obtains its nutritional needs by passive diffusion through interstitial fluid [17]. The outermost layer of the epidermis (10–20 μm) consists of dead cells, known as the *stratum corneum*; it acts as a rigorous barrier to the transdermal drug delivery. The second skin layer, dermis, contains an extensive nervous and vascular network [2]. The pain associated with drug delivery is due to damage to the nerves endings within dermis. Since only a few potent drug molecules having high lipophilicity and small molecular weight can be administered directly through passive diffusion [17], the primary goal of transdermal drug delivery research is to increase the number of drug candidates for effective transdermal drug delivery. The ideal of the transdermal drug delivery is to cross the intact stratum corneum layer without causing damage to nerve endings and vessels. In this way, a drug can be effectively absorbed through a painless process.

The development of transdermal drug delivery has taken place across three generations. The first generation gave rise to the transdermal patch-based delivery system. Medicated plasters, which were generally applied to the skin to treat localized ailments, can be traced back to ancient China (around 2000 BC) and are the early predecessors of today's transdermal patches [13]. These early plasters generally contained several medical herbs spread over an adhesive, a natural gum-rubber base, which was then applied to a backing support made of fabric or paper [13].

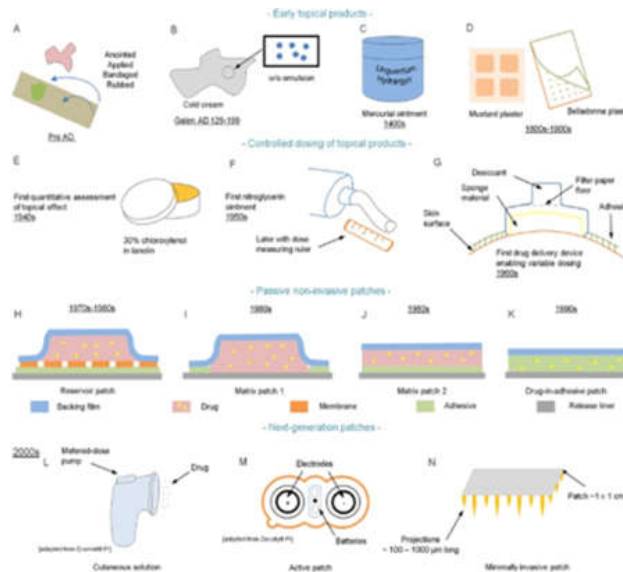


Figure 2. Historical development of patches. Early topical products: (A) Products from ancient times; (B) Galen's cold cream; (C) Mercurial ointment; (D) Mustard and belladonna plasters; controlled dosing of topical products: (E) First quantitative systemic delivery (Zondek's system); (F) Individualized delivery system: nitroglycerin ointment; (G) Topical delivery device (Wurster & Kramer's system). Passive non-invasive patches; (H) First patch system – the reservoir – introduced for scopolamine, nitroglycerin, clonidine and oestradiol; (I, J, K) Other types of patches – matrix and drug-in-adhesive (e.g. fentanyl and nicotine patches); Next-generation patches; (L) Cutaneous solutions (e.g. Patchless Patch®, Evamist®); (M) Active patches (e.g. iontophoresis, Zecuity®); (N) Minimally invasive patches (e.g. microneedles, Nanopatch®). [13]

In the early transdermal patches, the reservoir stored the drug. In the early 1970s (see Figure 2H) entrepreneur Alejandro Zaffaroni developed the first transdermal patch. In these designs, four layers were included: an impermeable backing membrane, a drug reservoir, a semi-permeable membrane that serve as a rate-limiting barrier to control the rate of transdermal delivery during the continues transdermal delivery, and an adhesive layer. Alejandro Zaffaroni's version of transdermal patch cannot met the market's need; a major limitation in this system is the potential for leakage from its sealed liquid reservoir that could arise due to manufacturing defects. Later, matrix patches with the matrix concept for nitroglycerin, oestradiol and testosterone overcame the limitations of Alejandro Zaffaroni's design. In these designs, the drug was incorporated into a solid polymer matrix, which simplifies manufacturing. The Matrix system may consist of three layers, after eliminating the semi-permeable layer compared to the former design. The three-layer Matrix patches were not only generally thinner and more flexible, and therefore more comfortable and had better adhesive properties, but they were also less expensive to manufacture. In addition, the state of the drug was not constrained. In later cases, for example the viscous adhesive (DIA) patches, a matrix could have just two layers by simply incorporating the drug entirely in a pressure-sensitive adhesive (PSA) [13].

There was a big concern about the usefulness of all first-generation patches: not all drugs were suitable for patch delivery. The first-generation transdermal delivery was limited by the stratum corneum. Drug transport across the stratum corneum involves diffusion

through the intercellular lipids. The transportation pathway is highly constrained by the structural and solubility requirements for solution and diffusion within stratum corneum lipid bilayers. Thus, the candidates for first generation delivery must be low-molecular weight, lipophilic, and efficacious at low doses, which largely limited the application of transdermal drug delivery.

The second generation of transdermal drug delivery systems is a more 'active' system which focuses on enhancing skin permeability, including conventional chemical enhancers, iontophoresis and non-cavitation ultrasound. However, the applications of second-generation systems are particularly effective for small molecules.

The third generation of transdermal delivery systems includes novel chemical enhancers, electroporation, cavitation ultrasound, microneedles, thermal ablation and microdermabrasion. The third-generation patches cause stronger disruption of the stratum corneum barrier, without affecting deeper tissues. Thus, the third-generation transdermal drug delivery system can be applied for both small and large molecules. Among the third-generation patches, the microneedle (MN) is a novel and very promising transdermal drug delivery method. It was first conceptualized for drug delivery in 1976 by ALZA Corporation, but only became the subject of significant research starting in the mid-1990's when microfabrication technology enabled their manufacture [15]. The first report to demonstrate MNs for transdermal delivery was not published until two decades later.

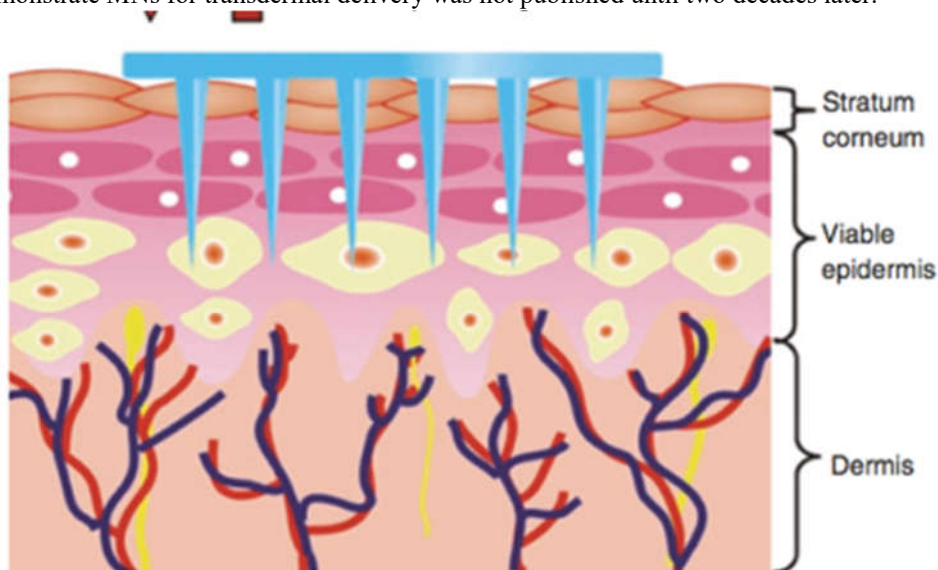


Figure 3. Concept of microneedles that can penetrate through the skin without touching the nerves and vessels. (Copyright 2011 Elsevier) [3]

To meet the goal of effective transdermal drug delivery, microneedles cause temporary disruption of the stratum corneum, creating microscopic aqueous pores as paths for drug to enter the dermal microcirculation for subsequent systemic absorption. Microneedles have an approximate length of 50-900 μm and an external diameter of no more than 300 μm . Microneedles are designed to penetrate through the epidermis up to a depth of 70-200 μm . Since microneedles are thin and short, they do not penetrate the dermis layer containing nerves and vessels (see Figure 3), and therefore painless application is

possible. With the simplicity of use and low cost, microneedles allow for easy and patient-friendly administration of medication to and across the skin. Since microneedles can be applied for both small and large molecules, with multiple varieties, microneedles can be widely used in different transdermal drug delivery situations.

There are generally four types of microneedles: solid microneedles, drug-coated microneedles, dissolving microneedles, and hollow microneedles. As seen in Figure 4, microneedles are first applied to the skin (A) and then used for drug delivery (B). Solid microneedles are used to penetrate the skin, and then a drug can diffuse through the residual holes in skin from a topical formulation (solid MN). After insertion of drug-coated microneedles into the skin, the drug coating dissolves off the microneedles in the aqueous environment of the skin (coated MN). Drug-loaded microneedles are made of water-soluble or biodegradable materials encapsulating the drug that is released into the skin upon microneedle dissolution (dissolving MN). Hollow microneedles are used to inject liquid formulations into the skin (hollow MN) [16].

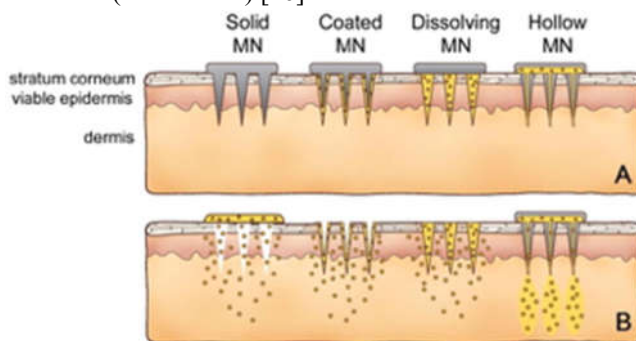


Figure 4. Different types of microneedles (MN). Solid MN, Coated MN, Dissolving MN, and Hollow MN. (Copyright 2012 Elsevier) [2]

Even with all the enhancements described above, the ideal solution for the transdermal drug delivery remains hidden. Although applications of microneedles already occur in the field of cosmetic, diagnosis, and drug delivery, there are limited studies and research in the field of traditional Chinese medicated plasters.

Traditional Chinese medicine, as a regular complementary health approach, originated in ancient China and has evolved over thousands of years. Nowadays, many of the Chinese medicines have been well known and habitually used around the globe due to their curative effects, convenience and inexpensiveness. Medicated plasters, a dressing that contains medicine aimed at relieving pain and swelling, has been an outstanding part of traditional Chinese medicine since its origin around 2000 BC. In popular traditional Chinese clinics, most traditionally medicated plasters are still being used. Certain herbs are blended together in different proportions and ground into powder. Then different liquids are added, such as honey, vinegar, and wine. After the prepared mixture is applied to the plastic backing support and stuck to the target skin area, the added liquid will permeate through the skin and then be absorbed. The limitations of Chinese plaster are due to its characteristics and include low absorption, as well as indirect and inaccurate transdermal drug delivery to the target area. However, by combining the current techniques and drug delivery enhancers, it is possible to partially fix the drug delivery limitations of medicated plasters.

2. Materials and Methods

2.1 Materials

Microneedles can be divided into three categories: solid, degradable/dissolvable and hollow. To meet the criteria of different microneedles, microneedles have been produced using glass, silicon, metals and polymers (see table 1).

Solid microneedles have been produced using plastic, or biodegradable polymers. Metallic microneedles are expensive, non-biodegradable and brittle. Polymer microneedles overcome the limitations of silicon and metal microneedles and bring multiple advantages such as low cost, mechanical strength, and safety in case of accidental breakage of the needle. Rapid-dissolving sugars and polysaccharides have been used to prepare dissolvable microneedles. Microneedles of dextrin can be prepared without any special fabrication. However, there are problems such as caramelization and difficulties in handling of molten sugar.

Table 1. List of materials used for preparation of microneedles.

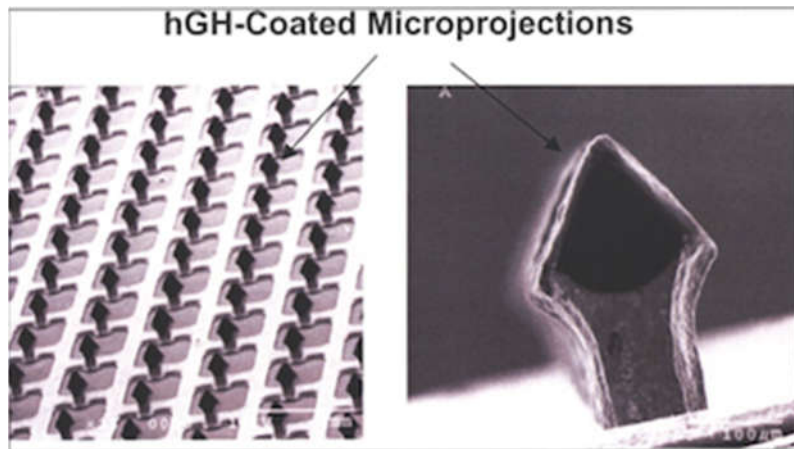
	Synthetic Polymers		
Metals	Biodegradable	Non-biodegradable	Neutral Polymers
silicon	polyactic acid (PLA)	polyvinyl acetate (PVA)	thermoplastic starch
stainless Steel	polyglycholic acid (PGA)	alginic acid	carboxy methal cellulose
titanium	polyactide-co-glycolic acid (PLGA)	Gantrez AN-139, a coploymer of methylvinyl ether and maleic anhydride (PMVE/MA)	amylopectin
mesoporous silicon	polycarbonate	Carbopol 971 P-NF	dextran, galactose, chondroitin sulfate
	polyvinyl pyrrolidone (PVP)		

2.2 Fabrication

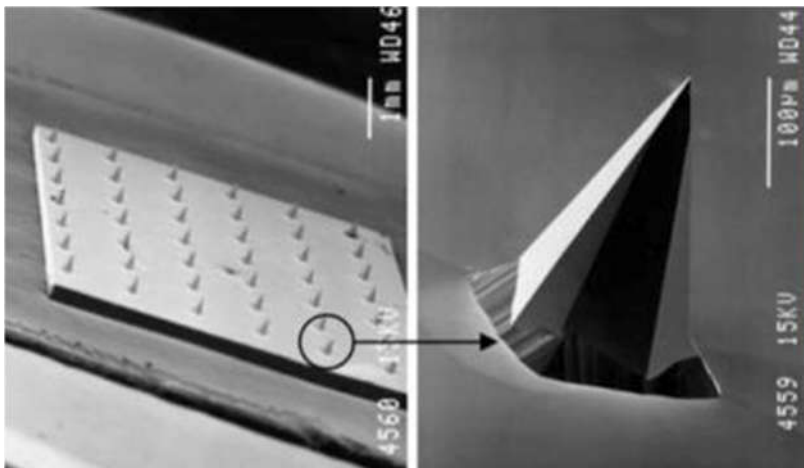
Microneedles can be produced by employing micro-electromechanical systems (MEMS) utilizing three basic processes: deposition, patterning and etching. First, thin films, which are between a few nanometers to about 100 micrometers thick, are created by deposition. Then by a process of patterning, a pattern is transferred onto the photosensitive film using lithography by selective exposure to a radiation source. This process can involve

photolithography, electron beam lithography, ion beam lithography or X-ray lithography. Finally, in the process of etching, dry etching or wet etching, strong acid or mordant is used to cut into the unprotected parts of a material's surface to create a design.

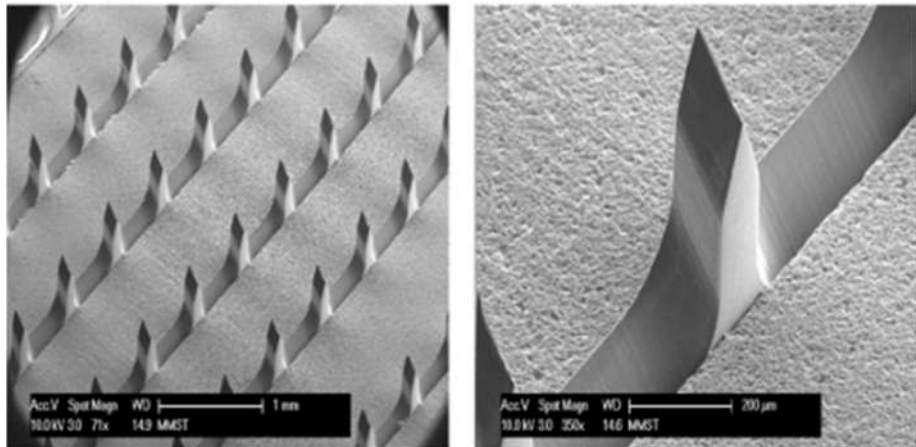
Microneedles are classified as in-plane, out-of-plane or a combination of both, as seen in *Figure 5*. Considering the in-plane designs (*Figure 5a*), the microneedles are parallel with the machined surface of the substrate (e.g. Si wafer); the major advantage of in-plane microneedles is that they can be easily and accurately controlled during the fabrication process to produce various lengths, whereas in out-of-plane designs (*Figure 5b*), the microneedles are perpendicular to the fabrication surface of Si wafer, and are easier to produce in arrays than in-plane designs [7].



(a)



(b)



(c)

Figure 5: Scanning electron microscope (SEM) images of (a) in-plane MNs, (b) out-of-plane MNs and (c) combined in-plane and out-of-plane MNs. (Copyright Elsevier) [7]

The tip shape of the microneedle is important for skin penetration because it will dictate the ease of application. microneedles can be classified on the basis of overall shape and tip (Figure 6), ranging from cylindrical, rectangular, pyramidal, conical octagonal to quadrangular, along with different needle lengths and widths. A sharper microneedle has a higher potential for penetrating the skin, but a larger tip diameter requires a higher insertion force, which may lead to bending or breaking of the needles in the skin. In addition, the shape of the tip of a hollow microneedle is essential for controlling the flow rate. The flow from a blunt-tip microneedle supports the application of a hollow microneedle less than a bevel-tip microneedle because a blunt-tip microneedle compacts the skin and thus has higher risk of clogging (the pathway of the hollow microneedle is blocked), which is one of the limitations for a hollow microneedle. To overcome this problem, the microneedle should have a very sharp tip with the bore of the microneedle off center or on the side of the microneedle. Increasing the number of bores in a hollow microneedle will increase the flow rate; nevertheless, this results in decreased strength of a microneedle and a reduction in sharpness [7].

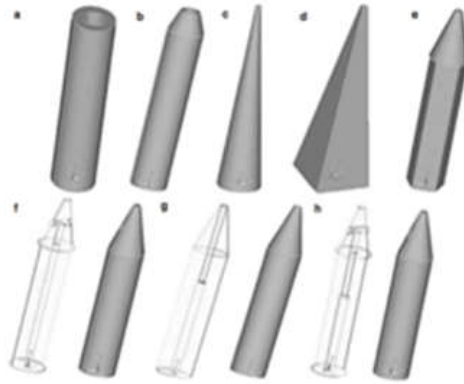


Figure 6. Shapes of microneedle (a) cylindrical, (b) tapered tip, (c) canonical, (d) square base, (e) pentagonal-base canonical tip, (f) side-open single lumen, (g) double lumen, (h) side-open double lumen (Copyright 2011 Ashraf) [7]

2.3 Microneedle Patches

The patches have four main components: a microdrug reservoir, a micropump, a micro flow sensor, and microneedles. Microdrug reservoirs contain the drug to be delivered through the skin. Micropumps inject the drugs from the reservoir to the needles. A flow sensor provided with a control system is used to monitor, control and regulate the flow of drugs through the needles. Microneedles act as the pathway to deliver drugs [1].

2.4 Drug Delivery

In the early research on microneedles, there is an approach known as “poke and patch” (see *Figure 7a*). An array of solid microneedles is pierced through the skin and then a medicated patch is applied to the treated skin surface. This method is similar to the traditional patches and may incorporate agents to help keep the pores open for a longer time to extend the duration of drug delivery. However, there is a safety concern of the breakage of microneedles in the skin if some hard and non-dissolvable materials are used to fabricate the microneedles.

Subsequent research focuses on the “poke and coat” method (see *Figure 7b*). The skin is pierced by solid microneedles coated with a drug solution before the release of the drug. The microneedles are coated using a dip-coating method. However, a limited amount of drug (only about 1 mg) could be coated over the microneedles and optimization is required for uniform coating in this method. Similarly, there will be the possibility of breakage of the hollow microneedles if certain materials are used.

Further research resulted in the development of a “poke and release” approach (*Figure 7c*). This kind of microneedle is made from polymers and polysaccharides that either slowly dissolve or degrade after administration. The advantage of this approach is that the drug release rate can be modulated using a variety of different dissolvable materials. However, a large amount of drug delivery is not feasible. Later, a new method, “poke and flow”, was developed (*Figure 7d*). After piercing the skin, the drug is allowed to flow through hollow microneedles. Once again, the microneedle may break with the non-dissolvable materials. Moreover, there may be a possibility of clogging (mentioned in the 2.2 fabrication), which may lead to ineffective applications.

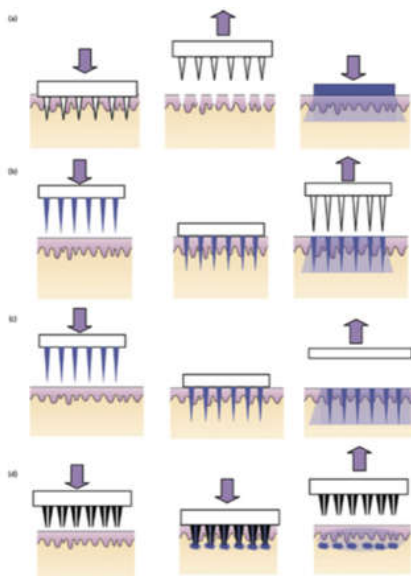


Figure 7. Approaches for drug delivery by different designs of microneedles. (a) “poke and patch”, (b) “poke and coat”, (c) “poke and release”, (d) “poke and flow”. (Copyright 2008 Elsevier) [5]

There are trade-offs when selecting a microneedle delivery method. The “coat and poke” and the “poke and dissolve” approaches are the most widely studied and have the advantage that the microneedle plaster is worn for just a few minutes and then discarded. For the Chinese medicated plasters, that means patients do not need to fix the plaster with a gauze bandage, which is the traditional method. The poke-and-dissolve approach also generates no sharp waste. However, these patches are constrained by the small size of the microneedle tips that generally limit dosing to less than <10 mg and preferably less than <1 mg. This may be an important consideration for many drugs with higher dosing ranges such as most Chinese medicine.

The “poke and patch” and the “poke and release” approaches can deliver larger doses because the drug is contained in a patch backing that delivers the drug continuously. The “poke and patch” method requires a two-step process and that may be cumbersome for some patients. In Chinese clinics, however, because of all the procedures going on, patients may not care about one more step as long as it will make the drug more effective. The “poke and release” method requires patients to wear a patch with in-dwelling microneedles for extended periods of time. Since the patients need to move with the large dose traditional Chinese medicated plasters, if the microneedles are not dissolvable there may be a possibility of breakage of the microneedles in the skin. In future research of microneedles, new dissolvable materials may be adequate, with degrading periods corresponding to the drug delivery time.

2.5 Microneedle Rollers

The microneedle Roller is the advanced application of microneedle. In this scenario, microneedles are mounted on a cylindrical surface that can be rolled across the skin. The microneedle rollers were introduced to treat large skin areas, and are most commonly used to cause microdamage to the skin, which induces collagen production

during the skin repair process for cosmetic purposes [16]. In modern medicine, microneedle has been used to increased permeability for transdermal drug delivery for pharmaceutical and cosmetic applications [9]. Microneedle rollers are desirable in Chinese clinics for transdermal drug delivery with multiple advantages demonstrated in experimental data.

3. Evaluations and Discussion

3.1 Permeability

Researches potently reveal the primary advantage for drug delivery---high drug delivery efficiency. Due to its hydrophilic nature, topical 5-fluorouracil (5-FU), which is approved for the treatment of superficial basal cell carcinoma and actinic keratosis, suffers from poor skin permeation. In an *in vitro* permeation test by the University of Texas at Austin, the feasibility of using microneedles to increase the skin permeability of 5-FU was tested. The hairless full thickness dorsal skin from mice sample was placed onto the flat surface of a balance, and the microneedle roller was rolled in 4 perpendicular lines over the skin surface, 5 times each for a total of 20 times, with an applying pressure of 600–800 g, which was constantly measured using the balance. There are 192 needles (500 mm in length, 50 mm in base diameter) on the roller. The skin was then mounted onto the Franz diffusion cells (device used to conduct the permeation studies with and without MNs) with the epidermis side facing upward. The donor compartment was loaded with 400 μg of 5-FU in 400 mL of PBS (pH 7.4, 10 mmol/L) while the receiver compartment contained 5 mL of PBS (pH 7.4, 10 mmol/L) and was maintained at 37 °C. At pre-determined time points (1, 2, 3, 6 and 18 h), samples (100 mL) were withdrawn from the receiver compartment and immediately replenished with fresh medium. The amounts of 5-FU at different time points are then recorded and shown in the graph. As *Figure 8* shows, the permeation of 5-FU through intact full thickness mouse skin is limited, with a transdermal flux of $8.93\pm 4.55\mu\text{g}/\text{cm}^2/\text{h}$. However, the permeability of 5-FU through mouse skin that was pretreated with microneedles was significantly higher, with a flux of $39.75\pm 18.50\mu\text{g}/\text{cm}^2/\text{h}$ [18]. Clearly, pretreatment of mouse skin with solid microneedles significantly increased the permeability of 5-FU through the skin. As a result, microneedle is an ideal method to increase the efficiency of transdermal drug delivery. The fact that microneedle system can be applied to both small and large molecule makes it more favorable.

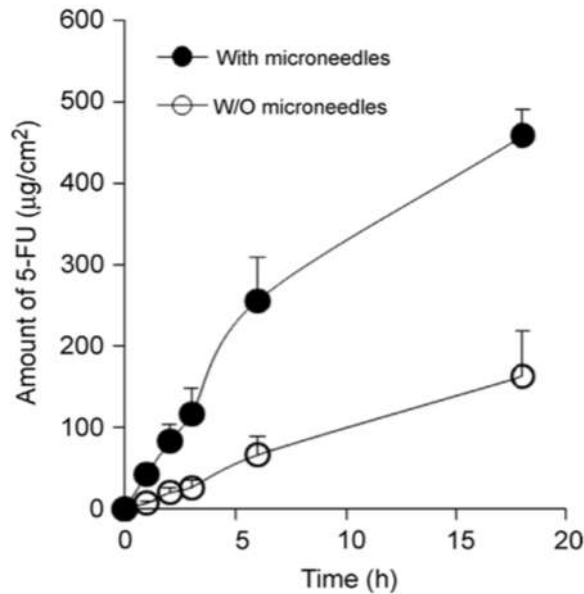


Figure 8. The amount of 5-FU in an aqueous solution diffused through full thickness mouse skin treated (●) or not treated (○) with microneedles as a function of time. Data shown are means \pm 7S.E.M (n1/43). At all the time points tested, with the exception of 0 h, the values between the microneedle-treated and -untreated groups are significantly different. [18]

In other studies, the relationship between the force of microneedle insertion and permeability has been investigated. MNs of 1200 μm and 1500 μm are used to conduct in vitro permeability experiments on porcine skin, using insulin.

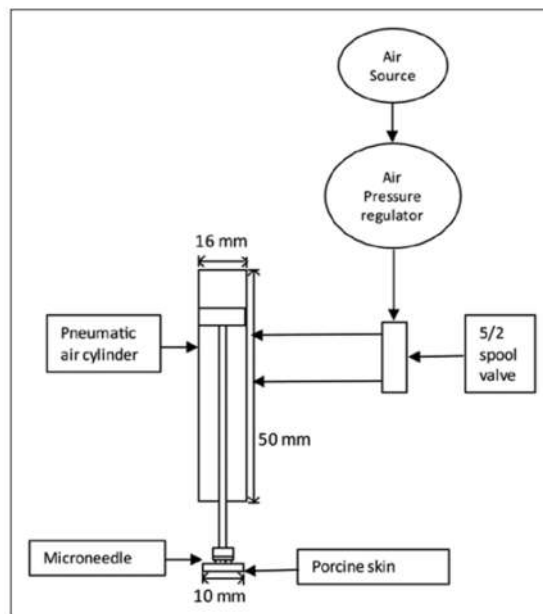


Figure 9. Schematic diagram of the force device equipment setup, which is used to apply a specified force on MNs for a given time duration. [19]

An in-house system has been manufactured specifically to provide a specific force on MNs for insertion. Following parts are used to construct the device (Figure 9): a double acting roundline cylinder, a 16 × 50 mm pneumatic air cylinder, an air pressure regulator, and a directional control valve (5/2 spool valve) [19]. Compressed air pressure is used to activate the pneumatic cylinder in the system, which acts like a piston. The amount of compressed air running through the system is controlled by an air pressure regulator, operated by a spool valve lever. The system is fitted with a directional control valve that has 3 settings. When the lever is in the central position (neutral position), the internal shaft of the pneumatic cylinder is free moving as the spool valve is open and under atmospheric pressure. In this position, the MNs can be attached to the holder and placed on top of the porcine skin without exerting force. When the lever is moved to the “on” position, the regulated compressed air is guided to the top of the pneumatic cylinder. This forces the air piston to move down as the air flows through and maintains a constant force onto the porcine skin for a required period of time. Turning to the “off” position, the regulated compressed air is channeled to the bottom of the pneumatic cylinder, which forces the air piston to move up to the top, with air flowing through. The imposed force and the MNs are perpendicular to the skin sample at the point where the MN contacts the sample [19]. Subsequently, the force exerted by the insertion device on the MN patch is calculated using the following equation:

$$F = \frac{p\pi(d_1^2 - d_2^2)}{4}, \quad (1)$$

where F is the force exerted (N), p is the air pressure (N/m²), d_1 is the bore piston diameter (m), and d_2 is the piston rod diameter (m).

The force device is used to apply different pressures onto the surface of the porcine skin (0.10, 0.20, 0.30, 0.35, and 0.40 MPa) which are converted to the total forces of 17.3 N, 34.6 N, 51.8 N, 60.5 N, and 69.1 N, respectively. Since the 1500 and 1200 MN patches consist of 31 and 43 MNs on each patch, the force acting on 1 individual MN is approximately 2.2 N and 1.6 N, respectively, assuming that the applied force is aligned in the same direction as the MNs and perpendicular to the skin. Processed porcine skin cut into 1.5 cm² pieces stored in a petri dish was aligned below the MN. The spool valve was set to the “on” position, and a constant force of MN was inserted into porcine skin for 5 minutes. The skin sample was then placed onto the receptor chamber of a diffusion cell. A total of 70 µg of insulin containing solution was placed into the donor chamber and samples extracted from the receiving compartment at time points of 1, 2, 3, and 4 hours [19].

The amount of insulin permeated for passive diffusion and MN insertion forces of 17.3 N and 34.6 N was almost undetectable. This may be due to the skin only being buckled or slightly pierced, therefore sufficient pathways cannot be created to let the insulin molecules pass through. Figure 10 illustrates the amount of insulin permeated when MN penetration forces of 51.8 N, 60.5 N, and 69.1 N were used to pretreat the porcine skin samples. The ratio of the insulin in the receiving compartment and donor compartment was shown to increase from 0.2% to 37.1%. The results show that after 4 hours the amount of insulin permeated was approximately 3 µg and 25 µg, respectively, for 60.5 N and 69.1 N insertion forces, but were nearly negligible when an insertion force of 51.6 N was applied. There was almost an 8-fold increase in the amount permeated using the larger insertion force. More important, it confirms that an increase in the insertion force results in an

increase in the amount of insulin permeated. The lack of insulin permeated when insertion forces less than 51.8 N were used could be the result of the “bed of nails” effect (explained in 3.2). Although the stratum corneum layer is likely to be disrupted, the high density MNs may not be able to pierce through the whole epidermis due to the elastic properties of the skin, greatly affecting the permeation of insulin. This resulted in a lower concentration of insulin, compared to the less dense MN.

As a result, it is shown that an increase in insertion force increased the amount of insulin permeated through porcine skin. It can also reveal that the insertion force of the microneedle is essential: An insufficient force may not help the insulin to pass through the skin. But, on a detectable level, the length of the MN and the force applied on the MNs are important factors that can greatly affect the permeation.

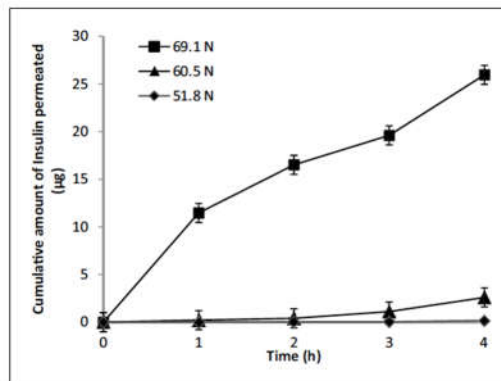


Figure 10. Cumulative amount of insulin (μg) permeated with a MN insertion force 51.8 N, 60.5 N, and 69.1 N for over 4 hours ($n = 3$). The amounts of insulin permeated for MN insertion forces of 17.3 N and 34.6 N and passive diffusion were undetectable. [19]

As the old population is one of the main group of beneficiaries and many old people consider Chinese medicine as a daily application, it is important to research into the efficiency of self-application. It shows that it is not hard for patients to use the simple equipment at home on their own before applying different drugs on their skin to relieve pain in the target area. In a study investigating the potential of microneedles for use in older populations by Queen’s University, 100% of the participants are willing and able to use the microneedle patches on their own after the instructions of the researchers [12].

3.2 Microneedle Mechanics

The advanced version of a microneedle array is the microneedle roller. Studies show that the force required to insert a microneedle array into the skin has a positive proportional relationship to the surface area of contact between the skin and the microneedle tip, as well as with the number of microneedles—this is known as the “bed-of-nails” effect. Thus, the planar microneedle patch is limited by the number of microneedles of a given sharpness that can be effectively pressed into the skin. However, the microneedle roller is able to treat much larger skin areas because only part of the microneedles is inserted at any given time when the roller is rolling across the skin [9].

In an experiment examining the mechanics of microneedle rollers, researchers pressed a 100-microneedle array against full thickness porcine skin with a 5N force, corresponding to 0.05 N per needle, and found that the skin was not pierced (shown in

Figure 11a). In contrast, when a roller covered with microneedles of the same geometry was applied to the skin with a force of 5N, or 0.17N per needle, the row of microneedles in direct contact with the skin penetrated fully, and the two adjacent rows of microneedles also penetrated, shown in *Figure 11b* [9].

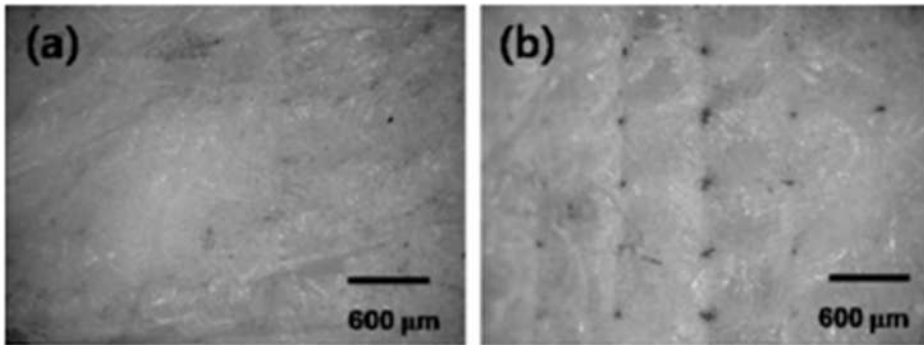


Figure 11. Human cadaver skin showing the sites of microneedle insertion. Microneedles measuring 600 μm in height and 250 μm in base diameter were pressed against the skin with a force of 5 N either (a) on a flat patch or (b) on a cylindrical roller. Microneedles on the roller were able to pierce the skin, whereas microneedles on the patch were not. [9]

Both solid microneedle arrays and rollers are close to the traditional method of medicated plasters, but contain an additional step. Both methods bring about significant increase in transdermal drug delivery. The earlier microneedle rollers were made of metals. In order to prevent the possibility of the breakage inside the skin, other dissolvable, hard materials were used such as biodegradable polylactic acid (PLA). For the dissolvable microneedle rollers such as water-soluble carboxy-methyl-cellulose (CMC), only one step is required.

Although larger insertion force can increase the permeability in some degree, the insertion force should be further studied as they may cause potential hazard of breaking microneedle. In one of the study considering insertion force and fracture force of microneedle, experimental measurements and theoretical modeling have shown that microneedle insertion force increases as a linear function of needle tip cross-sectional area, and increase strongly with increasing wall thickness and may increase modestly with increasing wall angle and tip radius. Data analysis suggests that skin does not deform into the lumen of hollow needles during insertion. Comparison of insertion forces to fracture forces showed that fracture forces were almost always greater than insertion forces over the range of geometries considered and that margins of safety of at least five-fold could be achieved using needles of small tip radius and large wall thickness [21].

3.3 Kinetics of Skin Resealing

The life-expectancy of the holes created by the microneedles need to be considered if the microneedle system will be used in the clinics. The research on the microneedle roller shows the lifetime of holes made in the skin using microneedles. After microneedles were applied across the forearm of a human subject, skin puncture holes were non-invasively imaged by applying a liquid bandage at 0 min, 15 min and 30 min after microneedle treatment to create an inverse replica of the skin surface, thereby showing the holes in the skin (*Figure 12*). The initial holes had a diameter of approximately 100 μm and appeared to be relatively deep, although the tip was poorly copied, suggesting that the bottom of hole

was wet due to the interstitial fluid in the skin. After 15 min, the opening of the hole was almost the same diameter, but its apparent depth had decreased, presumably due to elastic recoil of the tissue and possible active repair processes [9]. After 30 min, the hole was smaller still [9]. In the traditional Chinese clinics, these two methods can be applied, as people not only go to the clinics to cure pain, but also to maintain their health. Patients spend a long time in the clinic with each course of needle therapy, massage, or other traditional healing method. As a result, the doctor may use the novel equipment to increase the transdermal efficacy during the suitable time range.

Results from other studies indicate that in the absence of occlusion, all microneedle treated sites recovered barrier properties within two h, while occluded sites resealed more slowly with resealing windows ranging from three to 40 h depending on microneedle geometry. Results also show that occlusion significantly retards skin barrier resealing after microneedle treatment which corresponds to the case of microneedle patches. Then conclusion can be drawn that the use of microneedle is not limited inside the clinics with an extended life-expectancy of the pores. For future research, special agents and combinations may be found to help keep the pores open for a required period of time to idealize drug delivery process.

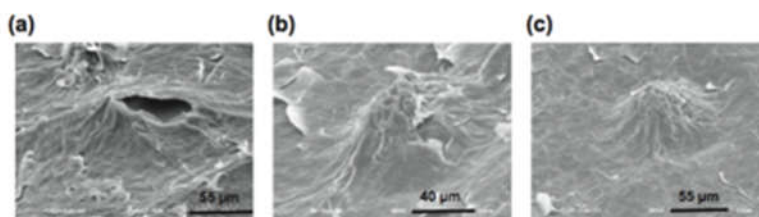


Figure 12. Scanning electron micrographs of inverse molds showing changes in hole size after inserting and removing microneedles into the skin of a human subject at (a) 0 min, (b) 15 min and (c) 30 min after microneedle application. As an inverse mold, the height of the protrusion in the image is a measure of the depth of the hole in skin. [9]

3.4 Drug Delivery

Comparing the two methods, microneedle rollers have advantages over the arrays. That is, microneedle rollers cover a larger treatment area and have a higher level of simplicity. For the future development of microneedle rollers, the “poke and patch” method for solid microneedle may be combined to simplify the method. The hollow microneedle roller may be able to apply medicine and create pores at the same time. The mixture of Chinese medicine can be added inside the roller, and the rotation of the roller may help the liquid and powder to mix better and therefore increase the efficiency of the drug.

For the hollow microneedles, there are multiple application devices available for use for traditional Chinese medicine (see Figure 13). The tiny patches may be used at different points on the body according to the Chinese iatrology although there will still be the possibility of breakage of the needles. The reusable devices can act as an alternative to both sprays and Chinese plaster in order to apply a liquid drug to different junctions and points. However, the sophisticated MEMS fabrication process is impractical for doctors in the clinics to prepare the patches containing different drug mixtures. In addition, the possible fracture caused by patients’ movement with microneedle patches should also be fully considered.

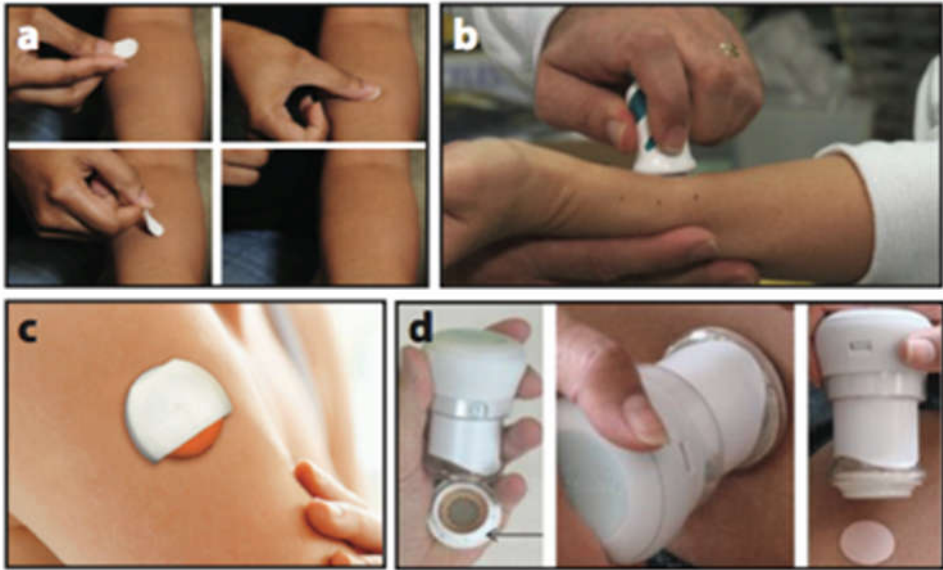


Figure 13. *Microneedle patch application to the skin. (a) Some microneedle patches are designed for manual application to the skin without an applicator. Other designs require a high-velocity applicator either as a separate, reusable device (b, d) or as a disposable, integrated part of the patch (c). [10]*

For coated and dissolving microneedles, small doses are not suitable for Chinese medicated plasters. The concept of coated microneedles may be applied to simplify the method of microneedle rollers. However, precise coating fabrication is required for the successful application of coated microneedles. Thus, a more precise and effective fabrication method should be developed in the future.

To overcome the limitations of solid, hollow, and dissolving microneedle patches, a large dose patch with safe needles should be created. Using the hydrogel-forming microneedles, delivered doses of drugs and biomolecules are no longer limited to what can be loaded into the needles themselves. Hydrogel-forming microneedles are microneedle arrays, prepared under ambient conditions, which contain no drug themselves. Instead, they rapidly imbibe skin interstitial fluid upon insertion to form continuous conduits between the dermal microcirculation and an attached patch-type drug reservoir. Such microneedles act initially as a tool to penetrate the stratum corneum barrier. Upon swelling, they become a rate controlling membrane. The fluid uptake range in one hour was 0.9–2.7 μL which is of the same order of magnitude as the effective rates of interstitial fluid uptake for hollow microneedles and microdialysis. That means, one of the limitations of dissolvable and coated microneedles---low doses---has been partially solved [5]. Other advantages of hydrogel-forming microneedles are that they can be fabricated in a wide range of patch sizes and geometries, can be easily sterilized, resist hole closure while in place, and are removed completely intact from the skin, which overcomes the limitation and safety concern of solid and coated microneedles [5]. This potential microneedle is a more advanced and suitable version for traditional Chinese plasters, which need sustained delivery for a relatively long period of time (within 24 hours) and also have high safety requirements of the needles.

4. Conclusion

Since the drug delivery of Chinese medicated plasters has high requirements—effective large dose delivery, appropriate self-application, and safety with daily movement—different areas of research should be considered together in the case of Chinese medicated plasters. From this paper, conclusions are partially available: the application of microneedle leads to increased skin permeability; the increased insertion force over certain threshold values leads to increased permeability; “bed-of-nails” effect makes the self-application of rollers easier than patches; the life-expectancy of the channels created by microneedle can be modulated; different method of transdermal drug delivery are applicable in different situations. In the future, integrated and overlapping studies regarding the above area should be finished.

However, more effort and research should focus on all the potential areas. The safety concerns about the possibility of breakage of certain needles inside the skin should be solved. More effort should be directed at widening the spectrum of materials that can be used to fabricate the useful transdermal drug delivery systems. Also, a simpler fabrication process may be developed, to make it easier for doctors in the traditional Chinese clinics to prepare patches with different drugs inside. Above all, there should be more research into the mechanics of microneedle patches when the patients move with the patches on their bodies to make the microneedle technology widely used in the field of Chinese medicine.

5. References

- [1] Antony, Rahul, Nidhin Sreekumar, and N. Selvaraju. "Overview of Micro Needle System: a Third Generation Transdermal Drug Delivery Approach." *Microsyst Technol* (2014): n. pag. Springer. Web.
- [2] Bariya, Shital H., Mukesh C. Gohel, and Tejal A. Mehta. "Microneedles: An Emerging Transdermal Drug Delivery System." Review. *Journal of Pharmacy and Pharmacology* (n.d.): 11-29. Web.
- [3] Garland, Martin J., and Ryan F. Donnelly. "Microporation Using Microneedle Arrays." *Percutaneous Penetration Enhancers Physical Methods in Penetration Enhancement*. By Emima McAlister. N.p.: Springer, 2017. 273-304. Web.
- [4] Gill, Harvinder S., and Mark R. Prausnitz. "Coated Microneedles for Transdermal Delivery." *NIH Public Access* (2007): 1-22. Web.
- [5] Ita, Kevin. "Transdermal Delivery of Drugs with Microneedles-Potentials and Challenges." Review. *Pharmaceutics* (2015): 90-105. Web.
- [6] Lee, Jeong Woo, Jung-Hwan Park, and Mark R. Prausnitz. "Dissolving Microneedles for Transdermal Drug Delivery." *NIH Public Access* (2008): 1-23. Web.
- [7] McMillan, Hannah, Karen Mooney, and Ryan F. Donnelly. "Fabrication of Microneedles." *Percutaneous Penetration Enhancers Physical Methods in Penetration Enhancement*. By Thakur Raghu Raj Singh. N.p.: Springer, 2017. 305-24. Web.
- [8] Nguyen, Ita, Parikh, Popova, and Bair, D.A. "Transdermal Delivery of Captopril and Metoprolol Tartrate with Microneedles." *Drug Deliv. Lett* (2014):236-243. Web.
- [9] Park, Jung-Hwan, Seong-O Choi, and Soonmin Seo. "A Microneedle Roller for Transdermal Drug Delivery." *European Journal of Pharmaceutics and Biopharmaceutics* (2010): 282-89. Web.
- [10] Prausnitz, Mark R. "Engineering Microneedle Patches for Vaccination and Drug Delivery to Skin." Review. (2017): n. pag. Web.
- [11] Prausnitz, Mark R. "Microneedles for Transdermal Drug Delivery." *ELSEVIER* (n.d.): 582-587. Web.

- [12] Quinn, Helen L., Carmel M. Hughes, and Ryan F. Donnelly. "In Vivo and Qualitative Studies Investigating the Translational Potential of Microneedles for Use in the Older Population." (2017): n. pag. *Springer*. Web.
- [13] Roberts, Michael S. "Transdermal Patches: history, Development and Pharmacology." Review. *British Journal of Pharmacology* (n.d.): 2179-2209. Web.
- [14] Verma, Garima. "Transdermal Drug Delivery System, Advance Development and Evaluation-A Review." *International Journal of Pharmaceutical Sciences and Research* 8.2 (2017): 385-400. Web.
- [15] Kim, Y., Park, J., & Prausnitz, M. R. (2012). Microneedles for drug and vaccine delivery. *Advanced Drug Delivery Reviews*, 64(14), 1547-1568. doi:10.1016/j.addr.2012.04.005
- [16] Li, N., Wang, N., Wang, X., Zhen, Y., & Wang, T. (2016). Microneedle arrays delivery of the conventional vaccines based on nonvirulent viruses. *Drug Delivery*, 23(9), 3234-3247. doi:10.3109/10717544.2016.1165311
- [17] Matteucci, M., Casella, M., Bedoni, M., Donetti, E., Fanetti, M., Angelis, F. D., .Fabrizio, E. D. (2008). A compact and disposable transdermal drug delivery system. *Microelectronic Engineering*, 85(5-6), 1066-1073. doi:10.1016/j.mee.2007.12.067
- [18] Naguib, Y. W., Kumar, A., & Cui, Z. (2014). The effect of microneedles on the skin permeability and antitumor activity of topical 5-fluorouracil. *Acta Pharmaceutica Sinica B*, 4(1), 94-99. doi:10.1016/j.apsb.2013.12.013
- [19] Cheung, K., Han, T., & Das, D. B. (2014). Effect of Force of Microneedle Insertion on the Permeability of Insulin in Skin. *Journal of Diabetes Science and Technology*, 8(3), 444-452. doi:10.1177/1932296813519720
- [20] Gupta, J., Gill, H. S., Andrews, S. N., & Prausnitz, M. R. (2011). Kinetics of skin resealing after insertion of microneedles in human subjects. *Journal of Controlled Release*, 154(2), 148-155. doi:10.1016/j.jconrel.2011.05.021
- [21] Davis, S. P., Landis, B. J., Adams, Z. H., Allen, M. G., & Prausnitz, M. R. (2004). Insertion of microneedles into skin: measurement and prediction of insertion force and needle fracture force. *Journal of Biomechanics*, 37(8), 1155-1163. doi:10.1016/j.jbiomech.2003.12.010



r-bonacci Numbers, r-Lucas Numbers and Their Identities

Yuheng Wu

Author Background: Yuheng Wu grew up in China and currently attends The Affiliated High School of South China Normal University in Guangzhou, China. His Pioneer seminar topic was in the field of mathematics and titled "Fibonacci Numbers and Visual Proofs."

Abstract

The well-known Fibonacci numbers are defined by $f_0 = f_1 = 1$ and $f_{n+2} = f_{n+1} + f_n$. Similarly, the Lucas numbers are defined by $l_0 = 2, l_1 = 1$ and $l_{n+2} = l_{n+1} + l_n$. There are various kinds of generalizations of the two famous numbers, and so far, mathematicians have shown a great many beautiful identities related to them. In this paper, we focus on two specific kinds of generalizations, the r-bonacci Numbers and r-Lucas Numbers, with a parameter r. We prove several identities of the two numbers, and derive the exact formula. Lastly, we present some conjectures found during our research.

Instead of pure algebraic calculation, we utilize combinatorial explanation to prove most of the identities. That is because although calculation and induction easily prove most of the identities, combinatorial explanation gives more insights into what the numbers and identities represent. For instance, the nth term of the Fibonacci numbers can be viewed as the number of ways to tile a board of length n with squares and dominos (which is a special case of an identity proved in this paper), and we can see that most identities are actually rearranging and classifying the decompositions. In this paper, we regard the nth term of the r-bonacci numbers as the number of ways to cover a board of length n with strips of length 1 and r, and prove most of the related identities by conditioning on different ways of covering.

Keywords: generalized Fibonacci numbers, generalized Lucas numbers, combinatorial proof

1. Definition

1.1 Cells, Boards, Bracelets, Tiles, Tilings, Squares and r-minos

A *cell* is a square block. A *board of length n* is a strip of n cells. For a board of length n we number the cells 1, 2... n, from left to right, and refer them to as the first cell, second cell, etc.

A *bracelet of length n* is a board of length n with cells 1 and n connected. For convenience, cell n refers to cell 1, cell n+2 refers to cell 2, etc.

A *tile of length k* also refers to a strip of k cells. A *square* is a tile of length 1, and an *r-mino* is a tile of length r.

To *tile a board (bracelet)* means to use the given tiles to completely cover the board, without overlapping or leaving uncovered cells. A *tiling* is a legal solution to tile a board (bracelet).

1.2 r-bonacci Numbers

The r-bonacci numbers, which we will call the x-sequence, are defined as the number of ways to tile a board of length n with squares and r-minoes, where r is an integer greater than 1. We define $x_0 = 1$ and $x_k = 0$ for $k < 0$. The first few numbers in the x-sequence are listed in Table 1.

Table 1: First few numbers in the x-sequence

n	-1	0	1	2	...	r-1	r	r+1	...
x_n	0	1	1	1	...	1	2	3	...

Take $r=4$, for example, then x_n represents the number of ways to tile an n-board with tiles of length 1 and 4. The first few number are listed in Table 2. Figure 1 shows all four different tilings for a 6-board with squares and 4-minoes.

Table 2: First few numbers in the x-sequence when r=4

n	0	1	2	3	4	5	6	7	8	9	10	11	12	...
x_n	1	1	1	1	2	3	4	5	7	10	14	19	26	...



Figure 1: Four different tilings for a 6-board with squares and 4-minoes

We consider an r-mino located at cell k, when it covers cell k, k+1...k+r-1. Easy to see that an r-mino cannot be located at cell k if $k > n-r+1$.

1.3 r-Lucas Numbers

The r-Lucas numbers, which we will call the y-sequence, are defined by the number of ways to tile a bracelet of length n with curved squares and r-minos. Specially, we define $y_0 = r$ and $y_k = 0$ for $k < 0$.

The first few numbers in the y-sequence are listed in Table 3.

Table 3: First few numbers in the y -sequence

n	-1	0	1	2	...	$r-1$	r	$r+1$...
y_n	0	r	1	1	...	1	$r+1$	$r+2$...

Figure 2 shows all five different tilings for a 4-bracelet with curved squares and 3-minoes. Note that there are four ways to tile the bracelet with a square and a 3-mino depending on the location of the square.

We define the *first* tile to be the tile that covers cell 1. The second tile is the next tile in the clockwise direction, and so on.



Figure 2: Four different tilings for a 4-bracelet with curved squares and 3-minoes

1.4 Breakability and Phases

Given a particular tiling, we consider a cell *breakable* when it is not connected to the cell following it, and *unbreakable* otherwise. Specially, the last cell of a board is always breakable. Easy to observe that, given a tiling, a cell is breakable if and only if it is a square or the ending cell of an r -mino.

Given a tiling to an n -bracelet, the bracelet is *out of phase* if cell n is unbreakable, and *in phase* otherwise.

2. Identities of the x -sequence

Identity 1 For $n \geq r$, $x_n = x_{n-1} + x_{n-r}$.

Proof Consider the number of tilings for an n -board. On one hand, by definition, it is x_n ; on the other hand, we condition on the last tile. If it is a square, removing it leads to an $(n-1)$ -board. If it is a r -mino, removing it leads to an $(n-r)$ -board. Summing up the two cases leads to $x_{n-1} + x_{n-r}$.

Identity 2 For $n \geq 0$, $\sum_{i=0}^n x_i = x_{n+r} - 1$

Proof Consider the number of tilings for an $(n+r)$ -board with at least one r -mino. First observe that there are x_{n+r} ways to tile the board, and excluding the case in which all tiles are squares gives $x_{n+r} - 1$. Then we condition on the position of the last r -mino. When it is located at cell k , there are x_{k-1} ways to tile the board in front of it. The tiles behind it

can only be squares, with only 1 way to tile. Summing up all cases leads to $\sum_{i=1}^{(n+r)-r+1} x_{i-1} = \sum_{i=0}^n x_i$.

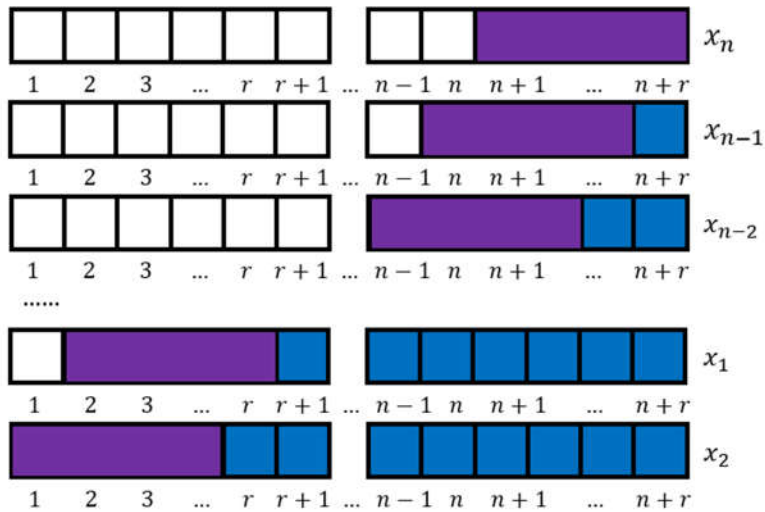


Figure 3: Tile an $(n+r)$ -board with squares and r -minoes and condition on the location of the last

Identity 3 For $0 \leq k < r-1$ and $n \geq 0$, $\sum_{i=0}^n x_{r+i+k} = x_{nr+k+1}$

Proof Consider the number of tilings for an $(nr+k+1)$ -board. On one hand, by definition, it is x_{nr+k+1} . On the other hand, since $(nr+k+1)$ is not divisible by r , there must exist at least one square. Condition on the number of r -minoes after the last square. When there are j r -minoes behind the last square, together they occupy $(rj+1)$ cells, leaving a board of length $(nr+k+1-rj-1)=(n-j)r+k$ to tile. There are $x_{(n-j)r+k}$ ways to tile it. Summing up all cases leads to $\sum_{j=0}^n x_{(n-j)r+k}$. Note that when $i=n-j$, the result is equivalent to $\sum_{i=0}^n x_{ri+k}$. Figure 4 shows the proof.

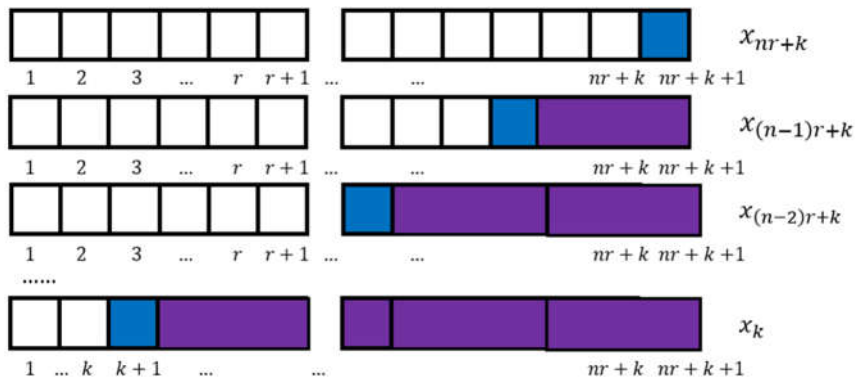


Figure 4: Tile an $(nr+k+1)$ -board with squares and r -minoes and condition on the location of the last square

Identity 4 For $n \geq 0$ $\sum_{i=0}^n x_{r\#(r-1)} = x_{rn+r} - 1$

Proof The proof is similar to the proof to Identity 2. Note that as $(n+r)$ is divisible by r , there exists a case in which all tiles are r -minos. To ensure that at least one square exists, we need to exclude this case. Thus, the right hand side should be $x_{rn+r} - 1$.

By employing Identity 1, 2, 3, we have another algebraic proof.

$$\begin{aligned} & \sum_{i=0}^n r\#_{i+(r-1)} = x_{rn+r} - 1 \\ \Leftrightarrow & \sum_{i=0}^n r\#_{i+(r-1)} + \sum_{k=0}^{r-2} \sum_{i=0}^n r\#_{i+k} = x_{rn+r} - 1 + \sum_{k=0}^{r-2} \sum_{i=0}^n r\#_{i+k} \\ \Leftrightarrow & \sum_{i=0}^{rn+r-1} i \# = x_{rn+r} - 1 + \sum_{k=0}^{r-2} r\#_{k+1} \\ \Leftrightarrow & \#_{n+2r-1} - 1 = \sum_{k=0}^{r-1} r\#_{k+1} - 1 \\ \Leftrightarrow & \#_{n+2r-1} = \sum_{k=0}^{r-1} r\#_{k+1} \end{aligned}$$

The last step is valid since

$$\begin{aligned} & x_{rn+1} + x_{rn+2} + x_{rn+3} + \dots + x_{rn+r} \\ & = x_{rn+2} + x_{rn+3} + \dots + x_{n+r} + x_{rn+r+1} \\ & = x_{rn+3} + \dots + x_{n+r+1} + x_{rn+r+2} \\ & = \dots = x_{n+2r-1} \end{aligned}$$

From the last step of the proof we can derive another identity.

Identity 5 For $n \geq 0$ $\sum_{i=1}^r x_{n+i} = x_{n+2r-1}$

Proof Consider the number of tilings for an $(n+2r-1)$ -board. On one hand, by definition, it is x_{n+2r-1} . This is the right hand side. On the other hand, we condition on the breakability of cell $n+r$:

- (1) Cell $n+r$ is breakable. Then we can divide the board into two boards of length $n+r$ and $r-1$ respectively. Note that $x_{r-1} = 1$, the total number of tilings is $x_{n+r}x_{r-1} = x_{n+r}$
- (2) Cell $n+r$ is unbreakable. Condition on the location of the r -mino covering that cell. When it is located on cell $n+k+1$, removing it gives two boards of length $n+k$ and $r-k-1$,

with $x_{n+k}x_{r-k-1}$ tilings. The r -mino ends at cell $n+k+1+r-1=n+k+r$, so to ensure cell $n+r$ is unbreakable, $n+k+r > n+r$ or $k > 0$. Also $n+k+1 \leq n+r$ gives $k < r$.

When k goes from 1 to $r-1$, $(n+k)$ goes from $(n+1)$ to $(n+r-1)$. $r-k-1 < r$ gives $x_{r-k-1} = 1$. Thus, in total there are $x_{n+1} + x_{n+2} + \dots + x_{n+r-1}$ tilings.

Summing up the two cases gives us $x_{n+1} + x_{n+2} + \dots + x_{n+r-1} + x_{n+r}$ which is the left hand side.

Identity 6 For $n \geq 0$

$$\sum_{k=0}^{\lfloor \frac{n}{r} \rfloor} r^k \binom{n-(r-1)k}{k} = x_n$$

Proof Consider the number of tilings for an n -board. On one hand, by definition, it is x_n . On the other hand, we condition on the number of r -minos. When there are k r -minos, there are $n-kr$ squares, together $n-(r-1)k$ tiles. There are $\binom{n-(r-1)k}{k}$ ways to arrange these tiles. Summing up all cases gives us the right hand side.

Figure 5 shows an example of tiling a 10-board with squares and 3-minosos.



Figure 5: There are $\binom{6}{2}$ ways to arrange four squares and two 3-minosos

Identity 7 For $n \geq 0$ $\sum_{k=1}^n \binom{n}{k} x_{(r-1)k-1} = x_{rn-1}$

Proof Consider the number of tilings for an $(rn-1)$ -board. First we observe that it is x_{rn-1} by definition. Then, note that $(n-1)r < nr-1$, which means there must be at least n tiles. As $rn-1$ is not divisible by r , there exists at least one square. We condition on the number of squares among the first n tiles. When there are k squares, there are $(n-k)$ r -minos, occupying $nr-(r-1)k$ cells, leaving a board of length $(r-1)k-1$. There are $\binom{n}{k}$ ways to arrange the first n tiles, and $x_{(r-1)k-1}$ ways to tile the rest of the board. Multiplying the two numbers and summing all cases up gives us the left hand side.

Identity 8 For $p \geq 0$ and $n \geq (r-1)p$, $\sum_{i=0}^p \binom{p}{i} x_{n-(r-1)i} = x_{n+p}$

Proof Consider the number of tilings for an $(n+p)$ -board. On one hand, it is by definition x_{n+p} . On the other hand, we condition on the number of r -minos among the first p tiles. When there are i r -minos, there are $p-i$ squares, together occupying $ri+p-i=p+(r-1)i$ cells, leaving a board of length $n+p-p-(r-1)i=n-(r-1)i$. There are $\binom{p}{i}$ ways to arrange the first p tiles and $x_{n-(r-1)i}$ ways to tile the rest of the board. Summing up, the product gives us $\sum_{i=0}^p \binom{p}{i} x_{n-(r-1)i} = x_{n+p}$

Identity 9 For $m, n \geq 0$, $x_{m+n} = x_m x_n + \sum_{i=1}^{r-1} x_{m-i} x_{n-r+i}$

Proof Consider the number of tilings for an $(m+n)$ -board. On one hand, by definition, it is x_{m+n} . On the other hand, we consider the breakability of cell m (See Figure 6).

- (1) Cell m is breakable. In this case, the board can be separated into two boards of length m and n respectively. There are $x_m x_n$ ways to tile.
- (2) Cell m is unbreakable. Condition on the position of the r -mino covering it. When it is located at cell $m-i+1$, removing the r -mino gives two boards of length $m-i$ and $n-r+i$ respectively. To ensure cell m is unbreakable, $m-i+1 \leq m$ and $m-i+r > m$, or $1 \leq i < r$.

Summing up all cases gives us $\sum_{i=1}^{r-1} x_{m-i} x_{n-r+i}$

Remark Identity 5 becomes a special case to this identity by replacing n by $n+r$ and m by $r-1$ and taking all " x_{m-i} " term to be 1.

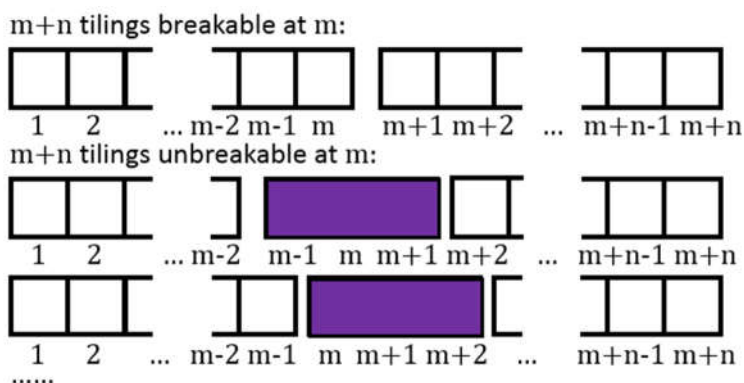


Figure 6: Count $(m+n)$ -tilings base on the breakability of cell m

Identity 10 For $n \geq 0$ and $i_1 i_2 \dots i_r \geq 0$, let $s = \sum_{k=1}^r i_k$

$$\sum_{i_1 i_2 \dots i_r \geq 0} \prod_{k=1}^r \binom{n - s + i_k}{i_k} = x_{rn+r-1}$$

Proof Consider the number of tilings of an $(rn+r-1)$ -board. On one hand, by definition, it is x_{rn+r-1} . This is the right hand side. On the other hand, as $(rn+r-1)$ is not divisible by r , there exists at least one square. More precisely, we may assume that there are d squares and e r -minos, then $rn+r-1 = d+re$. As $rn+r-1 \equiv 1 \pmod{r}$, we have $d+re \equiv r-1 \pmod{r}$ or $d \equiv r-1 \pmod{r}$.

Therefore, we can select $r-1$ squares from the tiling, split the board into r subboards beginning and ending with those squares, and ensure that there are exactly the same

number of squares (may be zero) in any of the sub-boards (The selected squares are not counted).

Now, how many ways can we tile the sub-boards when there are i_1 r-minos in the first sub-board, i_2 r-minos in the second... i_r r-minos in the last sub-board? First, we need to figure out the length of the boards. Assume there are c squares in one of the sub-boards. Then

$$c = \frac{(rn + r - 1) - (r - 1) - r\sum_{k=1}^r i_k}{r} = n - s$$

In the first sub-board, there are i_1 r-minos and in total $n - s + i_1$ tiles. Hence there are $\binom{n-s+i_1}{i_1}$ ways to tile. The number of tilings for the other boards are similar. As we are tiling the r boards at the same time, we need to multiply all cases to get the final result. Therefore, given $i_1 i_2 \dots i_r$, the total number of ways to tile is:

$$\prod_{k=1}^r \binom{n - s + i_k}{i_k} = x_{rn+r-1}$$

Summing all possible i up gives us

$$\sum_{i_1 i_2 \dots i_r \geq 0} \prod_{k=1}^r \binom{n - s + i_k}{i_k} = x_{rn+r-1}$$

which is the left hand side.

Note that the total length of r-minos will be no longer than the total length of sub-boards. Namely, $rs \leq rn$ or $s \leq n$. Other cases should be eliminated. Although i_k are unbounded in the expression (since it is hard to express the bounds explicitly), we can see that when $s > n$, $n - s + i_k < i_k$ gives $\binom{n-s+i_k}{i_k} = 0$. This eliminates all illegal cases.

Figure 7 gives an example of $n=4, r=3$.

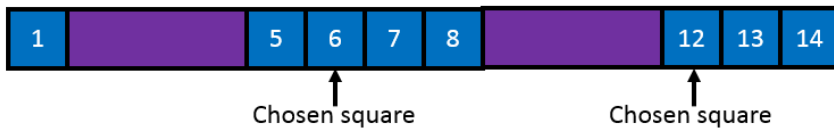


Figure 7: For $n=4$ and $r=3$, the two selected squares split the board into three intervals, in which the number of squares are equal

3. Identities of the y-sequence

Identity 11 For $n \geq r, y_n = y_{n-1} + y_{n-r}$

Proof Consider the number of tilings for an n-bracelet. First, we observe that by definition it is y_n . Then we condition on the tile that covers cell 1. When it is a square, removing it gives an (n-1)-bracelet with y_{n-1} ways to tile; when it is a curved r-mino, removing it gives an (n-r)-bracelet with y_{n-r} ways to tile. Together there are $y_{n-1} + y_{n-r}$ ways.

Identity 12 For $n \geq r, y_n = x_n + (r - 1)x_{n-r}$

Proof Consider the number of tilings for an n-bracelet. On one hand, by definition, it is y_n . On the other hand, we condition on the phase of the bracelet. When it is in-phase, it can be straightened into a board of length n, with x_n ways to tile. When it is out-of-phase, the ending cell of the r-mino covering cell n and 1 can be cell 1, 2... r-1\$, in total r-1 cases. For each case removing the r-mino, the bracelet can be straightened into a board of length n-r, with x_{n-r} ways to tile. Summing up all cases gives us $x_n + (r - 1)x_{n-r}$.

Identity 13 For $n \geq r, y_n + y_{n+r} = (r + 2)x_n + (r - 2)x_{n-r} + x_{n+r-2}$

Proof Consider the number of tilings for a bracelet of length n or n+r. On one hand, it is by definition $y_n + y_{n+r}$. On the other hand, we may create those tilings by matching the tilings to an n-board:

- (1) An in-phase n-bracelet by attaching cell n to cell 1. There are x_n ways.
- (2) An in-phase (n+r)-bracelet beginning with r inserted squares. There are x_n ways.
- (3) An in-phase (n+r)-bracelet beginning with an inserted r-mino. There are x_n ways.
- (4) An out-of-phase (n+r)-bracelet beginning with an inserted r-mino. As the location of the r-mino varies, there are in total $(r - 1)x_n$ ways.

Now let's consider which bracelets we have not created yet. In fact, we are missing out-of-phase n-bracelets and (n+r)-bracelets beginning with several squares followed by an r-mino.

To create an out-of-phase n-bracelet, we need to use an n-board ending with an r-mino, attach cell n to cell 1, and rotate it clockwise. There are x_{n-r} such boards, and as the location of the r-mino varies, there are in total $(r - 1)x_{n-r}$.

To create an (n+r)-bracelet that begin with k squares followed by an r-mino, we need a board starting with k squares, insert an r-mino after the k squares, and attach the last cell to the first one. There are x_{n-k} such boards. When k goes from 1 to r-1 we have in total $x_{n-1} + x_{n-2} + \dots + x_{n-r+1}$ tilings.

Summing up all cases gives $(r + 2)x_n + (r - 1)x_{n-r} + x_{n-1} + x_{n-2} + \dots + x_{n-r+1}$. Further we observe that

$$\begin{aligned} & (r + 2)x_n + (r - 1)x_{n-r} + x_{n-1} + x_{n-2} + \dots + x_{n-r+1} \\ &= (r + 2)x_n + (r - 2)x_{n-r} + x_{n-1} + x_{n-2} + \dots + x_{n-r+1} + x_{n-r} \\ &= (r + 2)x_n + (r - 2)x_{n-r} + x_{n+r-2} \end{aligned}$$

The last step is based on Identity 5.

Identity 14 For $n \geq 0$

$$\sum_{k=0}^n k y_{n-k} = (n+r) x_n$$

Proof We begin by constructing two sets.

Set 1. The set of tilings for an n-board. This set has size x_n .

Set 2. The set of ordered pairs (A,B) where A is a tiling to a k-board and B is an (n-k)-bracelet for $0 \leq k \leq n$. This set has size $\sum_{k=0}^n x_k y_{n-k}$.

Correspondence: We provide a 1-to-(n+r) correspondence between the two sets.

Given an n-board X, we check the breakability of cell k of X. When it is breakable, $X=AB$ in which A is a k-board and B is an (n-k)-board, and we associate the pair (A,B) where B is an in-phase (n-k)-bracelet. Otherwise, $X=ArB$ where A and B are two boards separated by an r-mino. Then we associate the pair (A,rB) where rB is an out-of-phase bracelet.

Specially, when cell k is unbreakable, we let cell k+1 to be the first cell of the bracelet. Then all kinds of out-of-phase bracelets are covered. This counts for n-1 tiling pairs.

When $r=0$, we associate the pair $(\emptyset X)$ where X is an in-phase n-bracelet. When

$r=n$, we associate r pairs of $(X\emptyset)$, since there are r kinds of 0-bracelet.

All together, each n-board tiling generates $(n-1+1+r)=(n+r)$ tiling pairs (A,B). This process can easily be reversed by examining the phase of bracelet B. Thus the process is indeed a bijection, so two sets have the same size.

4. Formula for the Sequences

In this section we derive formulas for the two sequences through similar approaches.

4.1 Formula for the x-sequence

4.1.1 Approximation

We may start with a lemma.

Lemma 1 For $r \geq 2$, the equation $f(x) = x^r - x^{r-1} - 1 = 0$ has exactly one real root in interval $(1, +\infty)$.

Proof Note that

$$f(1) = -1 < 0$$

$$f(2) = 2^r - 2^{r-1} - 1 = 2^{r-1} - 1 > 0$$

So there exists at least one real root between 1 and 2.

Further notice that

$$f'(x) = rx^{r-1} - (r-1)x^{r-2} = rx^{r-2} \left(x - \frac{r-1}{r} \right)$$

This tells that $f'(x) > 0$, or $f(x)$ monotonically increases, in interval $(1, +\infty)$. Clearly f has exactly one real root in it.

We may now present an approximation to the x -sequence.

Identity 15 Let $\theta \in (1, +\infty)$ satisfies $\theta^r - \theta^{r-1} - 1 = 0$ (such θ exists due to Lemma 1), then

$$x_n \approx \theta^n \frac{\theta^r}{\theta^r + (r-1)}$$

Proof We begin by randomly placing squares and r -minos with probability $1/\theta$ and $1/\theta^r$. This is valid since

$$\frac{1}{\theta} + \frac{1}{\theta^r} = \frac{\theta^{r-1} + 1}{\theta^r} = \frac{\theta^r}{\theta^r} = 1$$

Easy to observe that the probability of a tiling board beginning with a particular length- n sequence is $1/\theta^n$.

Furthermore, let q_n be the probability that a random tiling is breakable at cell n . By definition, there are x_n ways to tile an n -board with squares and r -minos. Thus

$$q_n = \frac{x_n}{\theta^n}$$

For cell n to be unbreakable, at least one cell among cell $n-1, n-2, \dots, n-r+1$ must be breakable, and the cell is followed by an r -mino. Thus for $n \geq r$

$$1 - q_n = \frac{q_{n-1}}{\theta^r} + \frac{q_{n-2}}{\theta^r} + \dots + \frac{q_{n-r+1}}{\theta^r}$$

and $q_k = 1/\theta^k$ for $0 \leq k < r$.

Assume $q = \lim_{n \rightarrow +\infty} q_n$, then

$$1 - q = \frac{q}{\theta^r} + \frac{q}{\theta^r} + \dots + \frac{q}{\theta^r}$$

Namely

$$q = \frac{\theta^r}{\theta^r + (r-1)}$$

Thus

$$x_n \approx q\theta^n = \theta^n \frac{\theta^r}{\theta^r + (r-1)}$$

The identity suggests a corollary:

Corollary 1. For θ defined in Identity 15, we have

$$\lim_{n \rightarrow +\infty} \frac{x_n}{x_{n-1}} = \theta$$

4.1.2 Exact Form

Before deriving the exact formula for x_n we introduce a theorem.

Theorem 1 (Vieta Theorem) Assume $\theta_1, \theta_2, \dots, \theta_n \in \mathbb{C}$ are n roots of polynomial $f(\theta) = a_0x^n + a_1x^{n-1} + \dots + a_{n-1}x + a_n$, then

$$\left\{ \begin{array}{l} \theta_1 + \theta_2 + \dots + \theta_n = -\frac{a_1}{a_0} \\ \theta_1\theta_2 + \theta_1\theta_3 + \dots + \theta_1\theta_n + \theta_2\theta_3 + \dots + \theta_{n-1}\theta_n = \frac{a_2}{a_0} \\ \theta_1\theta_2\theta_3 + \theta_1\theta_2\theta_4 + \dots + \theta_{n-2}\theta_{n-1}\theta_n = -\frac{a_3}{a_0} \\ \dots \dots \\ \theta_1\theta_2\theta_3 \dots \theta_n = (-1)^n \frac{a_n}{a_0} \end{array} \right.$$

Proof As $\theta_1, \theta_2, \dots, \theta_n$ are the roots of $f(\theta)$, the polynomial can be expressed as

$$\begin{aligned} f(\theta) &= a_0(\theta - \theta_1)(\theta - \theta_2) \dots (\theta - \theta_n) \\ &= a_0\theta^n - a_0(\theta_1 + \dots + \theta_n)\theta^{n-1} + a_0(\theta_1\theta_2 + \dots + \theta_{n-1}\theta_n)\theta^{n-2} + \dots \\ &\quad + a_0(-1)^n\theta_1\theta_2 \dots \theta_n \end{aligned}$$

Comparing the coefficients yields the Vieta Theorem.

We now derive the exact formula for x_n .

Identity 16. Let $\theta_1, \theta_2, \dots, \theta_r \in \mathbb{C}$ be n roots of the equation $\theta^r - \theta^{r-1} - 1 = 0$, then for $n \geq 0$

$$x_n = \sum_{i=1}^r \theta_i^r \frac{\theta_i^r}{\theta_i^r + (r-1)}$$

Proof First assume there are r geometric sequences $\{\alpha_{1n}\}, \{\alpha_{2n}\}, \dots, \{\alpha_{rn}\}$, where, for $1 \leq k \leq r$, $\alpha_{k0} = 1$ and the common ratio of $\{\alpha_{kn}\}$ is θ_k . We find that $\alpha_{kn} = \alpha_{k(n-1)} + \alpha_{k(n-r)}$. Define sequence $\{b_n\}$ by

$$b_n = \sum_{i=1}^r \alpha_{in}$$

where $\alpha_1, \alpha_2, \dots, \alpha_r$ are constants. Easy to observe that $b_n = b_{n-1} + b_{n-r}$. Thus, by taking appropriate α -sequences and α , we can ensure $b_n = x_n$ always holds.

Plugging in $n = 0, 1, \dots, n-1$ gives

$$\left\{ \begin{array}{l} \alpha_1 + \alpha_2 + \dots + \alpha_r = 1 \\ \alpha_1\theta_1 + \alpha_2\theta_2 + \dots + \alpha_r\theta_r = 1 \\ \alpha_1\theta_1^2 + \alpha_2\theta_2^2 + \dots + \alpha_r\theta_r^2 = 1 \\ \dots \dots \\ \alpha_1\theta_1^{r-1} + \alpha_2\theta_2^{r-1} + \dots + \alpha_r\theta_r^{r-1} = 1 \end{array} \right.$$

All we need to do is to solve for α . Without loss of generality, we merely solve α_1 . For all equations except the first one, subtract by the previous one multiplied by θ_r :

$$\left\{ \begin{array}{l} \alpha_1(\theta_1 - \theta_r) + \alpha_2(\theta_2 - \theta_r) + \dots + \alpha_{r-1}(\theta_{r-1} - \theta_r) = 1 - \theta_r \\ \alpha_1\theta_1(\theta_1 - \theta_r) + \alpha_2\theta_2(\theta_2 - \theta_r) + \dots + \alpha_{r-1}\theta_{r-1}(\theta_{r-1} - \theta_r) = 1 - \theta_r \\ \alpha_1\theta_1^2(\theta_1 - \theta_r) + \alpha_2\theta_2^2(\theta_2 - \theta_r) + \dots + \alpha_{r-1}\theta_{r-1}^2(\theta_{r-1} - \theta_r) = 1 - \theta_r \\ \dots \dots \dots \\ \alpha_1\theta_1^{r-2}(\theta_1 - \theta_r) + \alpha_2\theta_2^{r-2}(\theta_2 - \theta_r) + \dots + \alpha_{r-1}\theta_{r-1}^{r-2}(\theta_{r-1} - \theta_r) = 1 - \theta_r \end{array} \right.$$

Again, for all equations except the first one, subtract by the previous one multiplied by θ_{r-1} :

$$\left\{ \begin{array}{l} \alpha_1(\theta_1 - \theta_{r-1})(\theta_1 - \theta_r) + \alpha_2(\theta_2 - \theta_{r-1})(\theta_2 - \theta_r) + \dots + \alpha_{r-2}(\theta_{r-2} - \theta_{r-1})(\theta_{r-2} - \theta_r) = (1 - \theta_{r-1})(1 - \theta_r) \\ \alpha_1\theta_1(\theta_1 - \theta_{r-1})(\theta_1 - \theta_r) + \dots + \alpha_{r-2}\theta_{r-2}(\theta_{r-2} - \theta_{r-1})(\theta_{r-2} - \theta_r) = (1 - \theta_{r-1})(1 - \theta_r) \\ \alpha_1\theta_1^2(\theta_1 - \theta_{r-1})(\theta_1 - \theta_r) + \dots + \alpha_{r-2}\theta_{r-2}^2(\theta_{r-2} - \theta_{r-1})(\theta_{r-2} - \theta_r) = (1 - \theta_{r-1})(1 - \theta_r) \\ \dots \dots \dots \\ \alpha_1\theta_1^{r-3}(\theta_1 - \theta_{r-1})(\theta_1 - \theta_r) + \dots + \alpha_{r-2}\theta_{r-2}^{r-3}(\theta_{r-2} - \theta_{r-1})(\theta_{r-2} - \theta_r) = (1 - \theta_{r-1})(1 - \theta_r) \end{array} \right.$$

Repeat the procedure until one equation is left. Then we have

$$\alpha_1 = \frac{(\theta_1 - \theta_2)(\theta_1 - \theta_3) \dots (\theta_1 - \theta_r)\alpha_1}{(\theta_1 - \theta_2)(\theta_1 - \theta_3) \dots (\theta_1 - \theta_r)} = \frac{(1 - \theta_2)(1 - \theta_3) \dots (1 - \theta_r)}{(1 - \theta_1)(\theta_1 - \theta_2)(\theta_1 - \theta_3) \dots (\theta_1 - \theta_r)}$$

Define $f(\theta) = \theta^r - \theta^{r-1} - 1$ and $g(\theta) = (\theta - \theta_2)(\theta - \theta_3) \dots (\theta - \theta_r)$. By the definition of θ_k we know $f(\theta) = (\theta - \theta_1)(\theta - \theta_2) \dots (\theta - \theta_r) = (\theta - \theta_1)g(\theta)$.

The numerator of α_1 is $f(1) = -1$ and the denominator is $(1 - \theta_1)g(\theta_1)$. For the denominator, consider the derivative of $f(\theta)$. On one hand

$$f'(\theta) = r\theta^{r-1} - (r - 1)\theta^{r-2}$$

On the other hand

$$f'(\theta) = g(\theta) + (\theta - \theta_1)g'(\theta)$$

$$\text{So } r\theta_1^{r-1} - (r - 1)\theta_1^{r-2} = f'(\theta_1) = g(\theta_1) + (\theta_1 - \theta_1)g'(\theta) = g(\theta_1).$$

Moreover, since $\theta_1^r - \theta_1^{r-1} = 1$, we have

$$\begin{aligned} (1 - \theta_1)g(\theta_1) &= (1 - \theta_1)(r\theta_1^{r-1} - (r - 1)\theta_1^{r-2}) \\ &= -r(\theta_1^r - \theta_1^{r-1}) + (r - 1)(\theta_1^{r-1} - \theta_1^{r-2}) \\ &= -r(\theta_1^r - \theta_1^{r-1}) + \frac{(r - 1)}{\theta_1}(\theta_1^r - \theta_1^{r-1}) \\ &= -r + \frac{(r - 1)}{\theta_1} \end{aligned}$$

Thus

$$\begin{aligned}\alpha_1 &= \frac{-1}{-r + \frac{(r-1)}{\theta_1}} = \frac{\theta_1}{r\theta_1 - (r-1)} = \frac{\theta_1^r}{r\theta_1^r - (r-1)\theta_1^{r-1}} \\ &= \frac{\theta_1^r}{\theta_1^r + (r-1)(\theta_1^r - \theta_1^{r-1})} = \frac{\theta_1^r}{\theta_1^r + (r-1)}\end{aligned}$$

Applying the procedure above to other α proves the entire identity.

4.2 Formula for the y-sequence

The idea from 4.1 apply. We still assume $y_n = \beta_1\theta_1^n + \beta_2\theta_2^n \dots + \beta_r\theta_r^n$. Plugging the initial conditions gives

$$\begin{cases} \beta_1 + \beta_2 + \dots + \beta_r = r \\ \beta_1\theta_1 + \beta_2\theta_2 + \dots + \beta_r\theta_r = 1 \\ \beta_1\theta_1^2 + \beta_2\theta_2^2 + \dots + \beta_r\theta_r^2 = 1 \\ \dots \dots \\ \beta_1\theta_1^{r-1} + \beta_2\theta_2^{r-1} + \dots + \beta_r\theta_r^{r-1} = 1 \end{cases}$$

We may solve the equations like what we did for α , but here we have a simpler way.

First introduce a theorem.

Theorem 2(Newton's Identity) Assume $\theta_1, \theta_2, \dots, \theta_n \in \mathbb{C}$ are n roots of polynomial $f(\theta) = a_0\theta^n + a_1\theta^{n-1} + \dots + a_{n-1}\theta + a_n$, define $s_k = \sum_{i=1}^n \theta_i^k$, then

$$\begin{aligned}a_0s_k + a_1s_{k-1} + \dots + a_{k-1}s_1 + ka_k &= 0 \quad (k \leq r) \\ a_0s_k + a_1s_{k-1} + \dots + a_{k-n}s_{k-n} &= 0 \quad (k > r)\end{aligned}$$

In particular, when $f(\theta) = \theta^r - \theta^{r-1} - 1$, we know $a_0 = 1, a_1 = a_r = -1$ and $a_k = 0$ ($1 < k < r$).

Let $k=1$, we have $a_0s_1 + 1 \cdot a_1 = 0$, thus $s_1 = 1$.

Let $k=2$, we have $a_0s_2 + a_1s_1 + 2a_2 = 0$, namely $s_2 - s_1 = 0$ or $s_2 = 1$.

Let $k=3, 4, \dots, r-1$, we have $s_1 = s_2 = \dots = s_{r-1} = 1$.

Compare the results with the original equations, we discover that $\beta_1 = \beta_2 = \dots = \beta_r = 1$ is a solution. This derives the following elegant identity:

Identity 17. Let $\theta_1, \theta_2, \dots, \theta_r \in \mathbb{C}$ be n roots of the equation $\theta^r - \theta^{r-1} - 1 = 0$, then for $n \geq 0$

$$y_n = \sum_{i=1}^r \theta_i^n$$

5. Conjectures

In this section we present some conjectures discovered when deriving the formulas.

5.1 The Common Denominator

When reducing the fractions in 4.1 to a common denominator, we discovered that

$$\begin{aligned}
 (\theta_1^2 + 1)(\theta_2^2 + 1) &= 5 = 2^2 + 1^1 \quad (r = 2) \\
 (\theta_1^3 + 2)(\theta_2^3 + 2)(\theta_3^3 + 2) &= 31 = 3^3 + 2^2 \quad (r = 3) \\
 (\theta_1^4 + 3)(\theta_2^4 + 3)(\theta_3^4 + 3)(\theta_4^4 + 3) &= 283 = 4^4 + 3^3 \quad (r = 4) \\
 (\theta_1^5 + 4)(\theta_2^5 + 4)(\theta_3^5 + 4)(\theta_4^5 + 4)(\theta_5^5 + 4) &= 3381 = 5^5 + 4^4 \quad (r = 5)
 \end{aligned}$$

This pattern suggests the following conjecture:

Conjecture 1. Let $\theta_1, \theta_2, \dots, \theta_r \in \mathbb{C}$ be n roots of the equation $\theta^r - \theta^{r-1} - 1 = 0$, then

$$\prod_{i=1}^r (\theta_i^r + (r - 1)) = r^r + (r - 1)^{(r-1)}$$

Computer program (affiliated in the appendices) verifies the conjecture for $k \leq 14$. When $k > 14$, the mechanical error becomes significant and not negligible.

References

- [1] T. Benjamin and Jennifer J. Quinn, *Proofs that Really Count: The Art of Combinatorial Proof*, The Mathematical Association of America, 2003
- [2] 段学复, 数学小丛书—对称 (11-12), 科学出版社 (Science Press), 2002

Appendices

Program for Conjecture 1 (Matlab)

```

syms x;
r=12;
root=solve(x^r-x^(r-1)-1==0);
prod=1;
for i=1:r
    prod=prod*(root(i,1)^r+r-1);
end
double(prod-r^r-(r-1)^(r-1))
    
```



Social Networks in Entrepreneurial Opportunity Recognition Through Different Methods of Communication

Yuxin Wu

Author Background: Yuxin Wu grew up in China and currently attends Shanghai Pinghe Bilingual School in Shanghai, China. Her Pioneer seminar topic was in the field of business and titled “Examining the Factors of High-Growth High-Performance Entrepreneurship: Firms, Founders, and Ecosystems.”

Abstract

Different social networks can have influences on entrepreneurial opportunity recognition, and this is a highly concentrated part of entrepreneurial studies. This paper focuses on four social networks — family members, friends, acquaintances, third parties — and their influences on opportunity recognition. Difference of these influences through various channels of communications are also examined, especially online and face-to-face communication. This research shows that all these four types of social networks do influence on opportunity recognition, and generally positively. For different generations, different types of communication might play different role for social networks to influence on opportunity recognition, but both online and face-to-face communication are crucial.

Introduction

In contemporary society, entrepreneurs have emerged from every part of the world. Opportunity recognition has been acknowledged as an essential step in the entrepreneurial process. Entrepreneurial opportunity recognition means identifying opportunities to introduce new products or services, organize new events, start new businesses and so on.

The identification of opportunity depends on the social environment, as understanding the surrounding environment is the key to understanding demands, information, and opinions. The strongest links between entrepreneurs and the social environment are entrepreneurs' social networks. With the establishment of social networks, social capital may help entrepreneurs recognize new opportunities. Therefore, the importance of social networks in the opportunity recognition process is obvious.

However, there are various types of social networks, including strong ties with frequent interactions and weak ties with fewer interactions. Different types of social networks may have different degrees of influence on entrepreneurial opportunity recognition, but all of them provide valuable information or knowledge. To better understand how social networks influence entrepreneurial opportunity recognition, this research study will use questionnaires and deep interviews with entrepreneurs to gather their perspectives and experiences on the topic.

Besides, since online social media has developed rapidly in the modern era, it has also become a major platform for communication between entrepreneurs and their social networks. Consequently, this research also focuses on how entrepreneurs' social networks influence opportunity recognition with different kinds of communication methods. Moreover, as entrepreneurs of different ages could be influenced differently through diverse

kinds of communications in their social networks, this research also examines the difference in perceptions of post-millennials and generations X&Y.

Literature Review

Opportunity recognition has always been an essential process during an entrepreneurial startup. However, it is hard to define because it has different meanings to different people in the business world, and researchers have not come to consensus on whether opportunities emerge from the outside environment or inside entrepreneurs' minds (Miao, 2006). Consequently, there are various definitions of opportunity recognition. For instance, opportunity recognition has been defined as "perceiving an opportunity and creating an organization to pursue it" (Bygrave and Hofer, 1991). It is also supported by researchers that "opportunity recognition refers to the extent to which possibilities for new ventures exist and the extent to which entrepreneurs have the leeway to influence their odds for success through their own actions" (Lumpkin, 1999). Opportunities emerge from a complex pattern of changing conditions, which refer to changing technology, economic, political, social and demographic conditions. Nevertheless, most of the definitions have one common feature: opportunity recognition is the process of finding a new idea on which to found a new business or a new venture.

From the majority of past research, researchers nearly all agree on one other point besides the definition: opportunity recognition is crucial for any entrepreneurial processes from diverse standpoints. For instance, some of the research points out the importance of recognizing opportunities and various factors that may influence the process (e.g., Gaglio and Katz, 2001). In fact, opportunity recognition actually affects the entire domain of entrepreneurship (Taktak and Triki, 2015). Other research focus on the determinant forces of opportunity recognition in the final performance of the venture and how opportunities emerge in this ever-changing environment with different personalities and traits of particular managers (Shane, 2003). Actually, because entrepreneurs always create wealth not only for themselves but also for their societies, enhancing entrepreneurs' opportunity recognition has the potential to exert important social benefits on the whole world. Even for existing companies, opportunity recognition is still one of the vital processes for a business. For example, the recognition of opportunities may improve existing offerings or develop novel products through accomplishing first mover advantage (Gianiodis and Lisa, 2010). Nearly all research is in consensus that opportunity recognition is essential.

There is another common topic mentioned in the majority of the literatures besides the importance of recognizing opportunities in entrepreneurial startup: how important information is for opportunity recognition. For instance, researchers have concluded that better information is one of the reasons some individuals are able to discover opportunities before others (Shane, 2003). Similarly, information gathering is one of the cognitive activities that an opportunity seeker must undergo (De-koning, 1999). Two of the major factors that could have influences on opportunity recognition are information gathering and asymmetry (Alexander, Richard, Sourav, 2000). Access to priority information facilitates decision making in a context of uncertainty and pressure, and can be used to recognize the opportunity (Taktak and Triki 2015). Therefore, information gathering is the key to opportunity recognition.

One of the most convenient and effective methods to get useful information is through social networks. A social network is a network of social interactions and personal relationships, which is notably important in opportunity recognition. Connections between these networks are capable of providing access to resources, data, knowledge and investors for entrepreneurs. Most importantly, findings suggest that other persons often provide

entrepreneurs with information which is extremely useful in seeking viable new ventures during an opportunity recognition process (Ozgen and Baron, 2005). Also, as opportunities always emerge from a complex pattern of changing conditions, which refer to changing technological, economic, political, social and demographic conditions, the usage of social networks will help entrepreneurs to adapt the environment and identify the opportunities in the environment (Baron, 2004). In conclusion, social networking has become a necessary and important means of recognizing opportunities (Fatemeh and Ali, 2014). For example, research suggests that weak ties are useful recourses in finding opportunities, and they are able to reach a lot of people in the networks (Granovetter, 1973). Strong ties also play an essential role that cannot be ignored: they are able to provide entrepreneurs with accurate, reliable, low cost and proper resources, including information and communications between contacts (Jack, Anderson, 2004). Actually, one of the main methods of gaining better information is through social networks, which contributes to discovering an opportunity earlier than others (Shane, 2003). Therefore, social networks provide invaluable access to information for opportunity recognition.

This conclusion is not unprecedented, and it has abundant past research focusing on linking social networks to opportunity recognition. Evidence shows that people who are in social networks have large impacts on the process of identifying opportunities (Fatemeh, Ali, 2014). For instance, there is research which links network ties to entrepreneurial opportunity discovery by cognitive trust and examines how family members, informal networks, and mentors influence entrepreneurial opportunity recognition as sources of information (Ren and Shu and Bao and Chen, 2014). There is also research conducted in particular places, such as Nigeria, which suggests that social networks exert significant independent influence on Nigerians' entrepreneurial opportunity recognition (John, Salawu, Odunayo, 2014). Moreover, the impacts of weak ties are also examined in several research studies. Current research primarily focuses on three social sources of information that are relevant with opportunity recognition, including mentors, informal industry networks and participation in professional forums (conventions, conferences, seminars, workshops, etc.) (Ozgen and Barron, 2005).

To be more specific, research done on weak ties has already shown some convincing results. For instance, there are results which indicate that all the connections with mentors, informal industry networkers and participation in professional forums are capable of exerting direct and positive effects on opportunity recognition. Furthermore, the greater the extent to which entrepreneurs have mentors, informal industry networks, and participation in professional forums, the greater their alertness to new business opportunities (Ozgen and Baron, 2005). In addition to research on weak ties, there are also persuasive results found based on strong ties present in social networks. Results showed that strong ties between family members, close friends and entrepreneurs themselves will have a positive effect on entrepreneurial opportunity discovery and entrepreneurial opportunity exploitation due to affective trust (Ren and Shu and Bao and Chen, 2014).

In more recent years, social media has become a major communication media and technology in social networking. There is research which suggests that when firms use social networking technology to conduct several boundary-spanning activities, the likelihood of identifying new opportunities will be increased (Peter and Lisa, 2010). There is research which examines the difference of online communication and fact-to-face communication, which indicates that some internet-users will spend less time with their family members and friends (UCLA Center for Communication Policy, 2000). This would lead to the suggestion that poorer quality, weak tie, internet social relationships may be substituted for better relationships, or that time spent online could otherwise be spent

forming strong ties through face-to-face relationships. (Kraut et al., 1998) Obviously, through online communication or face-to-face communication, different effects on social networks are formed; however, little research concentrates on the differences between face-to-face communication and online communication as channels for opportunity recognition. Therefore, this research will not only examine the influences of social networks in opportunity recognition, but also pay attention to the use of social media and online communication, and how it plays a role in opportunity recognition.

Research Question

As social media is becoming increasingly important in contemporary life, it also contributes significantly to communication within social networks. Therefore, this research mainly focuses on how social networks affect entrepreneurial opportunity recognition through different approaches to communication. Besides, because the usage of social media is extremely popular in the millennial generation compared with relatively older generations, this research also takes into account the differences between generational perceptions of the impact that social media and social networks have on entrepreneurial opportunity recognition. The main generations studied in this research will include post-millennials, which generally refers to people who are born 1995 – 2012; Generation Y, which refers to people born from 1977-1994; Generation X, which refers to people born 1966 – 1976. This research examines both strong ties and weak ties in social networks. To be more specific, it examines four main parts of social networks, including family members, close friends, acquaintances who are not considered friends, and third parties, who stand for bankers, investors, entrepreneurs or colleagues in the same industry.

General Hypothesis: Social networks influence entrepreneurial opportunity recognition.

As I believe that social networks are able to provide crucial information and resources for entrepreneurs, social networks are capable of exerting impacts on opportunity recognition. Besides, both strong ties, with more interactions and higher cognitive trust, and weak ties, with higher diversity and more people in the networks, will influence opportunity recognition. Specifically, strong ties include family members – parents, spouses, siblings, children, and close friends — persons whom one knows and with whom one has a bond of mutual affection. Weak ties include acquaintances — persons known but not considered friends, and third parties — investors, mentors, bankers, entrepreneurs in the same industry, consumers, employees, and thought leaders.

Hypothesis 1: Family members influence on the opportunity recognition process, mainly through face to face communication.

<i>Hypothesis 1: Influences of Family Members in Opportunity Recognition</i>	
a) Face-to-Face Communication	Both post-millennials and generations X&Ys' opportunity recognition process may be largely influenced by family members through face to face communication.

<i>Hypothesis 1: Influences of Family Members in Opportunity Recognition</i>	
b) Online Communication	For both post-millennials and generations X&Y, online communication may be an auxiliary method of communication in family members' impacts.

As I believe that entrepreneurs will have a lot of communication with their family members, which facilitates information flow, family members may be able to provide valuable information for entrepreneurs. This information might help entrepreneurs to identify possible opportunities. Also, since family members are likely to be the people whom entrepreneurs trust most, this degree of trust ensures the richness of exchanging relationships and the exchanging of information between each other. By having access to more persons, resources from the relationships and information gathered from family members, entrepreneurs' opportunity recognition process is improved. Furthermore, since family members will have a lot of face-to-face communication due to the same geographical locations, face-to-face communication may be more important and useful for entrepreneurs in this context. Both post-millennials and generation X & Y are likely to consider face-to-face communication with family members helpful in identifying opportunities. Online communication may be an auxiliary way for family members to convey information and resources for entrepreneurs when they are in different geographical locations.

Hypothesis 2. Close friends have an influence on entrepreneurs' opportunity recognition process, mainly through face-to-face communication.

<i>Hypothesis 2: Influences of Close Friends in Opportunity Recognition</i>	
a) Face-to-face communication	For generations X&Y, face-to-face communication will be a major source of information from close friends which will help them to identify opportunities.
b) Online communication	For post-millennials, online communication will bring dominant sources of information from close friends which is essential for opportunity recognition.

Connection with close friends is also a type of strong tie in social networks. Therefore, the information transfer and relationship transfer are likely to happen between entrepreneurs and their close friends. This is mainly because the frequent interactions which take place between entrepreneurs and their friends is likely to produce a high degree of trust. Besides, as entrepreneurs of elder generations may spend a large portion of their time with their friends, especially friends who are in the same geographical location, face-to-face communication is a more useful and effective way to connect. However, post-millennials may prefer to contact their friends through online communication due to the convenience and diversity of social media. In conclusion, close friends are capable of influencing entrepreneurs when they are identifying opportunities; for generations X&Y, the influences

mainly come from face-to-face communication, while for post-millennials, the influences mainly come from online communication.

Hypothesis 3: Acquaintances will have influence on entrepreneurs' opportunity recognition, mainly through online communication.

<i>Hypothesis 3: Influences of Acquaintances in Opportunity Recognition</i>	
a) Face-to-face communication	Both for post-millennials and generations X&Y, face-to-face communication might be an effective way to gather information from acquaintances, which contributes to opportunity recognition.
b) Online communication	Both for post-millennials and generations X&Y, online communication could be as effective as face-to-face communication to provide information for entrepreneurs from acquaintances.

Acquaintances refer to people whom entrepreneurs know, but without friendships or deep relationships between each other. Although acquaintances may not be considered friends, they are still able to provide diverse information and relationships for entrepreneurs due to the frequent interactions between them. Both post-millennials and generations X&Y may spend a lot of time with acquaintances in their daily life, so face-to-face communication influences entrepreneurial opportunity recognition. However, due to the limited degree of trust, both post-millennials and generations X&Y may also prefer to use online technology to contact their acquaintances. Therefore, for acquaintances, both face-to-face communication and online communication could be effective ways to gather information for entrepreneurs exploring new opportunities, whether for post-millennials or elder generations.

Hypothesis 4: Third parties have influence on the opportunity recognition process of entrepreneurs, mainly through online communication.

<i>Hypothesis 4: Influences of Third Parties in Opportunity Recognition</i>	
a) Face-to-face communication	For post-millennials, face-to-face communication with third parties may help them to understand and explore opportunities more easily.
b) Online communication	For generation X&Y, online communication with third parties may bring a large amount of information and resources, which contributes significantly to their opportunity recognition process.

Third parties may include bankers, investors, entrepreneurs or colleagues in the same industry, employees, and consumers; these are weaker ties in social networks. Due to the possible resources that are able to be provided from connections with third parties, such as financial aid, recommendation from employees, and valuable information from entrepreneurs or colleagues in the same industry, third parties are able to exert impacts on the opportunity recognition process of entrepreneurs in diverse ways. Besides, for generation X&Y, since third parties may be easier to approach through online communication compared with face-to-face communication due to geographical separation and restriction, online communication may play an essential role in the transfer of information and resources between third parties and entrepreneurs. Nevertheless, post-millennials may have limited access to online contact with third parties. Therefore, face-to-face communication might be more accessible and important for post-millennials. In conclusion, third parties will influence entrepreneurial opportunity recognition: for generation X&Y, online communication is more essential; for post-millennials, third parties' influences mainly come from face-to-face communication.

Sampling and Data Collection

<i>Table 1. Results for Question 1 in questionnaire, presented in frequency and percentage.</i>		
If an entrepreneur is someone who initiates the pursuit of new opportunities without regarding personal risks, and is willing to give without expecting return, even anticipating failures, would you consider yourself an entrepreneur?		
Yes	49	49%
No	28	28%
Not sure	23	23%

This research adopts primary research as the major sources of results, including questionnaires and interviews. An online questionnaire with 16 questions is published through www.sojump.com in Chinese, in case there are respondents that are not able to read English. In August 2017, I sent the link of my questionnaires to my 518 contacts, including the majority of my peers, some members of generation X and Y, and several entrepreneurs in my social media (Wechat). According to Table 1, 49% of respondents consider themselves entrepreneurs, so this would be a proxy research based on this sample. To increase the rate of response, I sent the link five times in five following days. After one week of data collection, I received 100 responses in total. The response rate was approximately 19.31%. Since the sample of the questionnaire was quite limited to post-millennials, I conducted six individual interviews in total to gather information from entrepreneurs. Specifically, these six interviews included three interviews with post-millennial entrepreneurs who have started their own businesses, clubs, or events; two interviews with entrepreneurs from generation Y, and one interview with an entrepreneur from generation X. Clearly, the sample group is mainly my peers and mature entrepreneurs from generation X&Y.

Sources of Bias and Validity Threats

1. The majority of the participants who completed the surveys were post-millennials who were only able to represent the opinions of younger generations with entrepreneurial mindsets. It was difficult to draw conclusions about the opinions of entrepreneurs through the statistics of the questionnaire. However, interviewing mature entrepreneurs provided insightful opinions. Also, some of the younger generations already had an entrepreneurial orientation, which might represent the more established entrepreneurs.

2. The majority of the participants were in China, and only a few of them were from foreign countries. Therefore, the database was mainly based on China's environment, using online communication through specific technologies in China, which might differ from other countries' situations.

3. Recall Bias was caused by differences in the accuracy or completeness of recollections retrieved by study participants regarding events or experiences from the past. In my questionnaires and interviews, recall bias might arise, since I asked a question about "When was your last opportunity recognition? Who did you contact?" at the beginning of my surveys. This question might influence the answers in the following questions. However, this bias could be eliminated to some degree by the large number of the results and responses to the questionnaires and interviews. Therefore, it was acceptable in this research.

Findings and Analysis

There is a large number of findings both from the questionnaires and interviews, which are presented and analyzed in the following content.

Table 2. Percentage assigned by respondents to the degree of influences that each category of relationships that bring on opportunity recognition.

Type of Relationships	Percentage assigned for face to face communication	Percentage assigned for online communication
Family members	32.44%	24.92%
Close friends	32.2%	35.89%
Acquaintances	14.6%	16.42%
Third parties	14.82%	15.65%
Others	6.08%	7.12%

For Table 2, the first column shows different types of relationships. Vertical comparison will show the difference of different social networks through the same communication method, i.e., face-to-face communication and online communication. Horizontal comparison will show the difference between different methods of communication for the same type of social networks.

Table 3. Responses by Frequency and Percentage, The Degree of Influence of Relationship to Opportunity Recognition (F-F — face-to-face communication. Online — online communication.)

Relation-ship			Strongly Disagree		Disagree		Neither Disagree nor Agree		Agree		Strongly Agree		N
H 1	Family Member	a. F-F	4	4%	15	15%	20	20%	53	53%	8	8%	100
		b. Online	7	7%	19	19%	24	24%	45	45%	5	5%	100
H 2	Close Friends	a. F-F	3	3%	13	13%	21	21%	54	54%	9	9%	100
		b. Online	2	2%	9	9%	18	18%	60	60%	11	11%	100
H 3	Acquaintances	a. F-F	1	1%	24	24%	20	20%	47	47%	8	8%	100
		b. Online	2	2%	20	20%	21	21%	48	48%	9	9%	100
H 4	Third Parties	a. F-F	2	2%	16	16%	23	23%	51	51%	8	8%	100
		b. Online	1	1%	21	21%	22	22%	46	46%	10	10%	100

For Table 3, vertically, there are two columns under the five options available for respondents. The first column shows the number of respondents who choose this option, and the second column shows the percentage of respondents who choose this option as the total number of respondents. The last column N shows the total number of respondents, i.e., 100.

Analysis Method

To draw general findings from questionnaire, there are mainly several methods used to analysis. First, the overall relationship between entrepreneurial opportunity recognition and social networks would be analyzed from table two by referring to the percentage who choose “agree” or “strongly agree”. Additionally, for each network’s influences, data would be analyzed through vertical comparisons with other networks’ by comparing the number and percentage of respondents who choose “agree” or “strongly agree”. To examine the effectiveness of online communication and face-to-face communication as the medium for social networks to influence on opportunity recognition, comparison inside each category is necessary. Specifically, this will need to compare the number of respondents who choose for “agree” or “strongly agree” for face-to-face communication and online communication for each social network.

General Findings from Questionnaires

1. Influences of all four social networks

In this study, there are four types of social networks examined. From the results of the questionnaires, it is obvious that all four types of social networks, including family members, close friends, acquaintances and third parties, will influence entrepreneurial opportunity recognition. However, the degree of influence varies with different types of social networks, which can be concluded from the percentage that is assigned by respondents to each of the relationships. To be more specific, according to Table 3, friends always have the largest influence with a relatively higher percentage, and family members are close behind. Third parties have the third largest influences, and finally acquaintances may have the least influences in opportunity recognition among these four types of

relationships. However, it is true that all these four social networks have influence on entrepreneurial opportunity recognition, despite the different degree.

2. Face-to-face communication as a channel for social networks to have impacts on opportunity recognition, and comparison between these four social networks.

Face-to-face communication plays a crucial role for entrepreneurs to gather information from their social networks, from family members and friends to acquaintances and third parties. However, according to Table 1, in these four types of relationships, family members have the largest influence in face-to-face communication with a percentage of 32.44%. Close friends are merely 0.44% less than family members, which also indicates strong influence in face-to-face communication. This could be explained, as close friends and family members are most likely the people who entrepreneurs and respondents spend most time with, and they probably live in the same geographical locations. Then, 59% of respondents believe that third parties influence opportunity recognition process through face-to-face communication, while 55% of respondents believe that acquaintances influence the process through face-to-face communication.

3. Online communication as a channel for social networks to have impact on opportunity recognition, and comparison between all four networks.

Although social media has only developed in recent years, it already has a large impact on opportunity recognition by providing the platform for entrepreneurs to communicate with their social networks more conveniently. According to Table 3, close friends have the highest percentage (35.89%) in all four categories of relationships. The second highest relationship is family members, but with a percentage difference of 10.97% compared with close friends. The third is acquaintances, with a percentage of 16.42%, and lastly third parties with 15.65%. It is obvious, then, that through online communication, close friends have a dominant effect on opportunity recognition, and strong ties are more influential than weak ties.

Findings for Specific Relationships from Questionnaires

1. Family Members

Family members account for a large portion of influence in opportunity recognition. Through face-to-face communication, family members are able to have significant influence on opportunity recognition. This is because entrepreneurs likely have the most frequent face-to-face interactions and communications with their family members. Through these communications, valuable information is gathered. As for family members having the least influence through online communication, this might be because they do not often use social media to contact each other, which decreases the degree of communication online.

2. Close Friends

Close friends have the largest influence on entrepreneurial opportunity recognition both in face-to-face communication and online communication. This strong tie may be the most influential on entrepreneurial opportunity recognition due to the frequent interactions and information exchange. However, online communication is 8% higher than face-to-face communication, which shows the significant role of online communication.

3. Acquaintances

The influence that acquaintances bring would be quite moderate compared with strong ties. 55% of respondents suggest that acquaintances affect their opportunity recognition through face to face communication, while 57% of respondents suggest that the effects of acquaintances come from online communication. Therefore, online communication is slightly higher than face-to-face communication.

4. Third Parties

The influence that third parties have is slightly greater than that of acquaintances, but smaller than that of family members and close friends. Notably, third parties' influence through face-to-face communication is actually more significant than online communication, with a percentage difference of 3%. Therefore, face-to-face communication might be more important than online communication for third parties to affect opportunity recognition.

5. Questionnaires

From questionnaires, only statistics can be seen without explanation. Therefore, six deep interviews are conducted to further explore entrepreneurs' opinions with more details and possible reasons for both post-millennials and older generations.

Findings from Interviews with Post Millennials

From interviews with post millennials, there are several key points summarized from the results.

1. Social networks have influence on opportunity recognition.

2. For post-millennials to start their own business, family members are crucial for providing finance backup, and they may introduce some possible suppliers and customers for millennial entrepreneurs. The majority of the time, this information comes from face-to-face communication. Also, post-millennials mentioned that family members affect their opportunity recognition through another perspective: they provide education and business background. According to interviewees, both the education they received and their family helped them to identify business opportunities with more knowledge and experience from family.

3. Friends also ignite some post-millennial entrepreneurs to find new opportunities. According to post-millennials, since their friends have interesting ideas and information to share, they provide creative information and thoughts, which help post millennials to identify new opportunities. Also, post-millennials prefer to contact their friends online, because it is more convenient and informal, where they are able to talk various topics. Additionally, it is easier to get along with their friends. Post-millennials mentioned that online communication may enlarge their friend networks and bring diverse opportunities for them.

4. Post-millennials believe that acquaintances do have an effect on opportunity recognition, but since they do not trust acquaintances, the degree of influence is quite limited. Also, there may be information that they are not willing to share or exchange with acquaintances, which further restricts the degree of influence. They prefer to use online communication because they say that face-to-face communication with acquaintances can sometimes be awkward and embarrassing.

5. Third parties definitely have an influence on opportunity recognition by introducing possible events for millennials. However, information mainly comes from face-to-face communication. As post-millennials explain, face-to-face communication is more formal, and more appropriate for them to communicate with third parties like possible sponsors, teachers and so on.

6. Post-millennials generally consider online communication to be more important than face-to-face communication, because they actually know a lot of their friends online. Moreover, with social media, they are able to transfer and exchange information with even more people in different geographical locations. Also, online communication can be more relaxing for them.

Findings from Interviews with Generation X and Y

1. Generations X&Y also consider social networks to exert huge influence on opportunity recognition.

2. Family members provide information for these mature entrepreneurs. Also, some entrepreneurs mention that their family members even teach them how to identify possible resources and opportunities. Therefore, education provided by family is invaluable, and it is mainly from face-to-face communication.

3. Friends also contribute significantly when entrepreneurs are identifying opportunities. Some of the entrepreneurs mention that friends are actually their primary customers when they begin their businesses, so their friends have a lot of useful feedback for them to identify new opportunities. Also, friends are able to introduce some cooperators for entrepreneurs through face-to-face communication.

4. Acquaintances are also possible customers for entrepreneurs, so their opinions help entrepreneurs to improve the products and come up with new ideas about possible products. Entrepreneurs prefer to communicate with acquaintances through online communication because it saves time.

5. Third parties provide significant financial opportunities for entrepreneurs. They explain that third parties, like colleagues in the same industry, are possible cooperators for them. Also, some of the investors may help entrepreneurs gather information about the financial state, and help them to identify if there are any opportunities in the outside environment now. Online communication is more effective for generations X&Y to communicate with third parties. Because some of the third parties are also introduced by entrepreneurs' other ties like friends, social media becomes a crucial method for them to communicate.

6. Entrepreneurs in generations X&Y generally think that both online communication and face-to-face communication are important methods for their social networks to influence them. As entrepreneurs explain, there are more emotional interactions in face-to-face communication, which provides deeper knowledge and information transfer. Online communication helps them to contact more people who have different backgrounds.

Conclusions

Table 4. *Conclusions and Findings in Relation to Hypothesis (+) — Support Hypothesis. (-) — Contradict Hypothesis. (~) Partially Support Hypothesis.*

Conclusions and Findings in Relation to Hypotheses	
General Hypothesis	(+) Social networks will influence opportunity recognition, and all of the four types of relationships, including family members, close friends, acquaintances, and third parties have impact.
Family Members	(+) Family members have influence on opportunity recognition, mainly from face to face communication for both post-millennials and generations X&Y.
Close Friends	(~) Close Friends have influence on opportunity recognition. For post -millennials, it is mainly from online communication; for generations X&Y, it is mainly from face to face communication

Acquaintances	(+) Acquaintances have influence on opportunity recognition, mainly through online communication for both post-millennials and generations X&Y.
Third Parties	(+) Third parties have influence on opportunity recognition. For post- millennials, it is mainly from face to face communication; for generations X&Y, it is mainly from online communication

This research addresses the effects of social networks on entrepreneurial opportunity recognition. According to Table 4, nearly all findings match with the hypothesis suggested above. There are several notable conclusions from this research.

1. The persons in social networks play an important role in the process of identifying opportunities, no matter what type of social network they belong to. However, strong ties, such as close friends and family, may be more effective and influential compared with weak ties, including acquaintances and third parties. Different degrees of influence could be formed due to different degrees of trust between entrepreneurs and their strong ties or weak ties.

2. Generational impact exists in this research. Different generations sometimes have different channels to communicate with different people. For instance, post-millennials have more face-to-face communication with third parties due to the limited access of social communication technologies. Post-millennials also can present themselves better in face-to-face communication with third parties. Generations X&Y are contacting and communicating with third parties through online communication due to the diversity and the large scale of people that they are able to reach.

3. Family members have the largest influence on opportunity recognition as dominant factors in social networks. The flow of information between family members and entrepreneurs is easier due to deep and frequent face-to-face interactions and communications.

4. For friends, the findings are actually quite counter-intuitive. Post-millennials are actually contacting their friends mainly through online communication because they consider it less formal, and they are able to know more friends from the usage of online media.

5. Acquaintances are approached mainly through online communication, due to geographical separations and avoidance of embarrassment, and the influence that acquaintances bring is quite limited compared to strong ties. However, acquaintances still have a significant effect on providing information for entrepreneurs.

6. Third parties have influence on opportunity recognition, but their influence is relatively weak compared with strong ties like that of friends or family members. Their influence towards post millennials' entrepreneurial opportunity recognition is mainly face-to-face communication due to limited access for contact information. Their influence towards generations X&Ys' entrepreneurial opportunity recognition is mainly through face-to-face communication.

7. Post-millennials consider online communication to be slightly more important than face-to-face communication because it is less formal and more comfortable for them to use. Also, they are able to know more friends though online communication. Thus, when they are recognizing new opportunities, post-millennials think that for their social network to have an effect, online communication is necessary and crucial.

8. Generations X&Y consider online communication to be as important as face-to-face communication. Face-to-face communication conveys more emotion, while online communication gives entrepreneurs space and time to ponder the usage of the language when they are communicating to third parties.

Implications

Centering on social networks' impacts on entrepreneurial opportunity recognition, extant literature supports the importance of social networks on various business activities, specifically on entrepreneurial opportunity recognition. However, to my knowledge, there is very little literature focusing on the channels for social networks to influence opportunity recognition and generational impact in this process in contemporary background. Therefore, this research attempts to explore the question of through which channel, online communication or face-to-face communication, do social networks affect entrepreneurial opportunity recognition, and to what degree. Also, as a post-millennial, I personally sense the differences between generations X&Y and post-millennials, since most of the post-millennials are raised in the Internet era, while generations X&Y are not. This cultural difference also leads this research to pay attention to the generational influence on particular questions.

This research aims to enhance the existing literature on social networks. It further establishes the link between social networks and opportunity recognition, and channels for social networks to influence opportunity recognition. Analysis shows that both online communication and face-to-face communication are important platforms for social networks to influence opportunity recognition. Furthermore, there could be different amounts of reliance on these two methods of communication depending on different types of social networks, and generations.

Scholars are able to acknowledge the generational impact on entrepreneurial processes, particularly in identifying opportunities. Therefore, in future studies, there are questions that researchers could continue exploring. Because this research mainly identifies two generations' opinions towards social networks and technologies' role in opportunity recognition, will the opinion of other generations differ? To what extent do the development of the Internet and social media affect the influence of social networks on opportunity recognition? Both millennials and post-millennials are raised in the Internet environment, but do they perceive the importance of online communication for social networks to affect opportunity recognition similarly?

Entrepreneurs might seek help and advice from their social networks if they are struggling with identifying new opportunities. Besides, online communication is rapidly growing as important channels for post-millennials to use. For new companies, the adaptations of communication through online technologies or social media are essential to getting feedback from millennials and post-millennials, which could contribute to the creation of improved products and ideas. However, face-to-face interaction still remains crucial. Therefore, companies and entrepreneurs should find a balance between online and face-to-face communication when they are trying to gather information and recognize new opportunities in their business activities.

References

Alexander Ardichvili, Richard Cardozo, Sourav Ray (2000). *A Theory of Entrepreneurial Opportunity Identification and Development*. Center for Entrepreneurial Studies, Minneapolis.

- Bygrave, W. D., & Hofer, C. W. (1991). Theorizing about Entrepreneurship. *Entrepreneurship Theory and Practice*, 13-22.
- De Koning, A. (1999). Conceptualizing Opportunity Recognition as a Socio-Cognitive Process. Centre for Advanced Studies in Leadership, Stockholm.
- Eren Ozgen, Robert A., Baron (2005). Social Sources of Information in Opportunity recognition: Effects of mentors, industry networks, and professional forums. *Journal of Business Venturing*. 174-192.
- Fatemel Mohebi, Ali Rabiee (2014). Social Capital on Entrepreneurial Opportunity Recognition. *International Journal of Economics, Finance and Management*. Vol.3, No.4. 2307-2466.
- Gaglio, C., J., (2001). The Psychological Basis of Opportunity Identification: Entrepreneurial Alertness. *Small Business Economics* 16, 95-111
- Hills, G., Lumpkin, G.T. and Singh, R.P. (1997). Opportunity recognition: perceptions and behaviors of entrepreneurs. *Frontiers of Entrepreneurship Research*, Babson College, Wellesley, MA, pp. 203-218.
- Jack, S.L., Dodd, S.D. and Anderson, A.R., 2004. Social Structures and Entrepreneurial Networks: The Strength of Strong Ties.
- John Kolawole, Salawu Hassan, Odunayo Oluwasanmi. (2014). Social Network and Human Capital as Determinants of Entrepreneurial Opportunity Recognition in Nigeria. *International Journal of Public Administration and Management Research*. Volume 2, Issue 2.
- Kraut, R., M. Patterson, V. Lundmark, S. Kiesler, T. Mukhopadhyay and W. Scherlis (1998) 'Internet Paradox: a Social Technology that Reduces Social Involvement and Psychological Well-being?', *American Psychologist* 53(9): 1017-31.
- Mark S. Granovetter. 1973. The Strength of Weak Ties. *American Journal of Sociology*, Volume 78, Issue 6, 1360-1380.
- Nancy K. Baym., Yan Bing Zhang., 2004. Social Interactions Across Media. SAGE publications. Volume6(3):299-318
- Peter, Gianiodis., Lisa, Bosman. 2010. The Effect of Social Networking Technology on Firm Opportunity Recognition. Clemson University.
- Salima, Taktak., Mohamed Triki., 2015. The Importance of Behavioral Factors: How Do Overconfidence Affect Entrepreneurial Opportunity Evaluation. Sfax University, Tunisia Unity of Research GOVERNANCE.
- Shane, S., 2003. A General Theory of Entrepreneurship: The Individual-Opportunity Nexus. Edward Elgar, Cheltenham, UK.
- Sehngang, Ren., Rui, Shu., Yongchuan, Bao., Xiaohong, Chen. 2014. Linking Network Ties to Entrepreneurial Opportunity Discovery and Exploitation: The Role of Affective and Cognitive Trust. Springer Science, Business Media New York.
- UCLA Center for Communication Policy (2000). 'Surveying the Digital Future', URL: <http://ccp.ucla.edu/pages/internet-report.asp>.



Sofonisba Anguissola and Her *Self-Portrait at the Easel*

Xingzhi Jing

Author Background: Xingzhi Jing grew up in China and currently attends Hangzhou Foreign Languages School in Hangzhou, China. Her Pioneer seminar topic was in the field of art history and entitled "Methodologies of Art History."

Abstract

This research paper was on a particular painting, *Self-Portrait at the Easel (Painting a Devotional Panel)*, by the female Renaissance artist, Sofonisba Anguissola. In the image, the artist presents herself painting a devotional panel of the Virgin Mary and the Christ Child. In the paper, I interpret the work from a formal angle and compare it to self-portraits by her contemporaries (Dürer, Michelangelo, Tiziano Vecellio, Anthony van Dyck, Caterina van Hemessen and Lavinia Fontana), to show how Sofonisba used the portrait to demonstrate her religious devotion and equality with male artists of the period.

Introduction

Self-Portrait at the Easel (Painting a Devotional Panel) (Figure 1.), dated to 1556, is the work of Sofonisba Anguissola (1530/1532-1625) one of the very few influential female artists of the Renaissance. The oil painting is 26 inches × 22.4 inches in size and is currently kept in Zamek Lubomirskich i Potockich in Łańcut, Poland.

The combination of self-representation and religious iconography is evident in this painting, and may not be regarded as unconventional in Sofonisba's day. However, the depiction of a woman painting a religious scene is unusual and the intention behind her decision to paint herself as such is worth discussing. In this paper, I look at the painting from both a formal and feminist perspective by comparing it to self-portraits by other artists of the period, particularly those in which artists incorporate elements of Christianity. Hopefully, this self-portrait will allow readers to see how Sofonisba Anguissola drew on the standard theme of St. Luke painting the Virgin to create a self-portrait that deviates from those of her male and female contemporaries.

Sofonisba Anguissola: A Woman in the Renaissance

The involvement of women in the artistic world can be traced back to the medieval period when intelligent women started to receive support and protection from convents. However, it was not until the sixteenth century that the names, biographical information and works of female artists began to appear frequently (Frances Borzello, 41).

Sofonisba Anguissola was one of the pioneering professional female artists during the Renaissance period. She was the eldest daughter of Amilcare Anguissola and Bianca Punzona, both of whom came from the noble families of Cremona. Together with her younger sister, Elena, she was sent to Bernardino Campi in the year 1546 to learn painting officially after receiving a well-rounded education, and to Bernardino Gatti three years later when Campi moved to Milan (Sharlee Mullins Glenn, 296-297). The mentoring Sofonisba received from these established painters had a significant impact on her work. As she

matured, her fame spread and the demand for her portraits rose significantly. She was even recognized by Giorgio Vasari¹, who in his conclusion to the chapter about Properzia de' Rossi in the second edition of the *Lives of the Most Eminent Painters, Sculptors and Architects* of 1565 (127-128), referred to Sofonisba as someone who “has labored at the difficulties of design with greater study and better grace than any other woman of our time,” and also by Michelangelo, who examined, judged and praised her work (John Addington Symonds and Michelangelo Buonarroti).¹

As a result of her talent, effort and background, Sofonisba's reputation reached Spain in 1559 by recommendation of the Duke of Alba and impressed the court of King Phillip II. She was later employed as a court painter, lady-in-waiting, and painting instructor for the new queen Elisabeth de Valois, which demonstrates the Spanish court's appreciation of her artistic talent (Sharlee Mullins Gleen, 298). With the facilitation of the King, she married Fabrizio de Moncada, son of the Prince of Paternò, Viceroy of Sicily, in about 1570. They resided in Palermo until Fabrizio died in 1579. Two years later, while traveling to Genoa by sea, she fell in love with the ship's captain, sea merchant Orazio Lomellini, and married him shortly thereafter. They lived a comfortable life in Genoa until 1616-20, and Sofonisba died in 1625. Her final place of residence remains unclear and could have been either Genoa or Palermo (Mary D. Garrard, 618).

The year before Sofonisba died, Anthony van Dyck reportedly paid her a visit in Palermo and left some sketches of her. However, the young Flemish artist mistook her actual age of ninety-two as ninety-six, so the accuracy of the location of their meeting may also be wrong. In any case, he noted, “It was a great pleasure for her to have pictures placed in front of her . . . When I drew her portrait, she gave me several hints: not to get too high or too low so the shadows of her wrinkles would not show too much” (Betsy Fulmer, 33).

Sofonisba Anguissola was an extraordinary woman who not only took art as a career but also gained high renown and status. In the sixteenth century, when art was not considered an ideal profession for women, she was one of the very few exceptions.

Self-Portraiture in the Renaissance: A Reflection of Gender and Identity

In the Renaissance, as the public began to reject the concept that physical labor, artistic or otherwise, was socially inferior labor, much more emphasis was placed on artists' personal history and intellect. This significantly raised artists' self-awareness and gave way to the production of self-portraits – no matter whether the artists put themselves at the center or the margin of the canvas. Individual self-portraits were often presented as gifts to patrons in the hope of receiving commissions from them. Although the initial motivation behind self-portraiture could be associated with social class and status, it also granted artists a sense of freedom and flexibility as well as a unique opportunity to demonstrate their expertise and express their confidence (Thomas Kelley, 83-84).

Renaissance Women artists did not generally enjoy opportunities equal to those of their male counterparts. Although Jacob Burckhardt asserted that they “stood on a footing of perfect equality with men” (697), Kelly-Gadol and others have shown that women did not enjoy great freedom and opportunity (139). Nevertheless, there were surely at least some

¹ *The Life of Michelangelo Buonarroti: Based on Studies in the Archives of the Buonarroti Family at Florence*. Vol. 2. Philadelphia, PA, University of Pennsylvania Press, 2002. The evidence lies in the two letters sent by Amilcare Anguissola to Michelangelo on May 7th, 1557, and May 17th, 1558 to thank him for his “honorable and thoughtful affection that you have shown to Sofonisba,” and for being “kind enough to examine, judge, and praise the paintings done by my daughter Sofonisba.”

female artists who were recognized by their contemporaries for their achievements and who left a valuable legacy to the artistic world. Furthermore, according to their many self-portraits, it can be inferred that their self-awareness was just as high as that of their male counterparts.

Renaissance men and women produced self-portraits in different ways. Ambitious male artists seemed dissatisfied with painting straightforward portraits or self-portraits and more commonly devoted themselves to larger works, often with religious themes. They sometimes creatively portrayed themselves as characters in such religious scenes. For instance, in his famous *Last Judgment* (Figure 2.), Michelangelo displayed his own face on the skin of St. Bartholomew. The intention behind this arrangement remains controversial and may be associated with perishability and mortality (Marcia B Hall, 87), but the viewer can ascertain that the fresco is definitely beyond a simple self-portrait.

Tiziano was one of the very few male artists who produced multiple formal self-portraits in the 16th century. He typically depicted himself as a solemn man with a long, thick beard in a black hat and black robe, looking away from the viewer. For example, in his 1567 *Self-Portrait* (Figure 3.), he painted his profile in a priest-like way and showed himself holding a paintbrush or pencil with his right hand, revealing his occupation. His masterpiece *Allegory of Prudence* (Figure 4.), which he painted in the late 1560s, is also intriguing. Although there are some debates over the identities of the two younger characters in the painting (many scholars believe that they are Tiziano's son and nephew, while some doubt this idea), it has been agreed on the basis of his many self-portraits that the character on the left hand side is Tiziano himself. In other words, the *Allegory of Prudence* contains a hidden self-portrait of the artist. However, just as in the situation with Michelangelo, Tiziano's self-portrait has another layer of complexity, symbolizing the concept of old age and the past (Erwin Panofsky, 166).

Anthony van Dyck (1599-1641) also produced plenty of self-portraits. Unlike Tiziano's solemn self-portraits, his self-portraits are in a flamboyant aristocratic style, despite the fact that he was not born into the nobility. This is especially apparent in his *Self-Portrait with a Sunflower* (c.1633, Figure 5.). In this painting, van Dyck is in a red silk upper garment with a gold chain, which had been recently presented to him by his patron the English monarch Charles I, and his right hand is gesturing towards a large sunflower. It has been agreed that the theme of this self-portrait is van Dyck's devotion to the monarch, which is compared to the sunflower's natural inclination to the sun (Lisa Rosenthal). The painting not only raises his own fame and status but also demonstrates his loyalty to his patron.

Female artists in the Renaissance likewise found ways of representing themselves in self-portraits as well. In 1548 Caterina van Hemessen (1528-1587), the daughter and student of Flemish Renaissance artist Jan Sanders van Hemessen, produced the earliest self-portrait of an artist at work (Lilian H. Zirpolo, 246). In it (Figure 6.), she depicts herself as a respectable lady in a red and black velvet dress, holding a maulstick, a palette and paintbrushes and looking out of the painting. Her head is quite large in relation to her body, and her facial expression conveys a strong sense of modesty. On the other hand, Lavinia Fontana (1552-1614), who is said to be the first female artist to work within the same sphere as her male counterparts outside of a court or convent, painted female nudes, ran an entire workshop and supported not only her aging parents but also her husband and children (Katherine A. McIver, 3), generally depicted herself in a luxurious style as in her *Self-Portrait with Palette and Brushes* (c.1579, Figure 7.). Here she wears a delicate and expensive formal dress, with jewelry embellishing her head, the front of her chest and her

dress. The aristocratic collar and the patterns on her dress are depicted in detail, which highlights her nobility. The palette and paintbrushes remind the viewer of her profession.

Sofonisba Anguissola, the subject of this paper, seems to have produced more self-portraits than any of her female contemporaries. Taking a close look at this *Self-Portrait at the Easel*, the viewer may roughly divide it into two halves. On the right hand side is the figure of Sofonisba, who is depicted as an elegant and serious lady facing the viewer with a brownish up-do, smooth facial contours, deep-set eyes and a protruding nose. The artist is in a black dress with slits at the top of the sleeves and a white blouse with delicate lace decorating the cuffs and collar. On the left hand side is the “painting” that the artist is working on. Only Sofonisba’s upper body is painted to allow the viewer to take a closer look at the details in the “painting” as well as on her face. The “painting” depicts a woman and a child in a particularly affectionate position, perhaps whispering, or about to kiss each other. They are leaning towards one another, with the woman’s right hand holding the back of the child’s head and her left hand positioned between the two figures’ mouths and gently touching the child’s cheek. Although there are no halos, according to the red and blue garments of the woman, it can be inferred that the image depicts the Virgin Mary and the Christ Child. The artist holds a paintbrush in her right hand and a maulstick in her left, which prevents her from touching the unfinished painting and keeps her hand steady. The hands are small and unobtrusive relative to the body and face of the artist, which prevents the viewer from being distracted from the focus of the painting. The right hand of the artist creates a shadow on the bottom part of the “painting,” and on the tools next to the palette below the drawing board. The rectangular palette has four color areas on its upper part, including three black blocs and one red bloc, and a mixture of black, red, yellow and white on its lower part.

As previously mentioned, the painting uses a half-half composition, in which the artist and the “painting” each occupy half of the space. This composition puts an equal amount of emphasis on the “painting” and the artist, indicating that they are of equal importance. The artist’s hands are placed intentionally, not only perfectly fitting into the area covered by the blanket in the “painting” so that the figures can still be clearly shown to the audience, but also creating a connection between the left side and the right side of the painting. Interestingly, the “painting” inside the painting also utilizes a similar composition – though the two figures are not perfectly evenly placed on the left and right sides of the painting, the background could be divided into halves, with the right hand side as dark as the scene behind the easel.

The artist employs chiaroscuro – the contrast of brightness and darkness – in the painting. The light comes from the right front of the self-portrait and lights up the face of Sofonisba and the “painting” of the Virgin Mary and the Child. From the thick shadow created by the artist’s right hand, it can be deduced that the light from the source is relatively intense. However, the artist’s upper body does not create any shadow on the drawing board, which is unnatural and may be to ensure that the “painting” is complete and clearly presented. Additionally, the background of the self-portrait is so dark that the viewer cannot tell where the painting is being painted, which creates a sense of mystery and highlights the focus of the painting – the figures.

Contrast is also present in the figure’s movement. While Sofonisba’s facial expression conveys a sense of peacefulness, concentration, and stillness, her hands are moving actively in front of the easel. This adds complexity to Sofonisba’s persona as an elegant lady as well as a skillful artist. Furthermore, the artist is painting while she is looking at the viewer, which gives viewers the sense that the Virgin Mary and the Christ Child are actually by their side, and the artist is painting from sight rather than imagination.

The artist further demonstrates her expertise through her portrayal of texture. The highlights in Sofonisba's eyes make them glass-like, conveying a sense of translucency and girlish innocence. Her exposed skin is almost wrinkle-free and the transition between different colors seems very natural and smooth. The dress seems to reflect a reddish gloss from the light, which shows the texture of the fabric. The decorative lace on the collar and cuff is also depicted in detail.

With regard to colors, the left hand side of the painting is relatively more vibrant than the right hand side. The Virgin Mary's red and blue garments, together with the red paint on the palette, contrast with Sofonisba's black and white garments. The skin tones of all three figures are quite warm, but that of the Virgin Mary and the Christ Child is slightly paler than Sofonisba's. This may be due to the fact that in the painting Sofonisba is an actual person while the Virgin Mary and the Christ Child are painted figures on the canvas. The reddish cheek, ear and lips of Sofonisba echo the red paint and the red garment of the Virgin Mary and create a sense of consistency and uniformity.

In addition to the disparate skin tones of the figures, the self-portrait is painted in a different style from the Virgin Mary and the Christ Child. The self-portrait has more volume and provides a sense of solidity and three-dimensionality, while the figures in the "painting" appear to be more transparent, elongated and less substantial.

Last but not least, Sofonisba utilizes a range of strategies to create a sense of space. Besides the shadows and highlights, the overlap between the artist's right hand and easel shows Sofonisba's position in front of the painting. Moreover, the easel is placed at an angle with perspective, while the "painting" itself does not show a corresponding perspective. In addition, behind the Virgin Mary and the Christ Child, the dark half of the background may be regarded as a solid wall, while the bright half of the background shows the scenery behind the wall, extending the plane farther into the distance.

Some similarities and differences can be discovered among these three female self-portraits. Caterina appears to be somehow conservative and reserved with a relatively inert expression; Lavinia prefers a more flamboyant style, while Sofonisba falls in between, meeting society's expectations of women to be modest, while showing off her nobility and capability unreservedly. The slits on the top of her sleeves and the glossy fabrics symbolize aristocracy and nobility, yet the color and the dress's design prevent Sofonisba from appearing overly ostentatious – indeed, she seems like a serious scholar, dedicating herself to her work. If we take a close look at Sofonisba and Caterina, we can see that although the compositions of the two paintings are very similar, Sofonisba's face is much more vivid, with glowing eyes and rosy complexion, which makes her look younger than her actual age at that time. She also added complexity to her character by contrasting a girlish face with a serious expression. This contrast may be explained by Sofonisba's educational background and social status, which may have made her more mature. This is probably a result of her intention to show the viewer both youth and wisdom. Moreover, Sofonisba presents her almost-finished religious work clearly to the viewer, while Lavinia does not include an easel in her self-portrait at all and Caterina only shows a blurry face of the character on the canvas. The viewer may not find any trace of devotion to Christianity in the latter two paintings. Despite their differences, all three artists painted themselves in formal and upper-class clothes, which would not be worn in reality while working. This not only improves their external image but also allows them to demonstrate their exquisite technique.

If we compare Sofonisba's self-representation to that of her male counterparts introduced earlier in this section – Michelangelo, Tiziano to Anthony van Dyck – we likewise find similarities and differences. Presenting herself as a modest female artist, Sofonisba neither directly endows herself with a religious or symbolic identity as

Michelangelo and Tiziano do in *Last Judgment* and *Allegory of Prudence*, nor overly flaunts her aristocratic status as van Dyck does in his *Self-Portrait with a Sunflower*. However, she includes a religious painting in her self-portrait and wears a modest yet typical aristocratic outfit, which implicitly associates herself with both Christianity and the nobility. Although the strategy used by Sofonisba differs from those used by male artists, it would be reasonable to assert that she shares some traits with her male counterparts.

It is also worthwhile to compare the way in which Sofonisba and her male counterparts depict themselves. Examining Sofonisba's self-portrait and Tiziano's 1567 *Self-Portrait*, the viewer can tell that Sofonisba sits in a similar position as Tiziano does and her dark, formal outfit echoes Tiziano's, conveying a feeling of solemnity and formality. Nevertheless, Sofonisba applies a lighter and more reddish tone to her face than Tiziano does and hence creates a sense of youth and vitality, while Tiziano's wrinkled face emphasizes his age and serious personality. Also, Sofonisba's act of painting injects motion into the painting, which contrasts with the stationary aura of Tiziano's self-portrait. Moreover, while Sofonisba is gazing at the viewer, perhaps in order to engage the viewer into the painting, Tiziano appears to be relatively aloof, staring into space and showing no attempts at interaction. Anthony van Dyck engages in a form of interaction with the viewer that is similar to Sofonisba's and creates a similar sense of vitality through the use of light and rich color. However, at first glance, the viewer's attention may be drawn to the sunflower that he is pointing at or the golden chain that he is wearing. Van Dyck's pose is more complex and deliberate than Sofonisba's and conveys the sense that the main focus of this dazzling self-portrait is not the artist himself but his status and wealth, as represented by the sunflower and the golden chain. Sofonisba, although also an aristocrat, possesses a sense of modesty and reveals her nobility implicitly. She keeps the painting focused on herself, her profession, and the painting that she is working on.

Speaking generally, Renaissance men and women present themselves differently in their self-portraits, and the difference may lie in their self-perceptions and ambitions. Male artists seem to have been eager to associate themselves with religion and mythology, while female artists were more likely to focus on themselves and their profession, which they indicate through the inclusion of brushes or easels. Sofonisba could identify with both sides – while she painted her image vividly, highlighting her own identity and traits just as many other female artists did, she endows the self-portrait with a deeper meaning, as many male artists did, by including the painting of the Virgin Mary and the Christ Child. Besides showing her religious devotion, Sofonisba may also have wanted to highlight her expertise in a male-dominated genre – that of historical and religious painting.

Religious Paintings in the Renaissance

The role and significance of religious paintings changed in the Renaissance. Before this period, painting was seen purely as a tool to spread Christianity and control the minds of the public, particularly those who could not read. During the Renaissance, however, the aesthetic value of religious paintings was raised to a level equal to their spiritual quality. Artists still made fine religious works, but they endowed Biblical figures with more human traits rather than depicting them as immortal, invisible and unapproachable beings. There was also a revival of paganism, as if those masters were just “as happy painting scenes from Greek mythology as they were the Biblical stories and handle them in the same way,” writes Felix Arnott (42). Shedding the rigid restrictions of the previous era, aspiring artists were able to express their subjective thoughts in religious paintings and therefore seized opportunities to create them not only to draw attention from the court and the papacy but

also to demonstrate their expertise and creativity and to leave a stroke on the glorious history of the Renaissance.

However, due to the fact that Renaissance art was dominated by male artists, female artists lacked commissions to paint religious figures; rather, they were expected to specialize in portraiture, perhaps because, as Margaret L. King has argued, “executing a portrait required admitting a painter to a domestic setting for protracted sessions. Or perhaps it was thought that a woman painter had special empathy that would permit a truthful and intimate understanding of the subject” (274-275). Their modesty stopped them from chasing ambition, and their feminine traits were believed to give them an advantage in portraying their clients.

Sofonisba Anguissola, though a master in portraiture, boldly included an iconic scene of the Virgin Mary and the Christ Child in her self-portrait. The “painting” in the self-portrait may actually exist separately, but no document shows that it has been preserved. In contrast with Sofonisba’s solemnity, these two biblical characters are in a casual position, inclining to one another. The subject of the Virgin Mary and the Christ Child had been commonplace since the Middle Ages, and in this painting, the figures may remind the viewer of the medieval style; their bodies, especially the fingers and necks, are elongated and not as round as that of the artist herself. Indeed, Sofonisba painted herself in a different, typical Renaissance style under the influence of her mentors. The intimacy between the Virgin and Christ is depicted in many paintings with the Virgin holding Christ in her arms, and in some of them, the characters’ faces are very close to one another. Sofonisba turned this conventional gesture into a gentle kiss (or a lovely whisper). The portrayal of the Virgin Mary and the Christ Child is so vivid and tender that they seem to have walked out of the altar and turned into a normal yet loving mother-and-son pair. The sentiment between them is successfully “sensed,” “caught” and “preserved” by Sofonisba. She produced several other formal religious paintings during her apprenticeship with local artists in her mature years. Although the composition of the majority of her religious works was based on existing works by other artists (Sylvia Ferino-Pagden and Maria Kusche, 95), and only very few of them can be accessed – *Pietà* (about 1550, Figure 8.), *Holy Family* (c.1559, Figure 9.) and *Nursing Madonna* (c.1588, Figure 10.) – her mastery in depicting sentiment within Biblical figures, no matter whether it be sadness or joy, is still apparent.

The most surprising fact about Sofonisba Anguissola is not that she produced some religious paintings; it is her decision to place a religious painting in her self-portrait, to show herself painting the Virgin and Christ. Various intentions can be interpreted from it.

First of all, Sofonisba may want to express her devotion to Christianity, as introduced earlier. Portraying the Virgin and the Christ Child can be regarded as a means to praise them. This could also explain why she always depicts herself in a black outfit – black is the color of modesty, seriousness and elegance, and matches the principles of Christianity.

Secondly, perhaps she wanted to challenge the discriminative idea that only male artists could paint religious figures, and to demonstrate her capability and expertise. Although the word “feminism” did not exist in the Renaissance, this feminist idea may have come from the well-rounded education she received. Nevertheless, she was not really oppressed by men; rather, she received generous help and recognition from male masters and aristocrats such as Bernardino Campi, Bernardino Gatti, Michelangelo, Giorgio Vasari and the Spanish King Philip II during her lifetime. They were her mentors and sponsors, rather than competitors, so it is unlikely that Sofonisba held an aggressive attitude towards them. Regardless, it is possible she attempted to deliver such a message to the public.

Thirdly, by placing the Virgin, Christ and herself in the same picture, she may try to show that she shared the same virtue with them. To some extent, this can be regarded as a

less radical version of Albrecht Dürer's 1500 *Self-Portrait* (Figure 11.), in which he depicts and hence identifies himself with Christ. Sofonisba is not as bold as Dürer, yet the way she composes the painting – putting the brush right in front of the easel while staring at the viewer – gives people an illusion that she is painting from life, and that the biblical figures are present. By doing so, Sofonisba further endows herself with holiness and imply that she is beyond an ordinary artist – at least, she is able to “share space” with the Virgin Mary and the Christ Child and paint them with their “consent.”

Comparing Sofonisba's self-portrait with the Early Dutch painter Rogier van der Weyden's *Saint Luke Drawing the Virgin* (c.1435-1440, Figure 12.), we can find some intriguing similarities. The latter painting shows Luke the Evangelist, who is believed to be the first icon painter, sketching the Virgin Mary while she nurses the Christ Child. St. Luke's face is widely considered to be a self-portrait of van der Weyden (Chiyo Ishikawa, 54). By painting himself so, the artist associates himself both with Biblical figures and with the founder of icons. Both paintings include self-portraits and religious scenes; however, while van der Weyden presents himself as St. Luke, sketching the Virgin and Christ in their presence, Sofonisba neither endows herself with a second identity, nor clarifies whether she is painting from life or from imagination. Perhaps this is because Sofonisba is not able to paint herself as St. Luke due to her gender, but nonetheless she is confident enough to identify herself with biblical figures directly and show her identity as a skillful female artist who has expertise in both portraiture and religious works.

Epilogue

To conclude, although Sofonisba's fame is not comparable to that of her male counterparts, she did shine on the splendid stage of the Renaissance, demonstrating her outstanding talent with courage and confidence. Some critics may argue that a large proportion of her success can be credited to her noble background; however, from my perspective, it was her ambition and rebelliousness that made her not only famous but unique – she was never a conventional woman, confronting gender stereotypes in her works, marrying a man who was much younger than she was in her middle age and bearing no children. This *Self-Portrait at the Easel (Painting a Devotional Panel)* provides an insight for the contemporary viewer into her legendary life and extraordinary ideas about her own identity and the broader society. Together with other aspiring female artists, she set a successful example for gifted women in younger generations and opened a window for them to pursue a career in artistic fields. Her legacy continues today and will not diminish with time.



Figure 1. *Sofonisba Anguissola, Self-Portrait at the Easel (Painting a Devotional Panel), c. 1556. Oil on canvas, 26.0 × 22.4 in. Zamek Lubomirskich i Potockich, Łańcut, Poland*

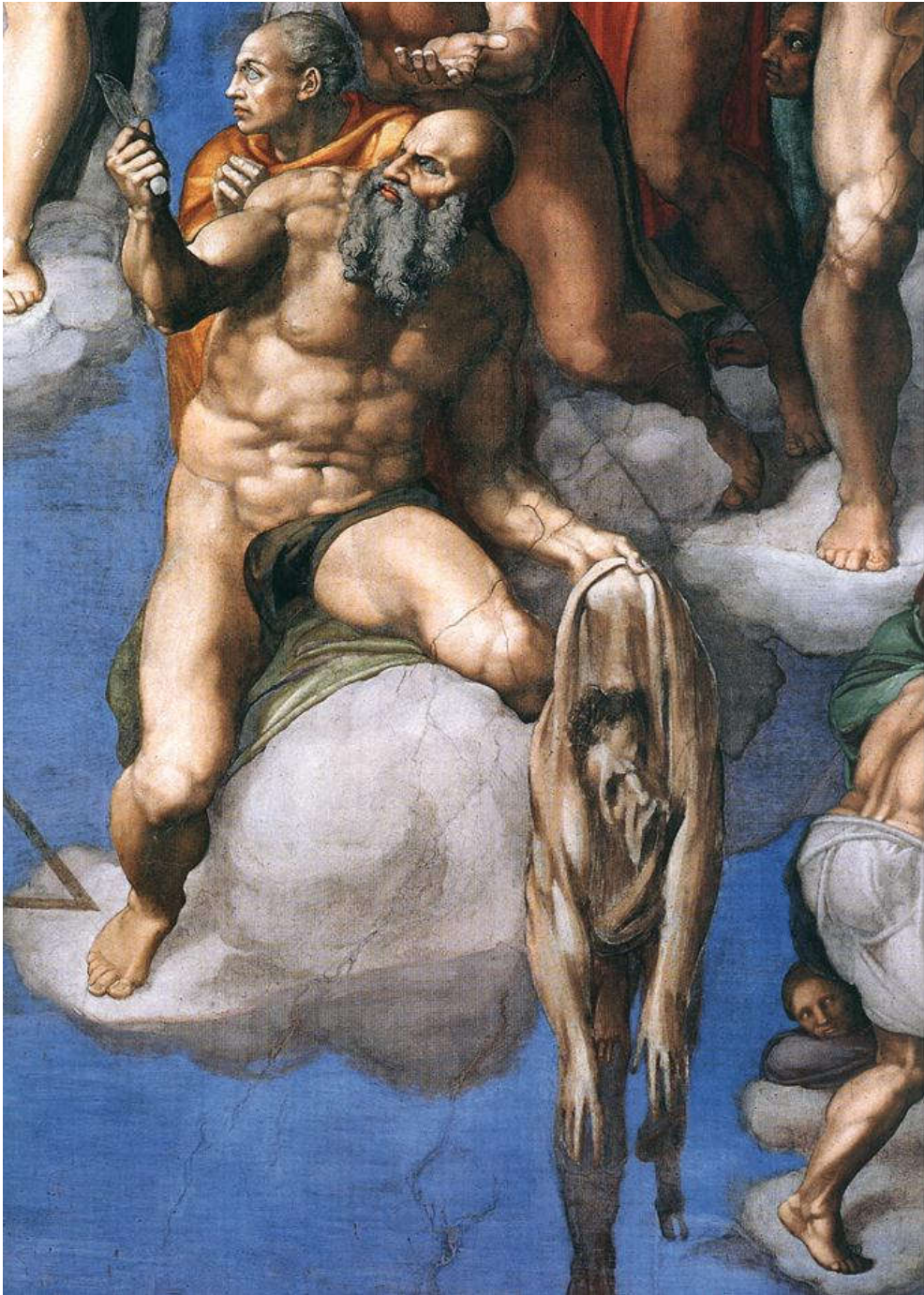


Figure 2. Michelangelo Buonarroti, *The Last Judgment*, c. 1536–1541. Fresco, 539.3 × 472.4 in. Sistine Chapel, Vatican City



Figure 3. Tiziano Vecellio, Self-Portrait, c. 1567. Oil on canvas, 33.9 × 25.6 in. Museo del Prado, Madrid



Figure 4. Tiziano Vecellio, *Allegory of Prudence*, c. 1565–1570.
Oil on canvas, 30.0 × 27.0 in. National Gallery, London



*Figure 5. Anthony van Dyck, Self-Portrait with a Sunflower, c. 1632-1633.
Oil on canvas, 23.6 × 28.7 in. Private Collection of the
Duke of Westminster*

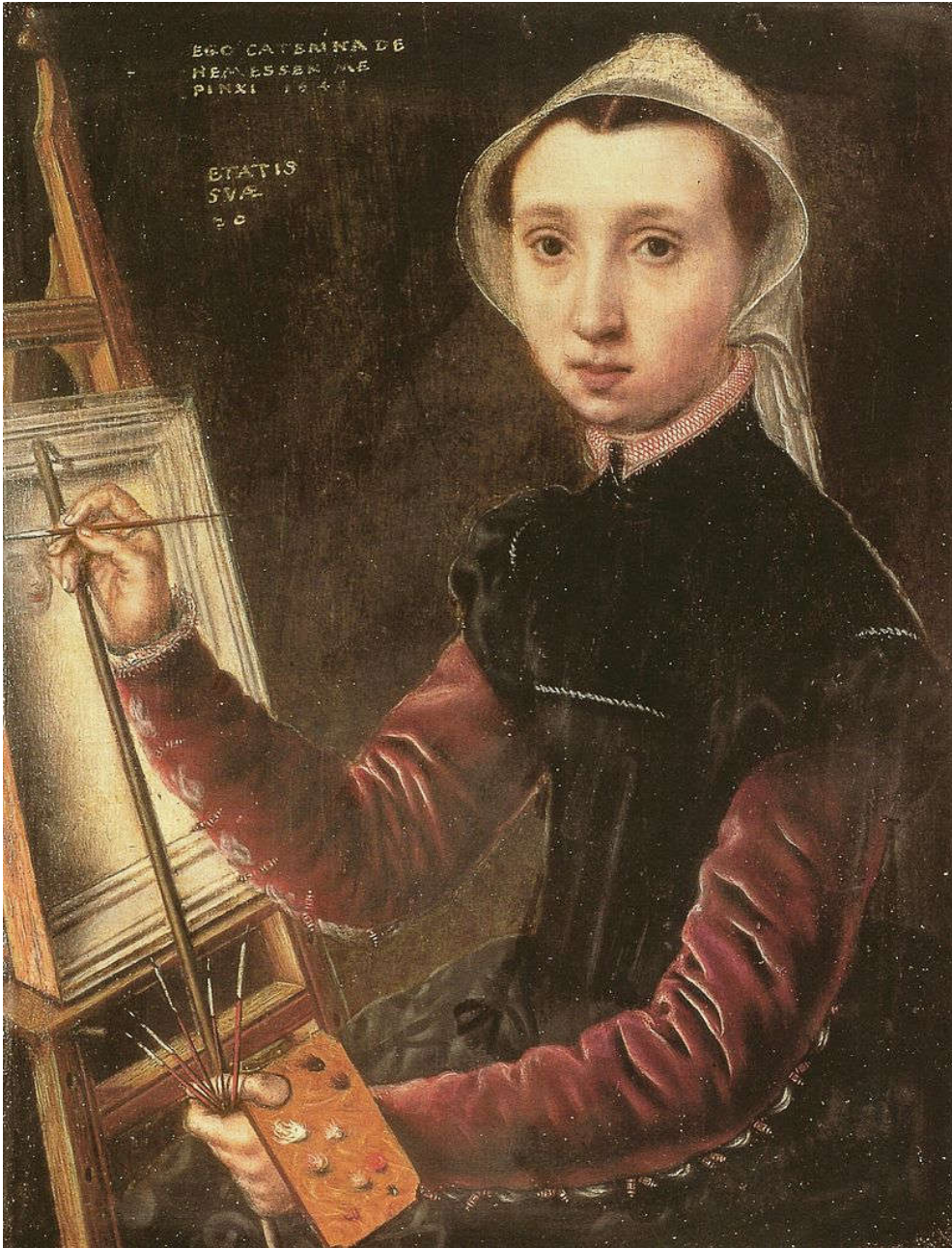


Figure 6. *Caterina van Hemessen, Self-Portrait, c.1548. Oil on oak, 12.1 × 9.6 in. Öffentliche Kunstsammlung, Basel*



Figure 7. *Lavinia Fontana, Self-Portrait with Palette and Brushes, c. 1579. Oil on canvas, unknown dimension Uffizi Gallery, Florence*



Figure 8. Sofonisba Anguissola, Pietà, c. about 1550. Oil on canvas, 18.3 × 12.8 in. Pinacoteca di Brera, Milan



Figure 9. Sofonisba Anguissola, *Holy Family*, c. 1559. Oil on canvas, 14.5 × 12.3 in. Accademia Carrara, Bergamo

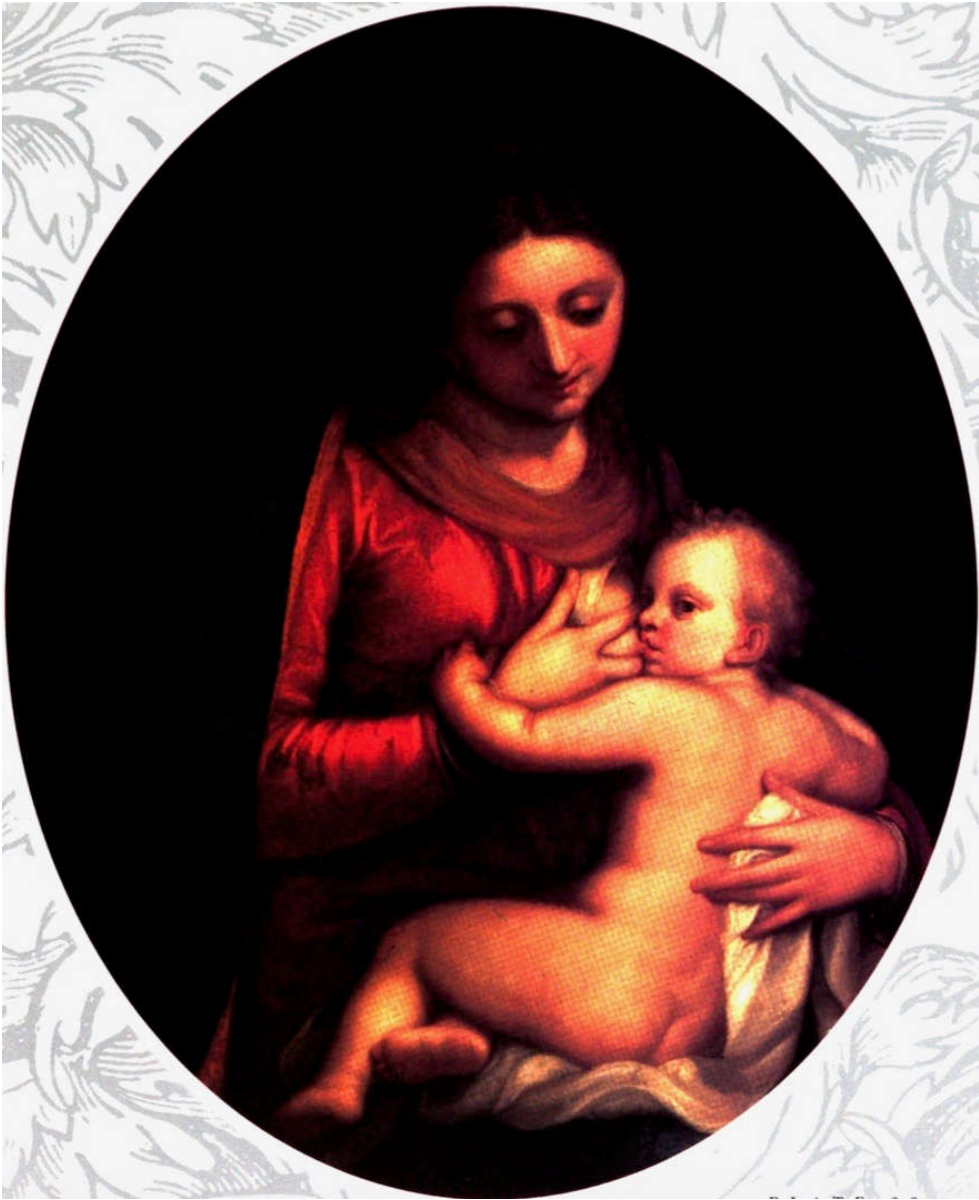


Figure 10. Sofonisba Anguissola, Nursing Madonna, c. 1588. Oil on canvas, 30.3 × 25 in. Szépművészeti Múzeum, Budapest



Figure 11. Albrecht Dürer, *Self-Portrait*, c. 1500.
Oil on wood panel, 26.1 × 19.3 in. Alte Pinakothek, Munich



Figure 12. Rogier van der Weyden, Saint Luke Drawing the Virgin, c. 1435-1440. Oil and tempera on oak panel, 54.1 × 43.6 in., Museum of Fine Arts, Boston

Sources

- Arnott, Felix. "Religious Art." *The Australian Quarterly*, vol. 28, no. 2, June 1956, pp. 36–46. *JSTOR*, <http://www.jstor.org/stable/41317770>. Accessed 11 July 2017.
- Borzello, Frances. *Seeing Ourselves: Women's Self-Portraits*. London, Thames & Hudson, 2016.
- Burckhardt, Jacob, and Samuel George Chetwynd Middlemore. *The Civilization of the Period of the Renaissance in Italy*. London, C.K. Paul & Co., 1878.
- Ferino-Pagden, Sylvia, and Maria Kusche. *Sofonisba Anguissola: A Renaissance Woman*. Washington, National Museum of Women in the Arts, 1995.
- Fulmer, Betsy. "Sofonisba Anguissola: Marvel of Nature." *Academic Forum*, no. 23, 2005, pp. 20–34.
- Garrard, Mary D. "Here's Looking at Me: Sofonisba Anguissola and the Problem of the Woman Artist." *Renaissance Quarterly*, vol. 47, no. 3, 1994, pp. 556–622. *JSTOR*, doi:10.2307/2863021. Accessed 18 June 2017.
- Glenn, Sharlee Mullins. "Sofonisba Anguissola: History's Forgotten Prodigy." *Women's Studies*, vol. 18, no. 2-3, 1990, pp. 295–308. doi:10.1080/00497878.1990.9978837.
- Hall, Marcia B. "Michelangelo's Last Judgment: Resurrection of the Body and Predestination." *The Art Bulletin*, vol. 58, no. 1, 1976, pp. 85–92. *JSTOR*, doi:10.2307/3049465. Accessed 1 July 2017.
- Ishikawa, Chiyo. "Rogier van der Weyden's 'Saint Luke Drawing the Virgin' Reexamined." *Journal of the Museum of Fine Arts*, vol. 2, 1990, pp. 49–64. *JSTOR*, www.jstor.org/stable/20519725. Accessed 21 July 2017.
- Kelley, Thomas. "Five Self-Portraits." *Log*, no. 31, 2014, pp. 82–85. *JSTOR*, www.jstor.org/stable/43630893. Accessed 1 July 2017.
- Kelly-Gadol, Joan. *Did Women Have a Renaissance?* Boston: Houghton Mifflin, 1977.
- King, Margaret L. *A Short History of the Renaissance in Europe*. North York, Ontario, Canada: U of Toronto Press, 2017.
- Mciver, Katherine A. "Lavinia Fontana's 'Self-Portrait Making Music'." *Woman's Art Journal*, vol. 19, no. 1, 1998, pp. 3–8. *JSTOR*, doi:10.2307/1358647. Accessed 2 July 2017.
- Panofsky, Erwin. *Meaning in the Visual Arts: Papers in and on Art History*. Garden City, NY, Doubleday Anchor Books, 1955.
- Rosenthal, Lisa. "Review: The Look of Van Dyck: 'The Self-Portrait with a Sunflower' and the Vision of the Painter." *CAA. Reviews*, 26 Mar. 2008, www.caareviews.org/reviews/1098#.WXgc79OGMnU. Accessed 26 July 2017.
- Symonds, John Addington, and Michelangelo Buonarroti. *The Life of Michelangelo Buonarroti: Based on Studies in the Archives of the Buonarroti Family at Florence*. Vol. 2. Philadelphia, PA, University of Pennsylvania Press, 2002.
- Vasari, Giorgio. *Lives of the Most Eminent Painters, Sculptors and Architects*. 2nd ed., vol. 5, London, Philip Lee Warner: The Medici Society, 1912.
- Zirpolo, Lilian H. *Historical Dictionary of Renaissance Art*. Lanham, Rowman & Littlefield, 2016.



The Effect of Political and Cultural Ideas on Clothing Reform, Chinese Hair, and Clothing Fashion, 1890-1910s

Yanyu Zhong

Author Background: Yanyu Zhong grew up in China and currently attends Shenzhen Middle School in Shenzhen, China. Her Pioneer seminar topic was in the field of history and titled "Qing China and the West: Conflict and Cultural Exchange."

Abstract

The conflict between Qing China and the western countries in the late nineteenth to early twentieth century led to an extensive cultural and ideological exchange between the east and the west. During this time, reformers and revolutionaries had debated over governmental, political, martial, and social issues, including the value of traditional social practice. This paper examines the effect that the then modern notions advocated by reformers and revolutionaries had on the clothing reform in the early twentieth century and how the 1911 Revolution helped define what clothing reform would become. It argues that the clothing reform was, to some extent, influenced by 1898 reformers' political and cultural ideas, as revolutionists in the twentieth century constantly referred to and modified the former reformers' arguments to promote the clothing reform to the public.

Introduction

Clothing is one of the most evident indicators of cultural influence. Sometimes, a change in fashion reveals the encounters between different cultures. Home to diverse cultural and ethnic groups, China has been through several significant sartorial revolutions over time. Specifically, two types of interaction propelled the revolutions—interaction among different ethnic groups and with various foreign, especially western, countries. Before the nineteenth century, changes in Chinese hair and clothing styles were, more or less, the result of internal cultural changes, sometimes driven by fashion and/or interactions between various ethnic groups. For example, the Northern Wei, one of the imperial states during the period of Five Dynasties and Ten Reigns, adopted Han-style clothing once it rose to power.

In addition, shifts in clothing styles were also driven by contemporary analyses of clothing's political and civilizational significance. Chinese literati and part of the public believed that clothing helped shape people's behavior, particularly in the nineteenth and twentieth centuries. One could change one's attitude, therefore, by changing one's garments or hairstyle.

After the First Opium War, China was forced to open its trading ports in Guangzhou, Xiamen, Fuzhou, Ningbo, and Shanghai for foreign trade. Frequent economic contacts stimulated cultural exchanges between China and other countries, such as Britain and Japan. Sensing advanced and sophisticated western technologies and ideas, some Chinese literati and government officials began to consider implementing public clothing

reform to shape a so-called “modern” and “civilized” nation. Consequently, foreign influence started to generate profound vestimentary change, which, according to Antonia Finnane, “over the next half century was to transform the clothing culture in urban China and in much of rural China.”¹ This transformation, however, did not result in publicly visible effects until the early twentieth century. In fact, during the late imperial period, although new notions and attitudes regarding clothing reform had already taken effect throughout the country, the costume of Qing officials and most ordinary citizens remained unchanged—long robes, loose sleeves with rich embroideries on the borders, hats for the officials, and plain shirts and loose trousers or skirts for commoners. Taken in the 1890s in Taigu, Shanxi, the photograph in Figure 1. features a typical Chinese family. The man sitting on the far right is Liu Facheng, a converted Christian. He wears a long gown (*changpao*), the typical garment for a non-laboring man, and is also clad in a “horse jacket” (*magua*). These two garments were, along with the *majia* (a sleeveless vest), the standard wear in the late imperial period. In addition to garments advocated by both reformers and revolutionists, men cutting queues remained unpopular before the 1911 revolution. Common people still considered queue cutting offensive. Lu Xun, one of the most influential writers in contemporary China, came back to China after finishing his study in Japan to teach in two middle schools. Seeing his short hair and western attire, the school officials once suspected that Lu attempted to encourage his students to cut their queues.² This incident demonstrates that even though modern ideas appeared in the late nineteenth century, they mostly influenced overseas students and officials, resulting in a gap between modern notions and their effect on clothing reform.

Although clothing reform was not apparent on the streets of late nineteenth-century China, the new arguments regarding traditional garments, queues, and foot-binding played an important role in accelerating clothing reform. Many scholarly papers and books have discussed clothing reform during the late imperial and early Republican era. Nonetheless, few have focused on the widely divergent reasons reformers had for advocating such changes, nor have scholars traced the impact of these ideas. This paper will examine the extent to which the clothing reform in the early twentieth century was influenced by the then modern notions promoted by reformers and revolutionaries and how the revolutions of the early twentieth century helped paint the picture of future clothing reform.

¹ Antonia Finnane, *Changing clothes in china: Fashion, history, nation* (New York: Columbia University Press, 2008), 2.

² Eva Shan Chou, “‘A Story about Hair’: A Curious Mirror of Lu Xun’s Pre-Republican Years,” *The Journal of Asian Studies* 66, no. 2 (2007): 423.



Figure 1. Source: Deacon Liu Fa-cheng and his family in Taigu, ca. 1890s, Oberlin College Archives

Arguments on Clothing Reform in Late Imperial China

Frequent contact with foreign traders, missionaries, and officials and failure in military conflicts with foreign powers engendered a new notion in Chinese society—*Xixue* (western learning). In 1895, after the fall of Weihaiwei Naval Base, China lost the Sino-Japanese War of 1894-1895. The Chinese imperial navy's failure was commonly regarded as marking the failure of the Self-Strengthening Movement in reviving Chinese military power, even though some argued that China did acquire certain advanced military technologies during the movement. At that time, Chinese literati who had advocated *Xixue* realized that learning solely from western technology was not enough. As a result, they turned to the study of western political ideologies as well. Some of them, including Tan Sitong, Kang Youwei, and Liang Qichao, believed that the Qing government needed reform. Despite this common core outlook, however, the reformers held relatively different viewpoints regarding clothing reform, including garments, queues, and foot-binding. While arguments on anti-footbinding emphasized gender equality,³ those on garments and queues

³ Some reformers and activists argued that unbinding women's feet allowed them to move deftly, and therefore they could perform more physically demanding works, such as reeling in textile factories. These factories offered female workers considerable amount of salaries (some offered as much as ten dollars a month, according to "Remarkable Progress of The Anti-Footbinding Movement" by Alicia Little) which they could never earn otherwise. Unbinding feet granted Chinese women at least a small degree of economic autonomy and social mobility and could be seen as a step towards gender equality.

were once endowed with political significance⁴ and, therefore, related more closely to the political revolutions in the twentieth century. For the sake of relevance and clarity, I will only focus on the change of garments and hairstyle in the paper.

Tan Sitong discussed the significance of queues in his work *Exposition of Benevolence*. His argument centered on what he saw as the unnecessary complexity of traditional Chinese clothing. He argued that traditional garments, including the *majia* and *magua*, restrained people from conducting events which were of greater importance, such as participating in military reform or learning western technologies and ideologies which were perceived as advanced. In addition, he accused the Manchus of confusing Han men's minds and stifling their intelligence by making them wear queues. From his point of view, wearing queues, along with other customs such as kneeling and bowing, was so complicated that "people have to spend all their efforts practicing [it] in order to prevent themselves from offending the authorities." Being occupied with "these irrational practices,"⁵ people, therefore, could not ponder the real issue—in this case, rebellion. Furthermore, seeing the success in political and military reform that Japan's Meiji Reform had, Tan assumed that clothing reform was the propeller of Japan's success. Consequently, by contrasting China's traditional long robes and wide sleeves with Japan's militarized new garments after the Meiji Reform, he suggested that his own nation ought to stop the shaving of foreheads and cut queues in order to stay away from "the northern barbarians—"⁶ Manchu people.

Kang Youwei, one of the most prominent political reformers and social idealists in the nineteenth century, supported clothing reform from another perspective—the promotion of the unity of humanity. According to Gongquan Xiao, Kang "was convinced that unity was an indispensable condition of political well-being," and that "peace and progress prevailed only when the empire was under a single rule."⁷ The various styles of Chinese clothing, therefore, made Kang conclude that conflicts China had with other countries in the nineteenth century were a consequence of excessive sartorial styles:

"...the myriad nations are all in communication and are as one moving towards veneration for oneness. It is just in our country that clothes are different, so that sentiments cannot be close and friendly relations between nations cannot be achieved."⁸

Consequently, Kang believed that Chinese people should unify their clothing in order to reach "peace and progress" both among different ethnic groups inside China and with various countries around China. Furthermore, Kang also believed that queue cutting, as

⁴ Qing vestimentary and hair styles were enforced by the Qing government, which mainly consisted of Manchu officials. As I elaborated later, the garments and queues became symbols of the Manchus. Consequently, oppositions to Qing clothing and hairstyle were later endowed with the political implication of opposing the Qing government in general.

⁵ Chan Sin-wai, *Tan Sitong, A Book of Benevolence Translation* (Hong Kong: Chinese University Press, 1984), 203.

⁶ Chan Sin-wai, 204.

⁷ Gongquan Xiao, *A modern china and a new world: K'ang yu-wei, reformer and utopian, 1858-1927*. Vol. no. 25. (Seattle: University of Washington Press, 1975), 99.

⁸ Luke S.K. Kwong, *A Mosaic of a Hundred Hays: Personalities, Politics and Ideas of 1898* (Cambridge: Mass.: Council on East Asian Studies, 1984), 201, 318. Cf. Edward J.M. Rhoads, *Manchus and Han: Ethnic Relations and Political Power in Late Qing and Early Republican China, 1861-1928* (Seattle: University of Washington Press, 2000), 65.

part of the clothing reform, could result in the rise of military power. He linked the western countries refined military weapons with the absence of queues on westerners and, therefore, reached the conclusion that “if everyone [in China] cuts his hair, the whole nation will become soldiers.”⁹ His views, however, did not become popular until the twentieth century. China in the early twentieth century was, for a long period of time, invaded by imperialist powers, including Britain, France, Germany, Japan, Russia, Italy, the United States, and the Austro-Hungarian Empire. Furthermore, with the introduction of democracy and a republican regime, Chinese feudalism, which the Qing government and its predecessors adhered to, gradually became out of date in a modern, European-dominated world in which capitalism and imperialism flourished. In order to protect the integrity of China’s territory and overthrow the Qing government and feudalism, Chinese revolutionaries gradually adopted Kang’s queue-cutting argument for enhancing military power.

In addition to queues’ purported disadvantages, cutting queues was an anti-Manchu symbol. Zhang Taiyan, one of the foremost pioneers of the anti-Manchu movement, believed that queues, together with shaved heads, were a symbol of the Manchu, the barbarians. Although historians were not certain about its origin, a commonly held explanation for the symbolism of the queue drew connection with the queue’s origin. Also known as *bian* (plaited cords), queues originated from northern Xiongnu tribes in the Han dynasty. From then on, other nomadic tribes of the far west, including the Manchus, adopted and preserved *bian* as a traditional hair-style.¹⁰ Consequently, queues were often seen as a symbol of the barbarians, calling up their Xiongnu origin. Zhang’s argument was similar to, yet more radical than, that of the other reformers. He argued that queues were not part of the Han tradition, and that Han people were not supposed to wear queues. It was the Manchus, i.e., the Qing dynasty, who made Han people wear queues as a form of enslavement. As a result, he cut his queue as a declaration of his rupture with the Qing government. Despite its aggressiveness, Zhang’s argument was popular among revolutionaries. For example, Sun Yat-sen, the provisional President of the Republic of China in 1912, wrote in his instruction to the Ministry of Internal Affairs about queue cutting that “the Manchu nation of looters and thieves changed our hats and attire and forcibly implemented the system of braided hair, so that we must completely abide by their putrid customs.”¹¹ From a contemporary perspective, such strong disparaging language was rather anti-ethnic. In fact, for the Manchus, wearing queues was not a “putrid custom” but rather a necessary practice. Originally nomadic people, the Manchus conducted most of their activities, such as grazing cattle, herding sheep or fighting, on horsebacks. In order to freely use their bow and arrow while galloping at full speed at the same time, the Manchu people invented a saddle on which one could put one’s feet. This invention allowed the Manchus to gallop freely while standing and shooting with arrows. The Manchu people’s long hair, however, could easily block their sight when galloping at full speed. Consequently, the Manchus had to shave off their front hair and braided their back hair in plaits to prevent it from hindering their movements on horseback. Given that wearing queues were of practical use, Sun Yat-sen’s pronouncement, indeed, vilified Manchu customs. Nevertheless, one could argue that it was

⁹ Li Zhigang (Lai Chi-kung), “Xianxiang yu yingzao guozu: jindai Zhongguo de faxing wenti” (Imagining and constructing nationhood: hairstyles in Modern China), *Si yu yan* 36, 1 (1998): 110.

¹⁰ Michael R Godley, “The end of the queue: hair as symbol in Chinese history,” *East Asian History*, no. 8 (December 1, 1994): 54.

¹¹ *Linshi zhengfu gongbao* (Bulletin of the Provisional Government), 29 (March 5, 1912).

necessary rhetoric to appeal to the public's disaffection towards the Qing government, which had just been overthrown, in order to stabilize the new Republican regime.

Though most supporters of clothing reform directly connected arguments on clothing with their political beliefs, some cut their queues and changed their outfits merely to follow the then current trend. In 1872, 120 schoolboys from the Chinese Educational Mission (CEM) set sail across the sea to the US to spend nine years for higher education. The first group of students who were supported by the Qing government to study abroad, were required to wear their queues at all times; any violation of the rule would result in dismissal.¹² This was because wearing queues had been part of Chinese tradition. When the Manchu assumed power from the Ming dynasty in the 1640s, the Qing government commanded the Han to wear queues in order to consolidate its sovereignty. This was because the wearing of the queue demonstrated that one accepted, or at least was open to, Manchu rule. In other words, it was a clear declaration of political allegiance to the Qing government. Since in Han Chinese Confucian tradition, hair was given by one's parents and, as a parental gift, ought not to be cut, some Han people resisted in both cutting their front hair and wearing queues on the back of their heads. Without any other means to popularize queues, the Qing government coerced the Han people—they would be executed if they refused to wear queues. One faced the choice of one's hair or one's head. The Han people, therefore, gradually became accustomed to wearing queues, and by the late nineteenth-century when reformers raised their objections, considered it a Han Chinese tradition. Taking wearing queues for granted, both the Han and Qing government officials insisted that people could not break this tradition. The long plaits and loose garments, however, made boys appear girlish in the eyes of their American peers. The boys were often laughed at and ridiculed by their American classmates. Consequently, most CEM students desired a clean hair-cut and a western outfit that could help them fit in or, at least, blend in with American society. Through the memoir of Major Fred G. Blakeslee of Harford, whose family received seven CEM students over the nine years, we can get a glimpse of the students' eagerness to change their appearance:

“When the boys arrived, they were dressed in Chinese costume, but they soon discarded this and appeared in American clothes. As it was a beheading offense to cut off the cue, they were obliged to retain their pigtails, but they made them as inconspicuous as possible by wearing them under their clothing.¹³”

Some of them did try to cut their queues—namely Zhong Juncheng and Zeng Pu. Both boys were expelled from the CEM program, although Zeng Pu somehow finished his education in the absence of a Chinese government scholarship. Similar to the CEM students, other overseas Chinese students later experienced the same situation in other countries. One of them was Duan Qirui, the president of the Republic of China from 1924 to 1926. Like other Chinese students, he hid his queue under his hat to reduce the uneasiness of being stared at by others when he was studying abroad in Germany. His behavior, however, was reproached by one of the Chinese officials. In his fury, Duan swore that he would become

¹² Edward J. M. Rhoads, *Stepping Forth into the World* (Hong Kong: Hong Kong University Press, HKU, 2011), 150-151.

¹³ Arthur G. Robinson, “Pilgrims to Western seats of learning: China's first educational mission to the United States: the breaking of Chinese intellectual isolation,” *Chinese Studies In History* 36, no. 4 (June 1, 2003): 70.

insane rather than be teased by his classmates.¹⁴ Similar to Duan, Sun Yat-sen changed his “Chinese attire for a European costume” and removed his queue to allow his “hair to grow naturally”¹⁵ when he was about to set off to London on a Japanese steamer. Furthermore, not only did overseas students experience such humiliation, Qing diplomatic officials were rather discomforted with their queues as well. It was reported that “all two hundred top officials at the Ministry of Foreign Affairs, quickly shed their plait” as soon as the Qing government permitted everyone to “freely cut their hair.”¹⁶

In contrast to their insistence on maintaining the queue, however, the Qing officials were rather tolerant of students changing Chinese garb for Western raiment. For example, the Chinese commissioners allowed the CEM students to wear western clothing, such as trousers and shirts, on a daily basis.¹⁷ The students’ experience demonstrated the importance of queues to Qing government officials and the significant influence that changing one’s environment could have on one’s attitude.

Similarly, overseas revolutionists mainly based their anti-queue arguments on the humiliation that queues caused them. For instance, one revolutionist who studied in Japan commented on the Han officials who wore queues, saying their dignified manner had “reached its nadir and will soon be completely smothered.”¹⁸

Clothing Reform in the Early Republic of China Queues

The 1911 revolution not only brought new government, but also new notions to the nation. Just as wearing queues demonstrated one’s acceptance of Manchu rule, cutting queues could be seen as one’s political allegiance to the Republic regime, bearing a symbol of the Republican identity. As a result, an “anti-Manchu” and revolutionary symbol, queue cutting became popular across the nation. According to Zhang Shiyong and Carissa Fletcher, it “was virtually the first resolute and speedy measure adopted by the military government of each province.”¹⁹ As soon as the Nanjing Provisional Government issued an order that men cut their queues, many provinces posted new policy on compulsory queue cutting. For example, Jiangxi province required its officials and adult males to cut off their queues “within five days receipt of the directive.”²⁰ Some provinces even established “queue-cutting teams,”²¹ which were designed to cut queues for people who refused to do it themselves. In Tongzhou (today’s Nantong), a small city in Jiangsu province, the first official announcement published in the local newspaper *Xingbao* was the “Queue-cutting Pronouncement,” requiring every man to cut his queue, otherwise, members of the “queue-

¹⁴ Hua Mei, *Chinese Modern Clothing History* (Beijing, China Textile & Apparel Press, 2008), 16.

¹⁵ Sun Yat-sen, *Kidnapped in London: being the story of my capture by, detention at, and release from the Chinese legation, London* (Bristol: J.W. Arrowsmith, 1897), 28.

¹⁶ Edward J. M. Rhoads, *Manchus and Han: Ethnic Relations and Political Power in Late Qing and Early Republican China, 1861–1928*, (Seattle: University of Washington Press, 2006): 209.

¹⁷ Edward J. M. Rhoads, *Stepping Forth into the World*, 149-150.

¹⁸ Zou Rong, “*Geming jun*,” *Zou Rong, Chen Tianhua ji* (Collected Works of Zou Rong and Chen Tianhua) (Panyang: Liaoning remin chubanshe, 1996), 14.

¹⁹ Shiyong Zhang and Carissa Fletcher, “Subversive laughter: Carnival in the 1911 Revolution.” *Chinese Studies in History* 46, no. 1 (September 1, 2012): 37-38.

²⁰ Edward J. M. Rhoads, *Manchus and Han: Ethnic Relations and Political Power in Late Qing and Early Republican China, 1861–1928*, 252.

²¹ Shiyong Zhang and Carissa Fletcher, 39.

cutting teams” would cut it for him by force.²² Similarly, the Hubei military government sent queue-cutting troops to search for people on the street who still had their queues behind their backs. A witness of the forced queue-cutting process in Hunan province once recalled:

“At that time, most of the city gates and thoroughfares had people armed with scissors who were tasked with cutting the queues. I remember a few days after the first uprising, a man from Xiangyang, Zheng So-and-So, returned to Hubei from Hunan, with a pigtail hanging down from his head: he was noticed by someone and seen as a traitor to China—they nearly used force, but stopped after someone’s mediation.²³”

This almost coerced queue-cutting experience demonstrated that the seemingly successful and widespread hair-style reform was, in some ways, unethical and did not win support among the public.

Despite the resistance, the general public opinions on queue-cutting and the official announcement centered on the humiliation which wearing queues ostensibly brought. This could be explained by unpleasant experiences overseas students and revolutionists, who constituted a large percentage of Republican revolutionists, had had. Studying abroad just as their older generation had done in the late nineteenth century, modern revolutionists experienced the same humiliation as their predecessors. Furthermore, as modern ideas on equality and nationality gradually spread across the country, overseas Chinese revolutionists harbored more national pride than their older counterparts. They believed that their appearance represented China’s appearance. As a result, if they were humiliated by foreigners, China as a nation was humiliated by foreigners as well. Believing that wearing queues was the reason behind such disgrace, the new generation of revolutionaries, including Duan Qirui and Sun Yat-sen, advocated queue-cutting in order to prevent personal humiliation and defend national pride.

In contrast to conservatives, revolutionaries, students, and other supporters of the Republic of China willingly cut their queues. According to a report in the *New York Times* in May 1910, a year prior to the revolution, many people had cut “off their queues, an action which constitute[d] an anti-dynastic [anti-Manchu] demonstration.”²⁴ Similarly, another report from the *New York Times* showed that “between 700 and 800 Chinamen” had cut their queues in public in only one afternoon in Shanghai’s Chang Suho tea garden. Some advocates who had already cut their queues made small addresses on the convenience of short hair to the assemblage. More than that, a large number of people had their queues removed in private houses on that day as well. It was reported that similar events would be held “on the fifteenth day of each month.”²⁵ After the success of the revolution, Republican soldiers were one of the first and most resolute queue-cutting groups. A few years prior to the revolution, for instance, military schools that trained new-style soldiers had already become “hotbeds of dissent over hair-cuts.”²⁶ Additionally, after the revolution, soldiers and

²² Lu Yuan, “*Xingbao* and Queue-cutting Pronouncement.” *Nantong Weekly*, October 16, 2014, Section C, Final edition.

²³ Hubei Academy of Social Sciences, ed., *Xinhai Geming zai Hubei shiliao xuanji* (Selected Historical Materials on the 1911 Revolution in Hubei), (Wuhan: Hubei renmin chubanshe, 1981), 125–36.

²⁴ “Grave Signs in China.” *New York Times* (1857-1922), May 18, 1910.

²⁵ “Queue Cutting in China.” *New York Times* (1857-1922), Feb 26, 1911.

²⁶ Antonia Finnane, 79.

police were recruited in “queue-cutting groups” and actively participated in other queue-cutting movements throughout the country. For example, in Wuxi, the soldiers helped cut almost 1,000 queues each day. In Zhenjiang, the soldiers cut more than 1,000 queues in an evening.²⁷ Under the influence of the queue-cutting groups and advocates, queue-cutting was popularized “in almost all the provincial capitals of central and south China and also in many prefectural level towns”²⁸ by 1912, when Yuan Shikai took office.

The fact that queue-cutting happened first in armies after the revolution could be seen as early twentieth-century revolutionists’ reflection upon previous reformers’ idea that cutting off the queue could enhance military power.

“Civilized Clothes” and Militarization

Along with queue-cutting, the 1911 revolution established what Henrietta Harrison called “standards of Chinese-ness” and “modernity.”²⁹ The mixing between western and traditional Han style became popular in the public. The integration was especially prominent in the textile industry. As a support for the national textile industry, many garments in the early Republican era were made of domestic fabric and were made in western style. During his brief presidency, Sun Yat-sen actively advocated clothing reform. He argued that changes of regime had always been accompanied by clothing reform in the past and, therefore, the Republic of China needed to transform Manchu-style clothing into a modern style—combining western and traditional Chinese Han—to push China forward to “civilized countries.”³⁰

In addition to clothing’s westernization, the humiliating failures China had in the late nineteenth century made Chinese elites and literati realize how weak Chinese military power was compared to that of foreign countries such as Britain and Japan. The failures, therefore, stimulated a strong determination for military reform in the public. In addition, Japan’s victory over Russia in 1905 demonstrated that a reformed Asian country could defeat western countries.³¹ Consequently, clothing reform, which was advocated by reformers in 1898, was gradually accepted by the public and took its trend in militarization. As described by Antonia Finnane, the militarization of public costumes as a notion of enhancing China’s military power infiltrated the nation:

“In place of the long robe, male students began to don trousers and jacket...Straw boaters or military peaked caps replaced the traditional round cap. School uniforms were often modeled directly on military uniforms, as was the case in Japan, and commercial suppliers of military uniform touted for custom among students.”³²

Furthermore, after the revolution, clothing reform took on the trend of simplification, which stemmed from the idea of militarization. Embroideries and trimmings were replaced by plain shirts, and bright colors on garments, especially student garbs, and

²⁷ Henrietta Harrison, *The Making of the Republican Citizen: Political Ceremonies and Symbols in China, 1911-1929* (Oxford: Oxford University Press, 2000), 25.

²⁸ Henrietta Harrison, 28.

²⁹ Henrietta Harrison, 34.

³⁰ Ni Liangduan, “Sun Yat-sen as provincial president of the Republic of China,” *Shiji Fengcai* 11, no.2 (October, 2016): 27.

³¹ Antonia Finnane, 73.

³² Antonia Finnane, 77.

were reduced to black, grey or white. A new clothing style called “*Wenmingxinzhuang* (civilized new clothes)” was popularized among female students. The plain blue or white shirt with wide sleeves and black skirt became the standard uniform in many schools throughout the country. Aside from the idea of raiment militarization, such simplification echoed 1898 reformers’ idea that traditional Chinese garments were overly complex.

Conclusion

From the time it was proposed in the late nineteenth century to its implementation in the twentieth century, clothing reform was endowed with different meanings and potential effects that it might have on Chinese society. Although commonly regarded as a change of fashion (*Shimao*, “styles of the times”), clothing reform held not only sartorial meaning but also political implications. The foremost idea which was held by many literati and commoners was that clothing reform equaled westernization. The beliefs about the effects of westernization, however, varied among the public. Some, such as Tan Sitong, believed that westernization could help simplify Chinese clothing and customs and, therefore, free people’s minds from Manchu enslavement. Others, similar to Kang Youwei, believed that clothing reform was an indispensable process to enhance China’s military power and to reach the ultimate unity of humanity. Another group of advocates, which mainly consisted of overseas students and revolutionists, centered on the “normality” westernization brought—cutting queues and wearing European clothes could keep the students from westerners’ humiliation.

In addition to westernization, clothing reform also symbolized being “anti-Manchu.” This meaning was popular both in the late nineteenth century and the early twentieth century. The popularization of modern ideas on republican government and democracy led to a gradual increase in the number of revolutionists who aimed to overthrow the Qing government. As a result, clothing reform was seen as a direct way to protest the Qing government and was clad with anti-Manchu meaning. After the revolution, later revolutionaries such as Sun Yat-sen emphasized clothing reform’s national political implication—the fall of the Qing government and feudalism and the rise of the Republic of China and republicanism. Under such circumstance, adopting modern Republican clothing was, in a way, accepting and adopting Republican identity.

Prior to the revolution, revolutionaries and their supporters often saw queue-cutting as anti-Manchu. After the revolution, however, more people chose to make their arguments objecting to the unnecessary complexity of maintaining queues. This change in argument could be tentatively explained by the change of audience. Before the revolution, the idea of queue-cutting largely spread within the elites, literati, and a small group of commoners. These people wanted to find alliances that also supported overthrowing the Qing government. Consequently, revolutionaries stressed the anti-Manchu symbolism of queue-cutting. After the revolution, however, more people came to know what queue-cutting was either through propaganda or personal experience. As a result, queue-cutting advocates needed to frame their arguments in a more neutral way, namely, the inconvenience of traditional Chinese clothing, in order to persuade as many different people from the social spectrum to cut their hair as possible. The change of argument on hair-style demonstrates that the 1911 revolution helped popularize the idea of clothing reform in the public and, therefore, accelerated vestimentary revolution in the early twentieth century.

The diverse arguments on queue-cutting demonstrated that the success of the 1911 revolution had influenced the interpretation of clothing and clothing reform. After the revolution, clothing reform advocates refashioned their arguments on the reform from being political and revolution-based to secular and sartorial based. These revolutionaries and

supporters of the new republican regime donned modern garbs and, therefore, transformed themselves into pioneers of clothing reform in the new era. Along with the new dress code that was issued by the republican government and its subsequent queue-cutting teams, revolutionary arguments and actions—cutting queues, wearing western clothing themselves, and making speeches on queue-cutting in tea houses—influenced many commoners who were previously neither against nor supportive of the republican government. A mixture of Chinese Han and western style, modern garments and hair style were gradually popularized throughout the country, creating a new fashion trend. Many people, regardless of supporting the new government or not, followed the trend, dressing up in modern outfits and cutting off their queues. Consequently, the 1911 revolution accelerated the clothing reform that was first proposed by reformers in 1898 and expanded the reform's sphere of influence to the whole nation. In addition to the former ideas in 1898 on clothing reform, which consisted of the enhancement of military power, reduction of unnecessary garment complication, and protest over the Manchu dominion, the revolution gave new meanings to clothing reform, including republican citizenship, the end of national disgrace, and vitality of the transformed nation. Ostensibly, clothing reform was rather successful in transforming the public appearance. Nevertheless, caution should be taken in assessing effects that the revolution had on the public. As Peter Carroll suggested in *Refashioning Suzhou: Dress, Commodification, and Modernity*, at the time when sartorial and political fashion took place simultaneously, one could question whether “adopting Republican identity [was] as simple as donning a set of new clothes.”³³

Various ideas on clothing reform allow us to catch a glimpse of the turmoil at the turn of the twentieth century. Although there is no definite causality between ideas that reformers in 1898 held and the actual reform which took place in the early twentieth century, we can still perceive a relationship between these two—twentieth-century clothing reform as an implementation of ideas in 1898. Specifically, revolutionists in the twentieth century took 1898 reformers' ideas and modified them to fit in China's society in the twentieth century—the victory of the republic regime over feudalism, the rising notion of nationality, and the frequent, intense contacts with foreign power. They assigned to new hair and clothing styles a political connotation—republican identity. To a large extent, both late-nineteenth-century ideas on hair and clothing reform and the 1911 revolution influenced the way people defined these areas of appearance in the twentieth century.

Bibliography

- “Grave Signs in China.” *New York Times (1857-1922)*, May 18, 1910.
- “Queue Cutting in China.” *New York Times (1857-1922)*, Feb 26, 1911.
- Carroll, Peter. “Refashioning suzhou: Dress, commodification, and modernity.” *Positions* 11, no. 2, (2003): 443-478.
- Chan, Sin-wai. *Tan Sitong, A Book of Benevolence Translation*. Hong Kong: Chinese University Press, 1984.
- Chou, Eva Shan. “‘A Story about Hair’: A Curious Mirror of Lu Xun's Pre-Republican Years.” *The Journal of Asian Studies* 66, no. 2 (2007): 421-59.
- Finnane, Antonia. *Changing clothes in china: Fashion, history, nation*. New York: Columbia University Press, 2008.
- Godley, Michael R. “The end of the queue: hair as symbol in Chinese history.” *East Asian History*, no. 8 (December 1, 1994): 53-72.

³³ Peter Carroll, “Refashioning suzhou: Dress, commodification, and modernity,” *Positions* 11, no. 2, (2003): 445.

- Harrison, Henrietta. *Making of the Republican Citizen, The Political Ceremonies and Symbols in China, 1911-1929. Studies on Contemporary China*. Oxford: Oxford University Press, 2000.
- Harrison, Henrietta. *The Making of the Republican Citizen: Political Ceremonies and Symbols in China, 1911-1929*. Oxford: Oxford University Press, 1992.
- Hubei Academy of Social Science, ed., *Xinhai Geming zai Hubei shiliao xuanji* (Selected Historical Materials on the 1911 Revolution in Hubei). Wuhan: Hubei renmin chubanshe, 1981.
- Li, Zhigang (Lai, Chi-kung), “Xianxiang yu yingzao guozu: jindai Zhongguo de faxing wenti” (Imagining and constructing nationhood: hairstyles in Modern China), *Si yu yan* 36, 1 (1998): 99-118.
- Lo, Jung-pang, and Association for Asian Studies. *K'ang yu-wei: A biography and a symposium*. Vol. no. 23. Tucson: Published for the Association for Asian Studies by University of Arizona Press, 1967.
- Mei, Hua. *Chinese Modern Clothing History*. Beijing, China Textile & Apparel Press, 2008.
- Murthy, Viren. *The Political Philosophy of Zhang Taiyan: The Resistance of Consciousness*. Netherlands: BRILL, 2011.
- Ni, Liangduan. “Sun Yat-sen as provincial president of the Republic of China.” *Shiji Fengcai* 11, no.2 (October 2016): 26-29.
- Rhoads, Edward J. M., *Manchus and Han: Ethnic Relations and Political Power in Late Qing and Early Republican China, 1861–1928*. Seattle: University of Washington Press, 2006.
- Rhoads, Edward J. M., *Stepping Forth into the World*. Hong Kong: Hong Kong University Press, HKU, 2011.
- Robinson, Arthur G. “Pilgrims to Western seats of learning: China's first educational mission to the United States: the breaking of Chinese intellectual isolation.” *Chinese Studies In History* 36, no. 4 (June 1, 2003): 63-87.
- Sun, Yat-sen. *Kidnapped in London: being the story of my capture by, detention at, and release from the Chinese legation, London*. Bristol: J.W. Arrowsmith, 1897.
- Xiao, Gongquan. *A modern china and a new world: K'ang yu-wei, reformer and utopian, 1858-1927*. Vol. no. 25. Seattle: University of Washington Press, 1975.
- Yuan, Lu. “Xingbao and Queue-cutting Pronouncement.” *Nantong Weekly*, October 16, 2014, Section C, Final edition.
- Zhang, Shiyong, and Carissa Fletcher. “Subversive laughter: Carnival in the 1911 Revolution.” *Chinese Studies in History* 46, no. 1 (September 1, 2012): 30-70.



The Effect of Viewing Psychopathic Characters on Television on the Development of Psychopathic Traits on Children

Selin Baydar

Author Background: Selin Baydar grew up in Turkey and currently attends Hisar School in Istanbul, Turkey. Her Pioneer seminar topic was in the field of psychology and titled "Clinical Psychology."

Abstract

The personality traits and behaviors that define adult psychopathy begin to disclose themselves in childhood, so intervention in the early stages of the socially devastating disease is crucial. Numerous studies have shown the correlation between watching media violence and aggression, yet there is limited research on the effect of media violence on prognosis of psychopathy. This paper proposes a five-year long study correlating viewing of psychopathic television characters with psychopathy in youth. The participants consist of 400 13-year old children from 10 different juvenile delinquency centers. Moreover, the study includes measures such as The Hare Psychopathy Checklist: Screening Version, The Hare Psychopathy Checklist: Youth Version, the Dirty Dozen (a measure of the dark triad), a specifically designed Violence Scale and a subjective diary. Through the study, we aim to correlate psychopathic tendencies with watching television shows that contain an increased level of violence, watching violent television shows for a longer duration, and the level of identification that the children demonstrate with the psychopathic characters.

Keywords

Psychopathy, media violence, youth, the dark triad, priming, observational learning

Introduction

What Is Psychopathy?

Psychopathy is a complex personality disorder defined by a constellation of affective, interpersonal, and behavioral characteristics. It began to emerge as a formal clinical construct in the 20th century, yet there have been references to individuals who are evidently psychopaths in biblical, classical, medieval, and other historical sources (Penny and Moretti, 2010). Throughout the 19th and the 20th century, the illness went through etymological changes, but many of the evident traits remained unchanged. The current diagnostic criteria were established by the writings of the psychiatrist Hervey Cleckley in his classic book, *The Mask of Sanity* (1941). In his 40 years of clinical work, he came to narrow the syndrome he called psychopathy to 16 defining characteristics (Kiehl, 2006, Hare, 1996, Davison, Neale, Martin and Oltmanns, 2015).

There is no basis for holding one component of psychopathy to be more essential than other components (Hare and Neumann, 2008). Psychopaths are defined by features such as lacking a conscience and feelings for others, limited overall empathy, difficulty describing feelings, callousness, superficiality, short-temperedness, externally oriented thinking, egocentricity, impulsivity, irresponsibility, glibness, shallow affect, promiscuity and pathological lying. They use charm, manipulation, and violence to satisfy their own selfish needs, and they persistently violate social norms and expectations without the slightest sense of guilt or regret (Jonason and Krause 2013, Kiehl 2006). Additionally, psychopaths have lower than normal levels of skin conductance, are less autonomically reactive when they face stressful or aversive stimuli and have lower resting heart rates, which correlate with the clinical description of psychopaths' non-anxious characteristics and with research using other measures of emotion, demonstrating that people who test high on psychopathy are generally less emotionally reactive. Furthermore, while their lack of negative emotions, especially anxiety, may prevent them from learning from their mistakes, their lack of positive emotions causes them to behave irresponsibly towards others. They base their actions not on needs like money but on things that generate pure excitement in their lives. Therefore, they enjoy making grandiose life plans, but they fail to follow through with them (Kiehl 2006, Davison, Neale, Martin and Oltmanns, 2015).

Psychopaths have problems with processing certain aspects of speech and face stimuli, so they have difficulty with affective voice and facial expression identification tasks. Also, they show impairment on response reversal or extinction response inhibition, response modulation and making decisions (Kiehl, 2006).

Regarded as vicious and cold-blooded "intraspecies predators," they make up 15% to 25% of the United States' male and female prison population, 10% to 15% of substance abuse populations, and are responsible for a distinctly disproportionate amount of serious crime, violence and social distress in every society. Yet, they make up a small proportion, 1%, of the general population (Hare, 1996). This population, nevertheless, impacts virtually everyone at one time or another because psychopaths compose an important portion of persistent criminals, drug dealers, pimps, confidence artists, murderers, spouse and child abusers, swindlers, con men, mercenaries, terrorists, cult leaders and gang members. However, they are not limited to the dark side of our world, meaning not all psychopaths are criminals.

Psychopaths might become prominent in the business and corporate world, politics, and law, where the rules and their enforcement are lax and accountability is difficult to determine. It is especially common to observe psychopathy in business settings compared to the general population because the ability to manipulate others and lie about coworkers may lead to success in a competitive corporate environment as well as in "white collar" crimes. Many business executives, professors, politicians, physicians, plumbers, salespeople, and bartenders have antisocial personality traits like the criminal population (Davison, Neale, Martin and Oltmanns, 2015). It is not surprising for psychopaths to emerge as saviors of societies that are under social, economic or political upheaval (Hare 1996, Slater and Pozzato 2012). Due to its massive influence on society, psychopathy has no equal in terms of the amount of social, economic, physical and emotional distress generated by other disorders.

One of the essential components of its terrible nature is that psychopaths suffer little personal distress and seek treatment only when it provides them advantages such as attaining probation or parole; therefore, it is extremely difficult to treat the socially devastating disorder.

These immensely difficult-to-cure individuals are portrayed as vile, inhuman, and qualitatively different than what we call “normal” individuals in the media. However, psychopathic personality traits in adults and adolescents are best viewed as existing on a continuum. Therefore, personality disorders like psychopathy are dimensional in nature. Furthermore, several recent longitudinal studies have provided significant evidence that the psychopathic characteristic is moderately stable across development (Hare and Neumann 2008), which is why the correct approach at an early age is essential.

Models for Psychopathy

The Four-Factor Model

Psychopathy is a latent construct, so it is not directly observable. Therefore, it is crucial to compose a structure for psychopathy that takes its latent nature into account. A number of studies, based on latent variable analyses of the PCL instruments, support a four-factor model of psychopathy across diverse and primarily very large samples of male and female offenders, forensic and civil psychiatric patients, youth offenders, and individuals from the general community. The four factors that are highlighted in this model are interpersonal (e.g. pathological lying, manipulation, and deception), affective (e.g. shallow affect and lack of remorse), impulsive lifestyle (e.g. irresponsibility, constant stimulus seeking and impulsivity), and antisocial tendencies (e.g. poor behavioral controls) (Hare and Neumann 2008).

The Dark Triad

The dark triad of personalities includes Machiavellianism, narcissism, and psychopathy. The construct of Machiavellianism, which means a manipulative personality, emerged from Richard Christie’s selection of statements from Machiavelli’s original books. He organized those statements into a measure of personality by showing reliable differences in participants’ agreement with his items. He concluded that participants who agreed with his statements were more likely to behave coldly and manipulatively in their lives. Machiavellians are strategic and cynical. They solely satisfy their own needs with little regard for norms and morals (Paulhus and Williams, 2002).

Narcissists are egocentric individuals with grandiosity, entitlement, and dominance who perceive themselves as superior to others in numerous aspects of life, but still show signs of insecurity. The items regarding narcissism were refined on large samples of students and assembled in the Narcissistic Personality Inventory (Paulhus and Williams 2002).

Psychopaths lack empathy and often engage in thrill seeking behaviors with no regard for the consequences. Their central character elements are high impulsivity with low empathy and anxiety, as discussed above. Also, they tend to overestimate their intelligence. They self-enhance and are low on neuroticism (Christie and Geis 1970). The self-report psychopathy (SRP) scale was assembled by Hare (1985) to differentiate psychopaths from non-psychopaths (Jonason and Krause 2013).

Machiavellians and psychopaths are low on conscientiousness, have low levels of affective empathy and are correlated with higher levels of externally oriented thinking. Additionally, narcissists and psychopaths exhibit self-enhancement, high extraversion and high openness scores. The highest correlation between all three of the dark triad traits is, in fact, between psychopathy and narcissism, which reaches .61. Furthermore, all three traits need a socially malevolent character with behavioral tendencies toward self-promotion, emotional coldness, duplicity, and aggressiveness. Also, all of them manipulate their

physical appearance to achieve social gains and are associated with lower levels of cognitive empathy. Finally, men scored higher on all dark triad traits than women did (Paulhus and Williams 2002).

Evolution

According to life history theory, trait variances in the dark triad traits may be one of the ways to maximize the likelihood of the survival of the offspring. The dark triad and limited empathy could, indeed, be adaptive so long as they afford individuals greater reproductive success and access to resources. Recent research has concluded that individuals high on dark triad traits use different types of “cheater strategies” to achieve interpersonal and social goals despite their antisocial personalities. These maladaptive traits can, in fact, provide a competitive advantage by facilitating behavior associated with the attainment of goals that require exploitation of conspecifics. Also, an externally focused thinking style, which is a prominent trait of psychopathy, may reduce the capacity to recognize and attend to both one’s own and others’ emotional states simultaneously, which might be an advantage in their fast-paced life course (Jonason and Krause, 2013). In the end, these seemingly maladaptive traits have managed to exist within our gene pool (Paulhus and Williams 2002).

Secondary and Primary Psychopathy

Present conceptualizations of primary psychopathy are similar to the Cleckley (1941) original description of the psychopathic personality. Primary psychopaths are described as extremely self-centered, manipulative, egocentric and callous, whereas secondary psychopaths are overtly antisocial and impulsive (Ferrigan et al, 2000).

Primary psychopathy, which is known for demonstrating core affective traits, is motivated by reward (instrumental behavior) and may be the result of low levels of anxiety. Additionally, primary psychopaths tend to neglect information about the negative consequences of their antisocial behavior. On the other hand, secondary psychopathy, which is more behaviorally derived, is motivated by emotion (reactive behavior) and may be the result of negative affect and impulsivity (Coyne et al. 2010). Furthermore, secondary psychopathy is believed to lead to disproportionate attention to information about the positive consequences of antisocial behavior (Ferrigan et al., 2000).

There are some self-report measures for the assessment of primary and secondary psychopathy in non-incarnated individuals. Two of the significant measures are the Primary and Secondary Psychopathy scales (PSP; Levenson et al., 1995), which were derived to operationalize the two-factor framework. These scales have shown adequate internal consistency coefficients and patterns of correlations with measures of other constructs that are concurrent with the two-factor framework (Ferrigan et al., 2000).

Childhood and Psychopathy

The personality traits and behaviors that define adult psychopathy begin to manifest themselves in childhood, so it is essential to intervene early if we hope to influence the development and behavioral expression of the mental illness. However, the prevention is complex due to the general failure to differentiate the budding psychopath from other children who display serious emotional and behavioral problems, especially those diagnosed with conduct disorder, attention deficit hyperactivity disorder or oppositional defiant disorder (Hare 1996). Nevertheless, it is important to note that adult antisocials have a history of behavior problems and conduct disorder during childhood. Therefore, these

disorders are essential in the identification of the disease (Davison, Neale, Martin and Oltmanns, 2015).

Environmental factors are important in the onset and prognosis of psychopathy. Children who grow up in physically abusive or neglectful households and have authoritarian parents, who are characterized by a mixture of restrictiveness with low warmth, are at increased risk for antisocial behavior in adolescence and adulthood (Davison, Neale, Martin and Oltmanns, 2015). Furthermore, mothers' antisocial behavior is linked to adolescents' antisocial behavior through the teens' awareness of their mother's actions such as drinking too much, breaking the law, having many conflicts with others, lying, and engaging in reckless actions, so the mothers serve as a model to imitate. Therefore, modeling is an important aspect of psychopathy in adolescents. Additionally, social factors play a vital role in the prognosis of the disease. Daily occurrences like school failure, peer rejection and identification with a deviant peer group are important predictors of adult psychopathy. Individuals developing in environments where levels of harshness or unpredictability vary can maximize fitness by adopting distinct behavioral patterns, like alternate Life History strategies. This LH theory predicts that lineages evolving and individuals developing in high-stress environments, which are the case in most psychopathic individuals, become faster LH strategists, devoting greater energy toward mating effort. Indeed, dark triad traits are important correlates of mating effort since they predict a higher number of sexual partners (Patch and Figueredo, 2017).

What Causes Psychopathy?

Biological Causes

There is evidence that broad genetic factors may account for a substantial portion of the variance and covariance of various sets of psychopathic traits. For example, there are genetic influences on the covariance of psychopathy scales reflecting emotional detachment and antisocial tendencies. Another example is Baker's (2007) study on a large sample of 9 to 10-year-old twins. He found a common antisocial behavior factor (composed of child psychopathy traits, aggression, and delinquency) that was highly heritable. Furthermore, Viding et al. (2007) discovered a common genetic component to the covariation between callous unemotional traits and antisocial tendencies in children. Moreover, some clinical and behavioral features of psychopathy such as impulsivity, poor response inhibition and difficulty in processing emotional material, are mirrored in brain function and brain structure (Hare and Neumann 2008).

Although early investigations regarding the biological components of psychopathy implicated relatively localized brain regions, like the amygdala, the hippocampus and the frontal cortex, current research concluded that psychopathy can be better understood in terms of complex interactions among various regions and functions (Hare and Neumann 2008). Some of these regions are the orbital frontal cortex, the insula, the anterior and posterior cingulate, the parahippocampal gyrus, and the anterior superior temporal gyrus. Therefore, the relevant functional neuroanatomy of psychopathy includes the limbic and paralimbic structures, which make up the paralimbic system.

Environmental Causes

Even though it is evident that biological factors play a significant role in the development of psychopathy, family factors, and unique environmental factors also play a significant part in the prognosis of psychopathy in individuals (Hare and Neumann, 2008). The environment that an individual is raised in is highly influential on the development of

that individual. Unpredictability in early life environment, thus the development of psychopathy, can depend on multiple factors, such as physical, sexual, or verbal abuse, physical or emotional neglect, growing up with a mentally ill or chemically-dependent parental figure, parental incarceration, witnessing or experiencing domestic violence, growing up with a low socioeconomic status or crime-ridden neighborhood, and malnutrition (Patch and Figueredo, 2017).

Cognitive Causes

Numerous social animal species have been known to inhibit violent attacks when a conspecific demonstrates submission cues. Blair (1995) has urged that humans possess a functionally similar mechanism that mediates the suppression of aggression when distress cues are displayed. He proposed that this is a violence inhibition mechanism (VIM). He has suggested that this mechanism is important for the development of the consistently observed distinction in an individual's judgments between moral and conventional transgressions (moral/conventional distinction). According to Blair (1995), psychopaths may lack this violence prevention.

Blair has considered VIM to be a cognitive mechanism that, when activated by a non-verbal communication of suffering, initiates a withdrawal response, which causes the individual to withdraw from the attack. Likewise, Camras (1977) has perceived that when a 4-to-7-year-old child displays a sad facial expression while resisting another child's attempt to take his or her possessions, the aggressor child terminates his or her attempts. However, this situation is not observed in individuals with psychopathic tendencies since psychopaths demonstrate an early commencement of extremely violent behavior that is not tempered by any guilt or empathy with the victim (Blair, 1995).

The Impact of Media on Psychopathy

Media has become a prominent part of our everyday life. Humanity cannot separate itself from the benefits of this technology, yet there are numerous disadvantages to our close association with media. Since media has a massive grasp over our lives, it is impossible to disregard the extensive impact of this growing culture. A significant impact of media is that on aggressive behavior. Previous research has suggested that viewing aggressive behavior in the media, such as violent television and films, video games, and music, can increase aggressive thoughts and behavior, both in immediate and long-term contexts (Anderson et al., 2003), though certain characteristics of viewers, social environment and media content can influence the degree to which the content affects a person's behavior. The evidence for this finding is the clearest within the most extensively researched domain of media, which is television and film violence.

It is important to note that the impact of viewing violence on milder forms of aggression is greater when compared to its effects on more severe forms of aggression. Nevertheless, the effects on severe forms of violence are also substantial when contrasted with the impact of other violence risk factors or even medical effects that are branded as significant by the medical community, such as the effect of aspirin on heart attacks. Therefore, viewing violence in any form is a significant risk factor for individuals (Anderson et al., 2003).

Moreover, the General Aggression Model that was created by Anderson and Bushman (2002) states that personality can influence the effect of media violence. Firstly, long-term exposure to media violence can shape an individual's personality, especially the aspects that are related to violent behavior. Secondly, specific personality traits appear to mediate the effect of viewing media violence on subsequent behavior in the short-term

context (Coyne et al., 2010). Additionally, media violence affects an individual's behavior and thoughts both through short-term and long-term exposure, yet the impacts of these exposures are dissimilar. Short-term exposure increases the possibility of physically and verbally aggressive behavior, aggressive thoughts and emotions, whereas long-term exposure in childhood causes the individual to form an association with aggressive behavior later in life (Anderson et al., 2003).

Watching aggressive images in the media can cause both long-term and short-term effects. Firstly, exposure to the violent media causes short-term consequences by priming existing aggressive scripts, increasing physiological arousal, and triggering an autonomic tendency of learning observed behaviors, which is called modeling. Secondly, an individual's exposure to violent images in the media causes long-term damages by numerous learning processes. These processes lead to lasting aggressive scripts, interpretational schemas, and beliefs that support aggression regarding social behavior and to desensitization, which reduces an individual's normal negative emotional responses to violent behavior (Anderson et al., 2003).

Since the most prominent effects of media violence are on youth, it is significant to note that a developmental perspective is essential. Some youth who have a tendency for aggressive behavior and engage in some forms of antisocial behavior do not grow up to be violent teens and adults. However, a significant proportion of aggressive children, indeed, go on to become aggressive adults. Anderson, Berkowitz, Donnerstein, Huesmann, Johnson, Linz, Malamuth, and Wartella (2003, p.83) highlight this point by stating "the best single predictor of violent behavior in older adolescents and young adults is aggressive behavior when they were younger." Consequently, childhood and early adolescence is a significant point for prevention efforts. It has been noted that parental supervision, interpretation, and control of children's media use has been effective, yet research has suggested that no one is completely immune to the massive grasp of media violence (Anderson et al., 2003).

Previous Studies

There have been previous studies regarding the effect of media violence on aggression in the past. However, each of them had significant limitations as well as reputable critiques. Therefore, it is important to note that each of these studies is susceptible to doubt and further research. Initially, Bjorqvist (1985) exposed 5-to-6-year-old Finnish children to either violent or nonviolent movies. Then in order to prevent experimenter's bias, the two raters that were observing the children in the room did not know which type of film the children had observed. They subsequently observed that youngsters who had viewed the violent movie were much higher on physical assault, hitting and wrestling other children, and on other types of demonstrations of aggression. The results for physical assault for children who had watched the violent movie were highly significant ($p < .001$) and the effect size was substantial ($r = .36$). Even though Bjorqvist obtained important results, he neglected the importance of a longitudinal study and observed the short-term effects of watching media violence.

Josephson (1987) conducted an experiment in which he randomly assigned 396 7-to-9-year-old boys to watch either a violent or a nonviolent film before they participated in a floor hockey game at school. Similar to Bjorqvist, he made sure that the raters did not know which boy watched which movie. They recorded the number of physical attacks, which he defined as hitting, elbowing, shoving another player to the floor, tripping, kneeling, pulling hair, and other assaultive actions that would result in a penalty, demonstrated by each boy. An important part of the study was the presence of a specific cue, which had appeared in the violent movie, throughout the game. This particular cue supposedly reminded the boys of

the movie they had watched previously. Josephson concluded that for aggressive boys who scored above average on a measure of aggressiveness, the mixture of watching the violent film and seeing the specific cue stimulated the most assaultive behavior ($p < .05$). This study highlighted the importance of the priming effect through the use of specific cues, yet the study lacked a longitudinal approach and a limited sample since it was made up solely of boys.

In a home for delinquent boys in Belgium, Leyens, Camino, Parke, and Berkowitz (1975) assigned boys in two cottages to watch violent movies every night for five nights while the boys in the other two cottages watched nonviolent movies. Later on, Leyens and his colleagues observed the boys each evening and rated them for their frequency of hitting, choking, slapping, and kicking other boys. The boys that were exposed to violent movies demonstrated significantly more physical assaults ($p < .025$) on other residents of their cottage. It is also important to note that there was a larger effect for the boys who were initially more aggressive compared to the boys who were initially less violent. This study is important because it took into account the initial personalities of the boys, showing that physically excited individuals are more apt to be aggressively stimulated, as well as demonstrating that violent films can produce serious physical aggression even in settings where violent behavior is strictly against the officially prescribed rules. Nevertheless, similar to Josephson's experiment, this study lacked a longitudinal approach and had a limited sample.

Paik and Comstock (1994) conducted a meta-analysis and examined cross-sectional surveys published between 1957 and 1990. For the 410 tests that they analyzed, they determined that viewing television violence is, indeed, positively correlated with aggressive behavior ($r = .19$). These surveys also supported the causal conclusions of the previous experimental studies, and indicate that the short-term effects observed in the laboratory can be generalized to long-term effects in the real world. Nevertheless, these surveys alone cannot indicate whether media violence is the cause of aggression, whether aggressive youths are especially attracted to media violence, or whether some other factor conditions the same youths to both watch more violence and behave in an aggressive manner compared to their peers. However, longitudinal studies have the opportunity to shed light upon these matters.

There have, in fact, been longitudinal studies conducted in the past. One of them was that of Milavsky, Kessler, Stipp, and Rubens (1982). Milavsky and his colleagues examined the effects of television violence on aggression in boys and girls aged 7 to 16 from two Midwestern cities. He used physical aggression and delinquency as a measure and surveyed the participants up to five times during a 3-year period. Within each time of the assessment, cross sectional correlations were obtained between viewing of television violence and concurrent levels of aggression (Anderson et al., 2003). The results were significant (.13 to .23 for boys and .21 to .37 for girls). Even though this study had a longitudinal approach and a diverse sample, it had confounding variables and focused on a broad term, aggression.

Another longitudinal study regarding the impact of TV violence was that of Huesmann and Eron (1986). Boys and girls aged 6 to 8 were examined three times in a 3-year period. Aggression level was assessed by peer nominations in response to questions regarding physical and verbal behaviors. Concurrent with Milavsky's study, the cross-sectional correlations of Huesmann's study demonstrated a positive correlation between aggression and exposure to violent TV. An important aspect of this study was that Huesmann re-interviewed the participants fifteen years after the onset of the study. The results suggested a delayed effect of media violence on serious physical violence. The

researchers found a significant correlation between TV violence in childhood and aggression during young adulthood for both women ($r = .19$) and men ($r = .21$). However, it was concluded that high aggression during childhood does not lead to frequent viewing of violent TV in adulthood. This study was, indeed, a highly effective and generalizable study, yet this study, similar to Milavsky's study, focused on a broad term in psychopathology, aggression.

Anderson and Bushman (2002) conducted a significant meta-analysis regarding the effects of media violence on aggression through a longitudinal study. Even though this analysis combined studies concerning all types of media violence, the vast majority of the research concerned violent television. They found a significant average effect size of .17 through 5000 participants, which underscores that high levels of exposure to violent television in childhood, indeed, promotes aggressive thoughts and behavior in adolescence and young adulthood. However, it is important to note that there are numerous critiques stating that media violence cannot affect youth and strongly oppose the views of Anderson and Bushman (Anderson et al., 2003). With this conclusion, it is significant to discuss the processes that contribute to the formation of an aggressive thought pattern and behavior subsequent to an exposure to violent media.

Observational Learning and Imitation

Observational learning and imitation affect both the long-term and the short-term effects of violent media. Humans begin imitating others at a very early age, and their observation is the likely source of many children's motor and social skills (Bandura, 1977). Humans have specific neurological systems designed for imitation that enable youngsters to attain rudimentary social behaviors. Social interactions that children have enhance the behaviors that they first acquire through observation, yet observational learning remains a significant mechanism for learning new social behaviors and the environments that are most suited to apply these behaviors throughout childhood and maturity. Furthermore, as children mature, their beliefs and attitudes are developed from the inferences they make about the social interactions that they observe. Children can learn from whomever they observe, including parents, siblings, peers or even television characters. Observational learning can, indeed, contribute to the effects of media, yet much of this learning takes place unintentionally and the children have no awareness that the learning has occurred (Anderson et al., 2003).

According to observational learning theory, there are various factors in modeling a figure. First of all, projection is heightened when the figure that is performing the observed behavior is similar to or attractive to the viewer. Similarly, when the viewers identify with the figure and feel a connection, modeling increases. Thirdly, when the context is realistic and applicable, the viewer has an easier time relating to the figure. Finally, when the behavior that is observed is followed by a positive consequence or the consequences of the negative action is disregarded, modeling increases (Bandura, 1977).

Observational learning can be used to explain the short-term effects of media violence, but the long-term effects are slightly more complicated. The consequences that the imitated behavior brings are highly significant in respect to long-term effects. Furthermore, children not only learn specific behaviors from models, but they also learn complex social scripts, which are clusters of rules that aid them in interpreting social interactions (Anderson et al., 2003). These scripts assume the role of a cognitive guideline for children and direct their behaviors. An example that Anderson and Huesmann (2003, p.95) gave was that "children may learn that aggression can be used to try to solve interpersonal conflicts. As a result of mental rehearsal (e.g. imagining this kind of behavior) and repeated exposure, this

approach to conflict can become well established and easily retrieved from memory.” This statement highlights the importance of another factor when it comes to the development of aggressive behaviors due to media violence, which is priming.

Priming

The human mind acts as an associative web in which ideas are constantly activated--primed--by associated stimuli in the surrounding environment. Observing an event or a specific stimulus can prime related concepts in a person’s memory without the person being aware of an influence. Exposure to violent scenes can prime a set of aggressive ideas and emotions and increase the accessibility of aggressive thoughts and scripts (Anderson et al., 2003). Frequently primed aggression-related emotions, thoughts, and scripts become automatically accessible to the individual. Therefore, these thoughts and scripts become a part of the normal mindset of the individual. Furthermore, the easy path towards these thoughts increases the probability of aggressive encounters in the individual’s lifespan, which may be the underlying reason behind the long-term effects of media violence. Moreover, Bushman (1995) has noted that high trait aggressive individuals are more susceptible to the effect of priming compared to low trait aggressive individuals (Ferrigan et al., 2000).

Emotional Desensitization

Emotional desensitization means that an individual demonstrates a reduced distress-related physiological reactivity when observing or thinking about violence. Simply, individuals who respond with less unpleasant physiological arousal, or fewer negative emotional reactions than they originally did are going through emotional desensitization. Since negative emotional reactions towards violence have an inhibitory influence on thought about violent behavior, or behaving violently, emotional desensitization may result in an elevated possibility of violent behaviors and thoughts. Moreover, even brief exposures to media violence can reduce a negative response to the sight of real-world violence and can decrease helpful attempts towards victims of aggressive behavior (Anderson et al., 2003). Therefore, emotional desensitization is an important factor regarding the effects of media violence.

Viewer Characteristics

The context and model are not the sole components when it comes to media influence. The characteristics of the viewer play a significant part.

Firstly, younger children, whose social scripts, schemas, and thoughts are less solidified than older children’s, are more sensitive to being influenced. Indeed, observational theory indicates that the viewer’s age and gender is significant when it comes to identifying with the portrayed violent characters, which increases enactment of the observed behavior.

Secondly, as discussed above, individuals with high trait aggression may have a lower threshold for media-violence induced aggression compared to less aggressive individuals.

Thirdly, children who identify strongly with an aggressive character or perceive violent scenes in television as realistic are specifically likely to have violent thoughts primed by the observed aggression to imitate the character in real life, or to attain numerous aggressive scripts. It is significant to point out that identification depends on the portrayal of characters as well as the viewer. Consequently, realistic depictions are more likely to heighten the viewer’s aggressive behaviors compared to more fantastic or fictionalized

characters (Anderson et al., 2003). Moreover, as Anderson (2003) stated regarding the ideology of Leyens and Picus (1973), “when people are asked to imagine themselves as the protagonist in a violent film, the effects of viewing the film are enhanced, perhaps because of the viewers’ relatively greater psychological involvement.” Furthermore, the viewers are apt to identify with aggressive characters that are similar to themselves in age, gender or race. Nevertheless, the attractiveness, power and charisma of the character may be more powerful over the viewers.

Finally, justifying violent behavior causes the viewers to believe that their own violent acts are also appropriate, so they are more prone to behave violently. Yet, violence does not need to be explicitly rewarded in order to heighten the possibility of aggressive behavior. Seeing the character get away with their aggressive behavior can also enhance learning of aggressive behavior and thoughts (Anderson et al., 2003).

As shown above, there have been various studies regarding the effect of watching violent media on aggression in individuals. These studies included correlation studies, experiments and longitudinal studies. They also included a wide range of samples and methods. However, there is a very limited number of studies regarding the effect of media, specifically television shows, on psychopathy. After searching through recent literature regarding this topic, I concluded that most of the studies are not longitudinal and are experiment-based, which leads to a question of generalizability. Therefore, I suggest a 5-year correlational study that focuses on the impact of viewing psychopathic characters in television shows on children aged 13. Within this study I plan to find various relationships between violent television and psychopathy. There are three hypotheses that I think this study will verify.

1. Children who watch television shows that contain an increased level of violence demonstrate more psychopathic traits in adulthood than children who watch television shows with a lesser level of violence.
2. Children who watch more hours of violent television shows per week show more psychopathic traits in their adulthood whereas children who do not watch a considerable amount of violent television per week show less of these traits.
3. Children who identify with the television characters due to similarity, attractiveness, charisma or familiarity demonstrate more psychopathic traits than children who do not identify with the television characters, who show fewer psychopathic traits.

Methods

Participants

I will recruit approximately 400 boys and girls aged 13 from 10 different juvenile delinquency centers around the US. The sample size will be kept large due to the possibility of dropouts. Furthermore, the children will be specifically chosen to ensure that they have a conduct disorder, a parent with a criminal background or an ASPD (antisocial personality disorder), or a criminal history themselves. I will be able to specify my sample since my study is correlational, which allows me flexibility. Furthermore, I will pay the children and their parents 10 dollars for their contributions every six months. Also, all of the children will be 13 since the study will take five years to complete and age will not be a confounding variable in the study.

Measures

PCL: YV

In order to determine the level of psychopathy of the children, Hare's Psychopathy Checklist: Youth Version (PCL: YV) will be administered. This version of the PCL was specifically designed to be more responsive to the roles and situations that characterize adolescents. Studies regarding the PCL: YV have shown that high-scoring individuals display more violent criminal behavior, earlier commencement of antisocial behaviors, more conduct disorder symptoms, and greater tendency towards substance abuse, compared to low-scoring individuals (Neumann et al., 2002).

The PCL: YV was designed to measure the same 20 items as the PCL-R. Nevertheless, the titles and descriptions of the dispositions assessed, references, and scoring criteria for most items were altered. Items are scored as 0 (consistently absent), 1 (inconsistent), or 2 (consistently present). These judgments require the integration of information provided by self-report measures, collateral sources (such as parents or teachers), and direct observation of behavior (Neumann et al., 2006, Neumann et al., 2002).

Neumann, Kosson, Cyterski, Steuerwald and Walker-Matthews (2002) conducted a study to test the reliability and validity of this measure. Based on the study there was a high internal consistency, reliability, of PCL: YV scores, which are indicated by alpha coefficients for averaged rating of .88. Also, the alpha coefficient for interviewers and observers was .79. Therefore, reliability of the measurement was high. As for validity, the pattern of correlations calculated based on all 20 PCL: YV items was equal to that reported for corrected PCL: YV scores in all cases. PCL: SV is strong in convergent, predictive and discriminant validity. Furthermore, this measure not only predicts antisocial behavior, but it also predicted other indicators of psychopathology, including all three DSM-III-R disruptive behavior disorders of childhood, and a lack of attachment when it came to their relationships with parents (Neumann et al., 2002). Therefore, this measure is extremely convenient when it comes to conducting a thorough longitudinal study with various aspects to consider.

PCL: SV

I will be using the Psychopathy Checklist: The Screening Version in my study. Since the PCL-R takes several hours to complete, an alternate, more practical, measure is necessary for my longitudinal study. The PCL: SV is theoretically and empirically correlated to the PCL-R and can be used for the assessment of psychopathy both in forensic and noncriminal populations. It has the same two-factor structure as the PCL-R. It measures the affective and socially aberrant aspects of psychopathy by six items, with a total of 12 items (Hare, 1996). Furthermore, previous validation studies have revealed a great correspondence between the PCL: SV and PCL-R, with $r = .80$ (Guy and Douglas, 2006).

The PCL: SV has high correspondence with external outcomes like drug and alcohol abuse, treatment noncompliance, and ASPD (Guy and Douglas, 2006). Much like the PCL: YV, the PCL: SV is an effective alternative for the commonly used PCL-R, and it is capable of assessing psychopathy in individuals successfully.

The Dirty Dozen

In addition to the PCL: SV and the PCL: YV, I will be administering the Dirty Dozen, which is used to measure the dark triad traits. The Dirty Dozen uses only four items per construct to measure Machiavellianism, Narcissism, and Psychopathy. During the longitudinal study, the participants will be asked how much they concur, from 1= strongly

disagree to 5=strongly agree, with a statement like “I tend to act remorsefully” (i.e. psychopathy). Then, the items will be averaged together (Jonason and Krause, 2013, Miller et al., 2012). This measure will add a not widely explored element to the traditional studies of violent media on psychopathy. Therefore, the Dirty Dozen will play a significant part in the study.

The Diary

The participants of the study will have a custom electronic diary in which they will rate the episodes of the TV shows that they watch throughout the day. The diary will be administered through a custom-designed webpage in order to prevent delays, which may occur in the case of a traditional hardcover diary. Moreover, this diary will be based upon the 5-point Likert scale (1=disagree a lot, 5=agree a lot), and the participants will be required to rate every episode they watch. They will be rating statements such as, “I identified with the main character,” “I saw physical and verbal aggression,” “I saw a plentitude of blood,” “I thought about this episode after I finished watching it,” “I thought that the main character was attractive/charismatic,” or “I think that the events are realistic.” Furthermore, the participants will be required to write down the episodes that they most vividly remember in their diary each Sunday night.

The Violence Scale

In order to have an objective rating of the impact of the television shows on the participants, I will use a Violence Scale. A team of 20 people will go through a two-week program, in which they will learn about the violence scale. These 20 individuals will code each episode that the participants will be noting down on their diary, which will be easily accessible for them through the Internet. First of all, the coders will note the duration of the episodes. They will also code every episode according to the amount of physical or verbal aggression, guns and harmful devices, and blood. They will also code for the amount of fictional and fantastic elements of the episode, and the amount of homicide, suicide, and genocide portrayed in the episode. Finally, the coders will rate the charisma and approachability of the main psychopathic character in the series.

Procedure

The participants will be recruited from 15 different juvenile delinquency centers around the U.S. All of the participants will be aged 13 in order to eliminate the effects of age on the development of psychopathic traits. Furthermore, each all participant and their parents or guardian will be informed about the longitude, requirements and the materials within the study. As I have noted above, the participants and their parents or guardians will be paid 10 dollars for their contribution every 6 months. Also, a consent form will be attained from the delinquency centers.

Once the participants are chosen, an initial PCL: YV will be administered. Interview questions for the participants will be written based on interviews used to assess psychopathy in adults (i.e. PCL-R). Moreover, there will be interviews with parents, guardians, or other members of the same delinquency center as a collateral source. Ten different individuals will serve as interviewers; these interviewers will undergo an extensive 2-week PCL: YV training and will be required to demonstrate good interrater agreement for practice ratings on the participants before they are permitted to rate the official scores. Also, there will be in-room observers, which will also undergo the PCL: YV training, in order to prevent any experimenter bias (Kosson et al., 2002). The PCL: YV will take 90-to-120 minutes and will be handscored by the interviewers. Moreover, PCL: YV will be

administered at the end of each year for 5 years, not including the initial assessment administration.

Besides the annual PCL: YV, the Dirty Dozen and the PCL: SV will be administered in order to have a continuous pattern on the development of psychopathic traits on the participants. The Dirty Dozen and PCL: SV will be administered to participants sequentially. The administrators will be trained prior to the assessments. The questionnaires will be administered every 6 months for 5 years. Therefore, the psychopathic traits will be observed without an excessive time needed for assessment, and dark triad traits will be measured as an additional aspect to the study.

Besides the scheduled measures, which are PCL: SV, PCL: YV and the Dirty Dozen, the participants will be required to fill out an online diary every day on a website that I will design. In the online diary, they will rate every episode that they watch throughout the day. This diary will only be accessible to them and the raters, and the participant's parents or guardians will not be permitted to view the ratings in order to secure confidentiality.

The Violence Scale will be administered for every episode that the participants submit in order to have a comprehensive study. The coders will be required to check the diary entries of the participants every 3 days. Moreover, the participants will not be informed about the results of the Violence Scale, regarding the episodes that they watch.

Data Analysis

In order to test my hypotheses, I will be using a multiple regression analysis since my study will be a correlational study. In order to eliminate the effect of some participants having higher PCL: YV scores at the onset of the study, my model will be $PCL: YV - End = PCL: YV - Start + \text{Amount of violence perceived for my first hypothesis}$. Therefore, while testing my first hypothesis, I will be predicting PCL: YV at the end of the study from violence perceived while also controlling the PCL: YV scores at the beginning of the study. For my second hypothesis, I will again be using a multiple regression analysis. My model will be similar to my first model except my independent variable will be projective identification with the television character modeled. Finally, for my third hypothesis, I will be using a multiple regression analysis and a similar model to the previous models with the exception of the independent variable. The independent variable in my final model will be the number of hours watching television per week.

References

- Anderson, C. A., & Bushman, B. J. (2002). Human aggression. *Annual Review of Psychology*, 53, 27–51.
- Anderson, C.A., & Bushman, B.J. (2002c). Media violence and the American public revisited. *American Psychologist*, 57, 448–450.
- Anderson, C. A., Berkowitz, L., Donnerstein, E., Huesmann, R. L., Johnson, J. D., Linz, D., Wartella, E. (2003). The Influence of Media Violence on Youth. *Psychological Science in the Public Interest*, 4(3), 81-110.
- Baker LA, Jacobson KC, Raine A, Lozano DI, Bezdjian S. 2007. Genetic and environmental bases of childhood antisocial behavior: a multi-informant twin study. *J. Abnorm. Psychol.* 116, 219–35
- Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice Hall.
- Bjorkqvist, K. (1985). *Violent films, anxiety, and aggression*. Helsinki: Finnish Society of Sciences and Letters.

- Blair, R. (1995). A cognitive developmental approach to morality: investigating the psychopath. *Cognition*, 57(1), 1-29.
- Bushman, B.J. (1995). Moderating role of trait aggressiveness in the effects of violent media on aggression. *Journal of Personality and Social Psychology*, 69, 950-960
- Camras, L.A. (1977). Facial expressions used by children in a conflict situation. *Child Development*, 48, 1431-1435.
- Christie, R., & Geis, F. L. (1970). *Studies in Machiavellianism*. New York: Academic Press
- Cleckley, H. *The Mask of Sanity*. St. Louis, MO: Mosby 1941.
- Coyne, S. M., Graham-Kevan, N., Grant, D. M., Keister, E., & Nelson, D. A. (2010). Mean on the screen: Psychopathy, relationship aggression, and aggression in the media. *Personality and Individual Differences*, 48(3), 288-293.
- Ferrigan, M. M., Berman, M. E., & Valentiner, D. R. (2000). Psychopathy dimensions and awareness of negative and positive consequences of aggressive behavior in a nonforensic sample. *Personality and Individual Differences*, 28(3), 527-538.
- Figueredo, A. J., & Patch, E. A. (2017). Childhood stress, life history, psychopathy, and sociosexuality. *Personality and Individual Differences*, 115, 108-113.
- Guy, L. S., & Douglas, K. S. (2006). Examining the utility of the PCL:SV as a screening measure using competing factor models of psychopathy. *Psychological Assessment*, 18(2), 225-230.
- Hare, R. D. (1985). Comparison of procedures for the assessment of psychopathy. *Journal of Consulting and Clinical Psychology*, 53, 7-16.
- Hare, R. D. (1996). Psychopathy. *Criminal Justice and Behavior*, 23(1), 25-54.
- Hare, R. D., & Neumann, C. S. (2008). Psychopathy as a Clinical and Empirical Construct. *Annual review of clinical psychology*, 4(1), 217-246.
- Huesmann, L.R., & Eron, L.D. (Eds.). (1986). *Television and the aggressive child: A cross-national comparison*.
- Jonason, P. K., & Krause, L. (2013). The emotional deficits associated with the Dark Triad traits: Cognitive empathy, affective empathy, and alexithymia. *Personality and Individual Differences*, 55(5), 532-537.
- Josephson, W.L. (1987). Television violence and children's aggression: Testing the priming, social script, and disinhibition predictions. *Journal of Personality and Social Psychology*, 53, 882-890.
- Kiehl, K. A. (2006). A cognitive neuroscience perspective on psychopathy: Evidence for paralimbic system dysfunction. *Psychiatry Research*, 142(2-3), 107-128.
- Kosson, D. S., Cyterski, T. D., Steuerwald, B. L., & Neumann, C. S. (2002). The Reliability and Validity of the Psychopathy Checklist: Youth Version (PCL: YV) in Nonincarcerated Adolescent Males. *Psychological Assessment*, 14(1), 97.
- Levenson, M. R., Kiehl, K. A., & Fitzpatrick, C. M. (1995). Assessing the psychopathic personality. *Journal of Personality and Social Psychology*, 68, 151-158.
- Leyens, J.P., & Picus, S. (1973). Identification with the winner of a fight and name mediation: Their differential effects upon subsequent aggressive behavior. *British Journal of Social and Clinical Psychology*, 12, 374- 377.
- Leyens, J.P., Camino, L., Parke, R.D., & Berkowitz, L. (1975). Effects of movie violence on aggression in a field setting as a function of group dominance and cohesion. *Journal of Personality and Social Psychology*, 32, 346-360
- Milavsky, J.R., Kessler, R., Stipp, H., & Rubens, W.S. (1982). Television and aggression: Results of a panel study. In D. Pearl, L. Bouthilet, & J. Lazar (Eds.), *Television and behavior: Ten years of scientific progress and implications for the eighties*, 2, 138-157

- Miller, J. D., Lynam, D. R., Few, L. R., Seilbert, L. A., Watts, A., & Zeichner, A. (2012). An Examination of the Dirty Dozen Measure of Psychopathy: A Cautionary Tale About the Costs of Brief Measures. *Psychological Assessment*, 24(4), 1048-1053.
- Moretti, M. M., & Penney, S. R. (2010). The Roles of Affect Dysregulation and Deficient Affect in Youth Violence. *Criminal Justice and Behavior*, 37(6), 709-731.
- Neumann, C. S., Kosson, D. S., Forth, A. E., & Hare, R. D. (2006). Factor structure of the Hare Psychopathy Checklist: Youth Version (PCL: YV) in incarcerated adolescents. *Psychological Assessment*, 18(2), 142-154.
- Oltmanns, T. F., Martin, M. T., Neale, J. M., & Davison, G. C. (2015). *Case studies in abnormal psychology*. Hoboken, NJ: Wiley.
- Paik, H., & Comstock, G. (1994). The effects of television violence on antisocial behavior: A meta-analysis. *Communication Research*, 21, 516-546
- Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality*, 36(6), 556-563.
- Slater, T., & Pozzato, L. (2012). Focus on Psychopathy. *FBI Law Enforcement Bulletin*, 81(7), 1-1.
- Viding E, Frick PJ, Plomin R. 2007. Aetiology of the relationship between callous-unemotional traits and conduct problems in childhood. *Br. J. Psychiatry* 190(Suppl. 49):33-38



The Impact of Emotions, Ambiguity, and Apathy on Decision-Making in Individuals with Alzheimer's Disease: A Proposal to Facilitate the Diagnosis of Alzheimer's Disease

Sammer M. Marzouk

Author Background: Sammer M. Marzouk grew up in Morocco and currently attends The University of Chicago Laboratory Schools in Chicago, Illinois. His Pioneer seminar topic was in the field of neuroscience and titled "The Decision-Making Brain."

Abstract

Alzheimer's Disease (AD) is the fastest-growing cause of dementia and mental illness for elders in the world (Alzheimer's, 2015). Because of this, the rush to diagnosis AD is leading to an increasing number of false positives. One-fifth of AD diagnosis are false positives. In the U.S, 18.2 billion hours of care and 230 billion dollars are given to the AD diagnosis process (Alzheimer's, 2015). In order to combat this, this five-part proposal (known as the Cumulative Diagnostic Test of the CDT) intends to change the current diagnosis method by creating indexes that only allow the most likely cases of AD to continue with further testing. This proposal tests the emotions, apathy, and ambiguity. These variables were chosen as they have the greatest influence over the decision-making process (Sinz et al., 2008). In this proposal, 1056 data points (n=1056) from 272 healthy elderly people (control, n=272) and 784 elderly people with DAT (experimental, n=784) were analyzed and scored in the CDT, with the average score of the experimental group being 6.10/11, and the average score of the control group being 2.4/11 (Appendix). These scores are positively correlated with risk for developing AD. With this proposal, more than 46 billion dollars of AD funding can be diverted from false positives into more advanced testing. And the U.S. as a whole could save about 3.72 billion hours from the diagnosis process.

Primer on The Brain, Memory, Decision-Making, Neurons, and Neurobiology

In the human brain, decision-making is a long and complex process that ends in a choice. At the start of the process, the senses of the human body must perceive an external stimulus (Murdock, 1972). When the senses perceive external stimuli, the nerves of the human body must send this information to the human brain. They do this by allowing the cell body to change its electrical output (Figure 1), that is, they increase or decrease their overall electrical charge, creating an electrical impulse (Murdock, 1972). This spreads through the myelin sheath to the axon terminals where it spreads to other nerves in the brain. When the brain is reached, the neurons start a similar process. When the charge spreads from the neurons, they create an electrical impulse (Figure 1). The neurotransmitter makes the inside of the neuron more positive through depolarization.

Depolarization is a process by which the membrane potential of the neuron is made positive through the movement of electrical charges triggered by neurotransmitters (Murdock, 1972). Enough depolarization can cause an action potential, which is a point in which the neuron has built up so much electricity that it sends out an electric impulse (Figure 1). Electrical signals continue through synapses by releasing neurotransmitters from synaptic vessels. They go across the synapse to receptor sites, which creates the electrical signal. The synapse is a small gap between each individual neuron in the brain (Murdock, 1972). Because electricity cannot jump over the gap, the neuron releases a neurotransmitter in order to transmit the impulse to the next neuron (Figure 1). A neurotransmitter is a chemical that is released by the electrical impulses at the axon terminal of a neuron. Each neurotransmitter has a unique shape that matches with a receptor site.

When the neurotransmitter connects to the axon terminal, it causes an electrical signal and continues the process. In terms of decision-making, the brain works in a similar way as a computer (Murdock, 1972). Neurons are binary cells, meaning that they can make one of two choices. They can either send out an electrical impulse, which is a 1 (Hochreiter and Schmidhuber, 1997), or not send out an electrical impulse, which is a 0. The results of doing this binary task hundreds of times, by thousands of neurons, leads to the decision. Decision-making depends on recognition and perception. Recognition is being able to put specific objects, based on they can be seen, into categories (Hochreiter and Schmidhuber, 1997).

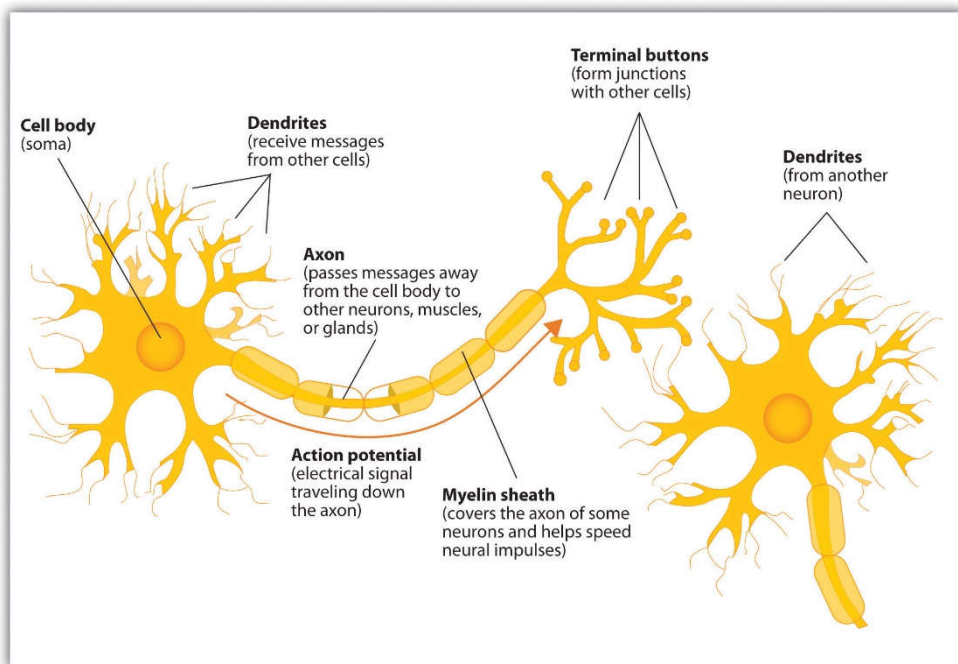


Figure 1. *The Anatomy of a Neuron.* Adapted from “Introduction to Psychology,” by University of Michigan, 2016, University of Michigan Libraries, Copyright (2016).

These categories may be general, such as “safe” or “dangerous,” or they may be more specific, focusing on shape, color, and function. Perception is the experience of sensing stimuli. That means when you put your hand on a table and feel how smooth it is, you are sensing stimuli. The other half of perception is remembering the experience, so that

even if you are not actually feeling the table, you are able to remember how the table felt (Hochreiter and Schmidhuber, 1997). The memories of these experiences are created through the process of long and short-term memories. In short-term memories, a stimulus activates a pattern of activity across neurons in a specific part of the brain. These neurons fire all of their neurotransmitters, allowing a specific memory a period of about 20 seconds to be the dominant memory.

In long-term memory, the same steps happen as in short-term memory. However, the protein kinase A is activated, which sets off CREB (cyclic AMP-response element binding protein). CREB activates genes, which then begin making specific proteins (Hochreiter and Schmidhuber, 1997). Depending on the specific memory, the proteins and the circuits of the brain that this happens in will be modified. A memory that involves motor skills will be different than a memory that requires visual stimuli. These proteins then bind to the neurons, allowing the memory to remain for that neuron's lifetime (Hochreiter and Schmidhuber, 1997).

Two of the proteins that help with the binding of memories to neurons are Amyloid- β and Tau. These two proteins also have the greatest impact on the development of AD (Hernández et al., 2010). Amyloid- β begins its life as a solitary molecule, but over time it starts to form small clusters that travel freely throughout the brain. As the age of the person increases, so does the concentration of Amyloid- β . As the concentration increases, Amyloid- β clusters start to bind to the receptors of neurons (Hernández et al., 2010). These clusters start to form Amyloid- β plaques. This sets off an intracellular process that erodes the synapses between neurons. The destruction of these synapses leads to the loss of memories. The increase of Amyloid- β also leads to the increased concentration of the Tau protein. The Tau protein is normally a protein that helps to stabilize neurons. However, the increase of Amyloid- β in the synapses leads to an increase of Tau proteins. These proteins form tangles within neurons called Tau tangles. Unlike Amyloid- β , which destroys neurons from the outside of the cell, Tau tangles destroy neurons from the inside. Tau tangles disintegrate the system that transports nutrients within the neuron, which leads to its death (Hernández et al., 2010).

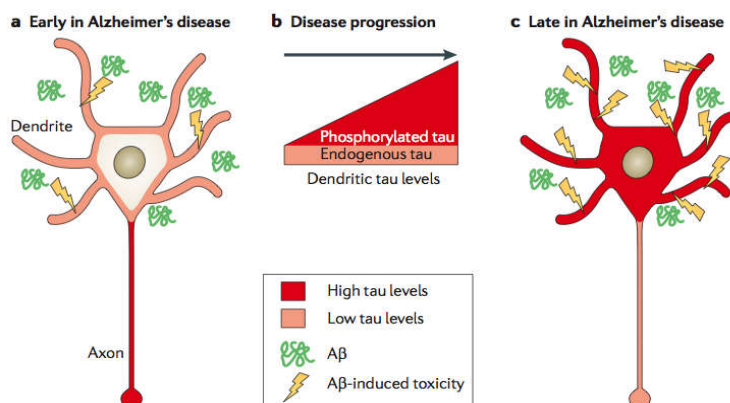


Figure 2. Role of Amyloid- β and Tau in Alzheimer's Progression. Reprinted from "Amyloid- β and tau—a toxic pas de deux in Alzheimer's disease," by J. Goetz, *Nature Reviews Neuroscience*, Copyright (2011).

1 Introduction

1.11 Introduction to Alzheimer's Disease

Alzheimer's Disease is a brain disease that causes the brain to slowly degenerate (Alzheimer's, 2015). It does this by slowly destroying brain cells. It is not infectious or contagious, but it is the most common cause of dementia in the world. Dementia, a symptom of Alzheimer's Disease, is a term that is used to describe the decline of a person's mental ability. This includes cognitive abilities, memory, and language. About 50% of people with dementia suffer from Alzheimer's Disease (Alzheimer's, 2015). Dementia as a whole affects 10% of people 65 or older and 20% of people 75 and older in the United States. Over 5.2 million Americans are estimated to have Alzheimer's disease. By 2050 this number is expected to reach 11 to 16 million (Alzheimer's, 2015). AD has many risk factors. These include smoking, high blood pressure, cholesterol, depression, previous brain trauma, age, and family history with AD. In developed countries, AD is one of the most costly diseases. In 2015, it cost the U.S 259 billion dollars (Alzheimer's, 2015).

1.12 Diagnosing Alzheimer's Disease

One of the reasons AD is increasing at such a rapid rate is because of the improvement in diagnosis (Mohandas, Rajmohan, and Raghunath, 2009). Even though it is easier to diagnosis AD now than it was decades ago, it is still difficult to do so. Diagnosing AD requires a complete medical assessment of a patient. There is no one test that can diagnose AD. In the complete assessment, the patient's medical history, mental status, and emotional stability are all compared with physical tests such as blood tests and brain scans. All of this information is then used in order to diagnose a patient (Mohandas, Rajmohan, and Raghunath, 2009). One of the earliest signs of AD is the appearance of dementia. This includes memory loss, difficulty problem-solving, and poor judgment in making decisions. After these appear, the doctors will interview a family member in order to see whether further testing is required. If it is agreed upon, the doctors use brain-imaging tests like CT scans and MRI's in order to see any abnormal changes. This culminates with a decision made by the doctor as to whether or not the patient has AD (Mohandas, Rajmohan, and Raghunath, 2009).

1.13 Treatment of Alzheimer's Disease

After a person has been diagnosed with AD, the treatment begins. Currently, there is no cure (Mohandas, Rajmohan, and Raghunath, 2009). However, both drug and nondrug treatments may help with dealing with the symptoms. To deal with memory loss, doctors may prescribe drugs such as cholinesterase inhibitors (Aricept, Exelon, Razadyne) and memantine (Namenda). These drugs are used to lessen or stabilize the effect that AD has on neurons (Mohandas, Rajmohan, and Raghunath, 2009). As for behavioral changes caused by AD, those require lifestyle changes to manage symptoms. Those around the patient may need to avoid causing emotional distress around the patient. They also have to maintain a calm environment, give the patient personal comfort, avoid being direct and confrontational, and keep the patient in familiar settings (Mohandas, Rajmohan, and Raghunath, 2009). Sudden changes in setting or people may be distressing to the patient. Antidepressants as well as anxiolytics and antipsychotic medications may also be recommended. Even though the treatment options of AD are limited, we well understand ways to prevent or decrease the risk of getting AD. Regular exercise can decrease the risk of developing AD by 50%. Social engagement is also important (Mohandas, Rajmohan, and Raghunath, 2009). Eating healthy

food, getting enough sleep, and managing stress are all ways that can help to reduce the risk of getting AD.

1.14 Neurobiology of Alzheimer's Disease

From a neurobiological perspective, Alzheimer's Disease is the neurodegeneration of neurons in the human brain. Dementia from AD is related to the irregular production of Amyloid- β and tau, peptides of amino acids commonly found in the brain. Amyloid- β movement in the brain follows a spatial progression pattern, starting at the basal neocortex, spreading throughout the hippocampus, and eventually reaching the rest of the cortex (Uday, 2015). The irregular amounts of Tau spread throughout neural networks, focusing on the primary areas of the neocortex. This irregular production causes amyloid plaques, specifically in the neocortex, which deals with higher-order reasoning. In normal brains, these irregularities would break down and be eliminated. In AD, they form hard, insoluble plaques called amyloid plaques (Uday, 2015). In AD, the buildup of Tau proteins leads to the destruction of the microtubule structures in neurons. In a neuron, the microtubules help to hold up the neuron and keep a definite shape (Mohandas, Rajmohan, and Raghunath, 2009). As they collapse, the size of the brain itself starts to shrink. This leads to the beginning of the symptoms (Uday, 2015).

1.15 Proposal

As explained above, AD is an ever-increasing problem in terms of scale and cost. As explained above, yet one that might be diagnosed ahead of time if many medical and emotional tests are done. However, many of these tests require a lot of time and money (Fischer et al., 2017). They also require large amounts of human body samples, which might be painful to obtain. This is why the production and implementation of a new, written questionnaire exam will help with the diagnosis of AD in a more general population (Alzheimer's, 2015). A written exam will question the emotional and neurological stability of the patients, saving cost and time. Patients who earn a score above a certain point will then be taken in for further testing. This saves people who do not have AD from continuing with the diagnosis process.

1.16 Hypothesis

The hypothesis of this paper is that a questionnaire will facilitate the current diagnosis process for AD. Currently, one-fifth of AD diagnoses may be false positives (Fischer et al., 2017). This wastes billions of dollars per year, not to mention the countless hours lost by these patients from testing and hospital visits (Fischer et al., 2017). This proposal aims to reduce that amount by saving the intense testing and stages of the diagnosis process for those who are most likely to have AD, saving those who are not time and money.

1.2 Literature Review

1.21 The Effect of Ambiguity on Decision-Making

Zamarian, Weiss, and Delazer investigated whether individuals with mild cognitive impairment (MCI) experience difficulties in decision-making in risky and ambiguous situations in the Iowa Gambling Task (IGT) and the Probability-Associated Gambling Test (PAG) (Zamarian, Weiss, Delazer, 2010). They found that people without MCI at first made disadvantageous choices, but over time they began to make more advantageous choices.

People with MCI did not prefer advantageous or disadvantageous choices. They did not improve over time. In the Probability-Associated Gambling Test, people with MCI made less advantageous decisions than those without MCI in low winning probability tests. It was also found that those with MCI have difficulties in making advantageous choices similar to those with mild dementia from Alzheimer's (Zamarian, Weiss, Delazer, 2010).

Delazer, Sinz, Zamarian, Wenning, and Benke conducted a study in order to see how ambiguity impacted decision-making in patients with AD (Sinz et al., 2008). The study tested the decision-making abilities by using the Iowa Gambling Task (IGT) and Probability-Associated Gambling (PAG). From these tests, it was demonstrated that people with AD made more disadvantageous decisions than those who do not have AD. Patients in ambiguous conditions also made random decisions and did not develop a clear strategy. In the PAG, patients with AD also demonstrated less advantageous decision-making. They would often bet and risky and likely deck with at random rates (Sinz et al., 2008). The patient's performance on tasks also correlates with early pathological cerebral changes and cognitive and emotional deficits. This indicates that as the number of deficits from AD increase from pathological cerebral changes, the performance in both the IGT and PAG decreases and causes the patients to make less advantageous choices. From the study, ambiguity is shown to amplify the number of disadvantageous choices made (Sinz et al., 2008).

Hot, Ramdeen, Borg, Balon, and Couturier conducted a study to see whether the decline of decision-making abilities in people with AD is correlated with the use of incorrect or no strategy (Hot et al., 2013). Using the IGT, patients with AD had impaired and decreased performance in the IGT; the patients that were generally happier performed better on the IGT than those who were not. Additional analysis demonstrated that decreased performances on the IGT was not due to memory functions. This allows the proposition of the idea that higher uncertainty level in patients with AD can be reduced through emotional responses (Hot et al., 2013).

1.22 The Effect of Apathy on Decision-Making

Bayard, Jacus, Raffard, and Nargeot conducted tests of patients with mild dementia from Alzheimer's Disease (AD) and with MCI (Bayard et al., 2014). They were tested using the Iowa Gambling Task and the Lars Apathy Scale. The aim of the study was to investigate decision-making based on emotional feedback processes. From the study, both people with AD and MCI had reduced performances if they had higher scores on the Lars Apathy Scale. Those with lower scores did better on the IGT. This related disadvantageous choices on the IGT to apathy levels. The correlation might lead to further studies using the IGT as a risk factor for increasing apathy, especially among older populations (Bayard et al., 2014).

1.23 The Effect of Reaction Time on Decision-Making

Delazer, Sinz, Zamarian, and Benke conduct a study to try and understand the strategy that people with mild Alzheimer's follow in decision-making (Delazer et al., 2008). This was tested by using a gambling test. The patients with AD shifted more frequently between safe and risky choices. They also demonstrated less consistent strategies and choices. Due to the frequent changes, this indicates that the choices were random and that there was no consideration of strategy. This is in comparison to those without AD, who favored safe choices and demonstrated a consistent strategy (Delazer et al., 2008). Those with AD also did not adapt their strategies as the problems changed. Those without AD demonstrated a clear change in strategy. These differences are not attributed to impulsive reactions since those with AD and those without AD have similar reaction times. This also

proposes that the gambling test could be used as an indicator for AD, because the frequency of changes between safe and risky choices could be used as a fair indicator (Delazer et al., 2008).

1.24 The Ability to Comprehend and Use that Information for Decision-Making

Karlawish conducted a literature review to develop a model that covers four decision-making concepts: understanding, appreciating, reasoning and choice (Karlawish, 2008). Karlawish tested the amount of information that those with AD could retain from reading. After being given a text to read, they would be asked a question referencing said text. Less than 40% of those with AD received a passing grade on the test. Another test for AD patients had an interview on the text with a supervisor (Karlawish, 2008). This time, 60% of the patients received passing grades. This indicated that the decision-making process is impacted by what senses are used to perceive the information. In general, the study found that even when information is explained or written to people with AD, they will often be unable to comprehend substantial parts. Also, those with mild dementia are more likely to retain more important information than those with mild AD (Karlawish, 2008).

1.25 The Effect of Alzheimer's Disease on Decision-Making in Giving Consent

High examined the effect that AD has on people's ability to give consent (High, 1992). Though there are already laws in place for this, there is a lack of uniformity in how well the laws are able to apply to people with AD. People with AD who have to give consent face emotional problems, problems of not having enough information, and fluctuating orientations of the mind. People under these circumstances are facing similar emotional and environmental stimuli as those taking the IGT or PAG (High, 1992). From these people, we see that those who are more uninformed or ignorant of the situation make the same number of disadvantageous choices as those who would be considered informed on the situation. Also, those who are more emotional, specifically those who are more apathetic, end up making more disadvantageous and more random choices than those who are not (High, 1992).

1.26 Significance

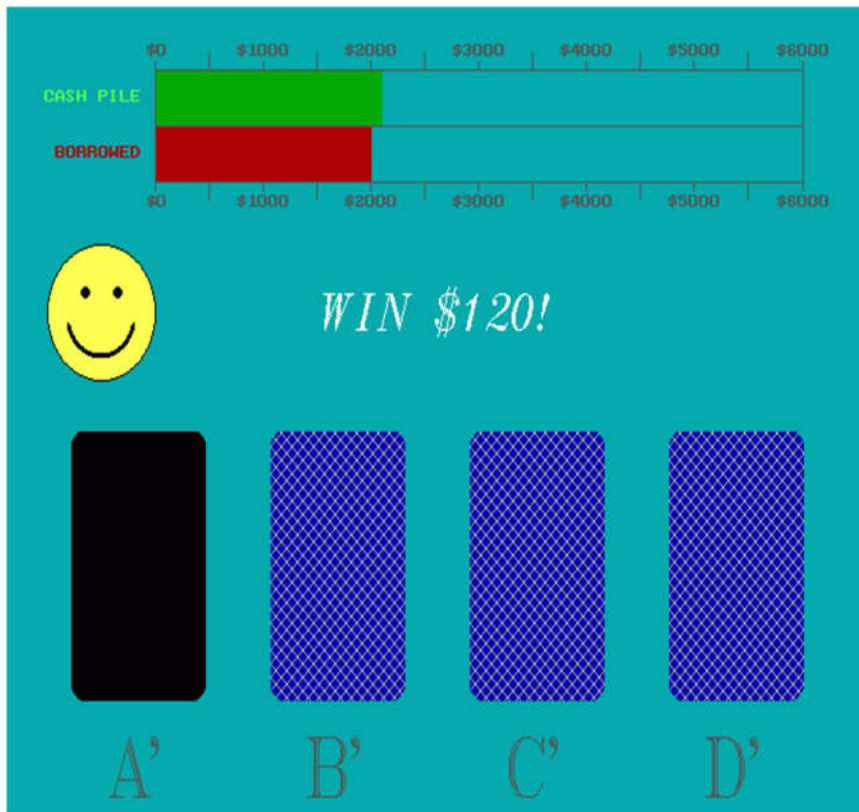
The proposed exam would be significant for the fact that it would reduce the time, cost, and investments needed to be diagnosed with or without AD (Alzheimer's, 2015). This proposal would make it so that the cheapest and quickest part of the diagnosis process, the questionnaire, would separate participants who need further testing from those who do not. Instead of making all participants conduct the costly and time-intensive testing, the testing could be reserved for those who got above a specific score on the exam. This makes it so those who score below do not have to be bothered with additional testing. More than 1/5 AD diagnosis may be false positives (Fischer et al., 1992). This means 1/5 people who have been diagnosed with AD may be spending money, time and recourses on medicine and testing they do not need. These resources could be directed to people that actually do have AD (Fischer et al., 1992).

2 Established Methodology

2.1 Iowa Gambling Task (IGT)

The Iowa Gambling Task (IGT) is a psychological test that is designed to stimulate real-life decision-making (Li et al., 2010). In it, the participants are given four virtual decks

of cards on a monitor (Brevers et al., 2013). They are told that every deck has cards that either reward or penalize them. They are given game money, and if they are rewarded, they gain money. If they are penalized, they lose money (Brevers et al., 2013). In the study, there were 20 patients with AD, 20 participants with MCI, and 20 healthy people who were the controls (Figure 3). All participants completed the Iowa gambling task (IGT). The goal of the task is to earn as much money as possible. The decks are different as each deck has a different ratio of reward to penalty decks. Therefore, there are “bad” and “good” decks (Li et al., 2010).



*Figure 3. The Home Screen of the Iowa Gambling Task.
Reprinted from the Iowa Gambling Task Homescreen.*

2.2 Probability-Associated Gambling Task—Revised (PAQ)-R

The PAG-R is a psychological test that is designed to test the ability of the participant to adapt to changing probabilities overtime (Sinz et al., 2008). In the task, the participants are given four decks. Each deck is designated a rating. Either it has a high-winning probability, or it has a low-winning probability (Sinz et al., 2007). If you win, you gain game money. If you lose, you lose game money. However, a high-winning deck it not guaranteed to win, and vice versa. The goal of the task is to earn as much money as possible (Sinz et al., 2008).

2.3 Lars Apathy Scale (LAS)

The LAS is a written questionnaire that examines the amount of apathy participants have (Sockeel, 2006). The examination itself has 18 statements, each of them being answered on a four-point scale (Sockeel et al., 2006). For example, the statement of "Meeting my family is important for me" might be displayed. The participants will then circle a number. The higher the number, the more they disagree with the statement. The smaller the number, the more they agree. A higher score means they are less apathetic, and vice-versa (Sockeel, 2006).

2.4 Lille Apathy Rating Scale (LARS)

The LARS task is a structured interview to test levels of apathy (Ross, 2017). It includes 33 items, divided into nine domains. In the interview, the questions are based on day-to-day tasks. Each section is designed to test a certain part of emotional intelligence, including emotion and curiosity. Responses are scored on a dichotomous scale (Ross, 2017).

2.5 Mild Cognitive Impairment (MCI)

Mild cognitive impairment is a middle stage between the cognitive decline of normal aging and the more-serious decline that comes with dementia (Boyle et al., 2012). It can involve problems with memory, language, thinking, and judgment that are greater than normal age-related changes. People with MCI are at a greater risk of getting more advanced stages of dementia. MCI itself is not severe, as people affected by it can still function in a normal setting (Boyle et al., 2012).

2.6 Alzheimer Smell Test

In the development of Alzheimer's, the sense of smell is the only sense that is not affected by the deterioration of the brain (Duff, McCaffery, and Salomon, 2002). Because the frontal lobe is the center for the sense of smell, it is not affected by the effects of AD. The other senses are in the cerebellum, which is heavily affected by AD. In the AD smell test, the participants are given scent cards that smell of peanut butter, soap, and mouthwash. They are asked the participants to describe the object. People who pass will not have to worry about AD developing for four years (Duff, McCaffery, and Salomon, 2002).

3 Demographic and Clinical Data

3.1 Demographic Data

Demographic and clinical data are reported on, in detail, in Table 1 (Bayard et al., 2014). The table refers to the IGT, PAG-R, and the LAS participants. AD and MCI participants were, on average, older (resp., $P < 0.001$ and $P = 0.043$) and less educated (resp., $P = 0.09$ and $P = 0.027$) than the controls. In terms of education and age, there were not many differences between AD and MCI participants (all P values = 1) (Bayard et al., 2014). All groups were matched for gender. As expected, AD participants performed worse than MCI participants ($P = 0.006$) and controls ($P < 0.001$) on the MMSE (Mini-Mental State Examination), as a significant difference was also observed between MCI participants and controls ($P = 0.018$). Finally, there was no significant difference between groups with reported depression symptom severity using the BDI total score (Bayard et al., 2014).

Table 1. Demographic and Clinical Data for the CDT. Data from “Apathy and Emotion-Based Decision-Making in Amnesic Mild Cognitive Impairment and Alzheimer’s Disease” by S. Bayard, 2014, Behavioral Neurology.

	Healthy controls (n = 20)	MCI participants (n = 20)	DTA participants (n = 20)	Statistics	P value
Demographic and clinical data					
Age, mean (SD)	73.5 (6.7)	78.25 (6.9)	80.9 (5.4)	$F = 8.7$	<0.001*
Sex, n (women)	11	11	12	$\chi^2 = 0.13$	0.93
Years of education, mean (SD)	11.1 (2.7)	7.9 (2.4)	8.3 (3.1)	$F = 7.7$	0.003*
Mini-mental State Examination ^a , mean (SD)	28.5 (0.9)	27.15 (2)	24.8 (2.3)	$F = 11.6$	<0.001*
Beck depression inventory					
Total score, mean (SD)	10.6 (6.5)	12.2 (7.8)	13 (7.44)	$F = 0.57$	0.57
Moderate (>18), n (%)	2 (10)	4 (20)	3 (25)	$\chi^2 = 1.55$	0.45
Severe (>19), n (%)	1 (5)	0	0	$\chi^2 = 2.03$	0.36
Lille apathy rating scale ^c , mean (SD)					
Intellectual curiosity	-2.6 (0.76)	-1.4 (0.96)	-1.5 (1.01)	$F = 5.6$	0.006*
Emotion	-3.6 (0.59)	-2.5 (1.29)	-2.7 (1.30)	$F = 1.9$	0.15
Action initiation	-3.5 (0.64)	-2.2 (1.15)	-2.0 (1.74)	$F = 5.1$	0.009*
Self-awareness	-2.9 (1.14)	-2.4 (1.27)	-2.5 (1.31)	$F = 0.1$	0.86
Total score, mean (SD)	-2.8 (4.18)	-17.8 (6.80)	-19.1 (6.40)	$F = 9.6$	<0.001*
Lille apathy rating scale-cutoff					
Lightly apathetic to severely apathetic, n (%)	1 (5)	12 (60)	12 (60)	OR = 3.62 95% CI 1.6 to 7.7	<0.001
IGT disadvantageous profile (net score <0)					
Total net score, n (%)	6 (20)	13 (45)	10 (35)	$\chi^2 = 5.64$	0.02*
Blocks 3 to 5 (trials 41-100), n (%)	5 (18)	12 (42)	11 (39)	$\chi^2 = 6.66$	0.03*
Executive function assessment ^d , mean (SD)					
Hayling Test (time)	4 (2.34)	7.5 (4.53)	6.8 (3.67)	$F = 1.66$	0.19
Hayling Test (error)	2.7 (1.92)	8.3 (7.34)	9.8 (5.90)	$F = 3.27$	0.04*
Trail Making Test (time)	112 (50)	191 (60)	198 (46)	$F = 5.98$	0.004*
Trail Making Test (error)	0.5 (1.05)	3.1 (4.58)	5.3 (7.14)	$F = 1.96$	0.14
Updating memory task	3.3 (0.83)	2.6 (0.64)	2.4 (0.64)	$F = 1.56$	0.21

AD: Alzheimer’s disease; CI: confidence interval; IGT: Iowa gambling task; MCI: mild cognitive impairment.
*Controls ≠ (MCI ≠ DTA); ^aControls ≠ MCI; ^bControls ≠ MCI; ^cDTA; ^dadjustment for age and education.

3.2 Results of Impact of Emotion on Decision-Making in Patients with Alzheimer’s Disease

In the IGT, both the MCI and the AD patients had significantly reduced results. Both MCI’s and the AD patients were more emotional compared to the controls, but there was not a difference between the MCI’s and the AD patients in terms of emotion (Bayard et al., 2014). A group effect was demonstrated for the IGT net win ($F = 6.83$, $P = 0.002$), with a lower final outcome in MCI’s (mean = 492, SD = 255) and AD participants (mean = 630, SD = 199) than compared to controls (mean = 1416, SD = 280; resp., $P = 0.003$, Cohen $d' = 3.63$ and $P = 0.014$, Cohen $d' = 3.27$) (Figure 4). There was no difference found between MCI and AD Patients ($P = 0.1$) (Bayard et al., 2014). It was also discovered that MCI and AD participants and controls displayed different decision-making patterns during the task (Figure 4) ($F = 2.42$, $P = 0.025$, $\eta^2 = 0.07$). In trials 41 to 100, a group effect was noted ($F = 3.43$, $P = 0.042$, $\eta^2 = 0.10$). Controls performed better than MCI and AD patients did (resp., $P = 0.048$, Cohen $d' = 1.18$ and $P = 0.043$, Cohen $d' = 1.21$) (Bayard et al., 2014). There was no difference observed between MCI and AD patients ($P = 0.8$). Overall, the percent of participants with a more disadvantageous record was higher in MCI and AD groups than controls (Table 1, all P values < 0.05). There was no significant difference observed between AD and MCI participants (all P values > 0.8) (Bayard et al., 2014). In summary, patients with AD and MCI made less advantageous choices than those in the control group. Also, there is a negative correlation between the level of emotions that the patient’s exhibit and the score they get on the IGT (Bayard et al., 2014). This makes emotion an influential part of the decision-making process.

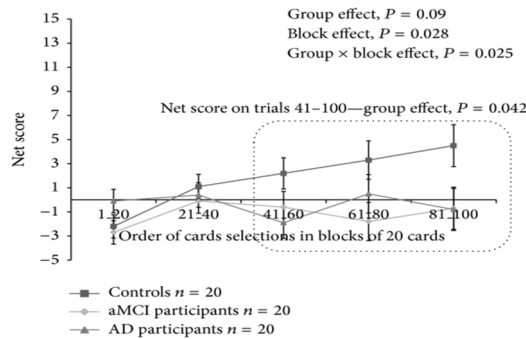


Figure 4. Demographic and Clinical Data for the CDT. Data from “Apathy and Emotion-Based Decision-Making in Amnesic Mild Cognitive Impairment and Alzheimer’s Disease” by S. Bayard, 2014, Behavioral Neurology.

3.3 Results of Impact of Ambiguity on Decision-Making in Patients with Alzheimer's Disease

Overall, Mild Alzheimer's and Dementia (DAT) and MCI patients in the IGT selected the advantageous decks less frequently than healthy controls (Bayard et al., 2014). DAT patients chose the more advantageous decks more frequently than healthy controls in the first deck ($p < 0.01$), whereas the opposite pattern was found for decks 3, 4 and 5 (t-tests, all $p < 0.05$) (Bayard et al., 2014). In general, the frequency of advantageous choices significantly increased over the task for healthy participants (block 1 $p < 0.001$), while no relevant difference between decks was detected for DAT patients (Table 2). In the PAG-R, the low probability decks had chances between $p = 0.125$ and $p = 0.375$ to win money. The high probability decks chances are $p = 0.625$ and $p = 0.875$ to win (Bayard et al., 2014). DAT patients chose to gamble more often than the controls in the low probability decks, whereas the opposite pattern was found for the high probability condition (Table 2). In the low probability conditions, AD patients gambled more often than healthy participants, and for the patients with AD, the ambiguity of the decks led them to make significantly worse choices than the control group ($p = 0.125$, $p < 0.01$; condition $p = 0.375$, $p < 0.01$). In contrast, they gambled less frequently in the highest winning probability condition (condition $p = 0.875$, $p < 0.01$; in the condition $p = 0.625$, differences were not significant; Table 3) (Bayard et al., 2014). In general, AD and MCI patients gambled more than healthy patients. Also, there was a correlation between the ambiguity of the decks and the amount which patients gambled (Table 2). This leads to the conclusion that ambiguity is a major factor in the decision-making process (Bayard et al., 2014).

Table 2. Demographic and Clinical Data for the CDT. Data from “Impact of ambiguity and risk on decision-making in mild Alzheimer's disease” by H. Sinz, 2008, Neuropsychologia.

	M (S.D.)		t-Test (p-value)
	DAT patients	Controls	
IGT			
Advantageous choices (C + D) (frequency)			
Total sum	49.4 (4.4)	58.5 (8.6)	0.0001
Block 1	9.5 (1.7)	7.5 (2.4)	0.005
Block 2	9.6 (1.6)	10.8 (2.3)	n.s.
Block 3	10.8 (1.9)	13.0 (3.0)	0.01
Block 4	10.0 (1.9)	13.7 (3.6)	0.0001
Block 5	9.2 (1.6)	13.5 (4.3)	0.0001
Shifts between single decks (frequency)			
Total sum	92.0 (10.6)	59.0 (24.6)	0.0001
Block 1	19.6 (0.5)	13.9 (5.0)	0.0001
Block 2	19.3 (1.2)	12.9 (5.3)	0.0001
Block 3	17.8 (4.2)	11.4 (6.0)	0.001
Block 4	18.5 (2.0)	11.0 (5.6)	0.0001
Block 5	16.7 (4.0)	9.8 (5.5)	0.0001
Shifts between good (C + D) and bad (A + B) decks (frequency)			
Total sum	61.2 (9.4)	35.7 (19.1)	0.0001
Block 1	13.4 (2.6)	8.0 (4.4)	0.0001
Block 2	13.1 (1.9)	8.6 (4.1)	0.0001
Block 3	11.5 (3.4)	7.0 (5.0)	0.003
Block 4	12.4 (2.6)	6.8 (4.8)	0.0001
Block 5	10.9 (3.8)	5.3 (3.9)	0.0001
Net win (€)	1834 (731)	1967 (902)	n.s.
Final borrow (€)	3059 (1029)	2455 (858)	0.053
PAG			
Gambles (frequency) ^a			
p=0.125	1.2 (1.0)	0.3 (0.6)	0.0001
p=0.375	1.6 (1.4)	0.5 (0.8)	0.004
p=0.625	3.2 (1.1)	3.8 (1.1)	n.s.
p=0.875	3.7 (1.3)	4.7 (0.6)	0.002

Legend: M, mean; S.D., standard deviation; n.s., not significant.
^a Trials corresponding to different fix sums (+20, -20) are collapsed.

3.4 Results of Impact of Apathy on Decision-Making in Patients with Alzheimer's Disease

Of the 40 participants with AD or MCI, 29 participants were identified as having a disadvantaged profile (IGT < 0); 28 demonstrated a preference for disadvantageous choices for decks 3 to 5 net score (trials 41–100; IGT < 0) (Sinz et al., 2008). Table 1 indicates that participants with an advantageous profile at the IGT (net score > 0) were less apathetic than participants who demonstrated a preference for disadvantageous choices (net score < 0). Overall, this represents a negative correlation between apathy, and decision-making. In the survey, 29/30 participants with AD or MCI reported having an increase in making disadvantageous choices. This means that the view that participants bring to the task impacts the way they perform it (Sinz et al., 2008). Those who are more apathetic towards a particular task or goal are more likely to make less advantageous choices that relate to that task. Those who are less apathetic are more likely to make more advantageous choices. And this difference is significantly increased in people with AD and MCI. This means that apathy is an important part of the decision-making process (Sinz et al., 2008).

4 Proposal

4.1 Discussion of Presented Paper Results

The data presented in the previous proposals shows that the impact of emotions is a significant variable in decision-making. When making a choice, emotions will impact the

way the brain will think before a final decision is reached. Because emotions are linked to memories, the increase of proteins linked with emotions block past experiences (Sinz et al., 2008). The blockages of these neurons in the frontal lobe leads to a decrease in logical and level-headed thinking (Tabert et al., 2005). In people with AD, this effect is increased dramatically since these people already have a limited number of neurons left, due to the accumulation of Amyloid- β . It is shown that the ambiguity of questions affects the decision-making process. The IGT shows that the ambiguity of the task leads to confusion and anger. This ambiguity leads into emotions and apathy, two other influential factors in the decision-making process (Sinz et al., 2008). This impact is due to the confusion about unknown details, as well as the required skill of needing to use the frontal lobe. Since patients with AD have limited control of their frontal lobe, they cannot use logic to solve ambiguous problems. Apathy is a condition that makes patients lose interest in doing well (Tabert et al., 2005). This is caused by a multitude of factors. One reason is that due to natural human emotions, when confronted with a difficult task with seemingly no solutions, humans tend to abandon the task. Since the patients were not confronted with this option, they chose to finish the task but did so apathetically. This led to the same reaction times, but to less thought being devoted to each problem (Tabert et al., 2005). Also, as apathy is an emotion, the rise of apathy also leads to the influence of general emotions in the decision-making process. This increases the severity of the impact on decision-making, as demonstrated by the increased difference between the control and the experimental groups.

4.2 Proposed Methodology

The current diagnosis process for AD in America and the world at large is inefficient, costly, time-wasting, and inaccurate. More than $\frac{1}{5}$ patients diagnosed with AD may not have AD, or may not have a severe case of AD that requires massive testing and medicine. These cases result in billions of dollars and countless hours being wasted (Alzheimer's, 2015). These resources could be redistributed to patients with likely or current cases of AD. To combat this waste, this proposal intends to diagnosis suspected patients of AD through inexpensive, accurate, general tests and screenings. These tests will decrease the number of false positives for the diagnosis process, as only those deemed "likely" to have AD will go on and get further, expensive testing. Those that score below a certain percentile do not have to worry about developing AD for the next four years. Those within a warning range should be monitored and should come to retake the test in six months to a year.

The proposal is to use the Cumulative Diagnostic Test (CDT). The CDT is a 5-stage questionnaire and interview test. The 5-stages covered are emotions, interview signs, apathy, sniff test, and the IGT (ambiguity task). The reason specific variables are covered is that from previous studies, it has been shown that emotions, apathy, and ambiguity are the largest influencers in the decision-making process. In the emotions test, the patients respond to a questionnaire. The range for this exam is -8 to 8, where an eight is the maximum and negative eight is the minimum. The higher the score, the lower the emotional intelligence. The average range for healthy people is from -2 to 2 (Appendix). The apathy exam is given in an interview-like setting. In the interview, the patients are presented statements in which they are asked to rank how much they agree or disagree. This ranking is on a scale from 1-4, where 4 is strongly disagrees, and 1 is strongly agrees. The range is from 18 to 72, where the normal range for healthy adults is from 18 to 28 (Appendix). In the smell test, the patients would be blindfolded. They would then be presented with three strips. One strip would smell like beans, one like mouthwash, and one like soap (Duff et al., 2002). After patients

smells the strip, they would then be asked to describe a memory associated with that smell. They would be asked about who, what, where, when, and why. For each one that they fail to describe, they gain a point. There are 15 points in total, with the average for healthy elders being from zero to two.

The next test would be the interview signs test (Appendix). This would be in conjunction with the apathy test and the smell test. During the interview during these two exams, the interviewee would notice and check off if the patient exhibited any asocial or awkward motions. The scale is from zero to five, where five means most of the aforementioned signs were exhibited. The final test is the IGT. The IGT tests the patient's ability to deal with ambiguity and ambiguous situations (Sinz et al., 2008). As described in the methods, the IGT's main goal is to gain as much money as possible from choosing the more beneficial decks. In the IGT, the range of scores that a patient can get is from 0 dollars to 4500 dollars. The normal range for a healthy elder is from 2250 dollars to 4500 dollars.

Results from the five tests are compiled and put into a singular, comprehensive formula that reduces and approximates all of the results into one number. This number ranges from zero to 11, with the normal range being zero to four. This means that the patient does not have to worry about developing AD for the next four years. From four to six, there is a warning range, which means that that the patient should retake the test in six months to a year (Seeber et al., 2008). Six to eight indicates a likely case of DAT. This range will require further, more physical testing to prove (Alzheimer's, 2015). Eight to eleven indicates a severe case of AD, for which more physical testing is required, as well as some medication.

From this proposal, the percentage of false positives will decrease from $\frac{1}{5}$. In the U.S, 18.2 billion hours of care and 230 billion dollars are given to the AD diagnosis process (Alzheimer's, 2015). With this proposal, more than 46 billion dollars of AD funding can be diverted from false positives into more advanced testing. And the U.S as a whole could save about 3.72 billion hours normally given to the diagnosis process to help patients with known cases of AD.

4.3 Analysis of Data

Since the CDT is an accumulation of established and new AD tests, the data from individual studies of these exams can be averaged out and cultured into an average for the CDT. For the Iowa Gambling Task, 100 people were divided into two groups: the control and the experimental group. There were 60 controls, of people with good mental health. There were 40 people in the experimental group, with MCI's or AD. After taking the IGT, it was shown that the average score for the control group was 1415, while the average score for the experimental group was 473 (Sinz et al., 2008). For the interview signs, in a group of 159 patients with AD and 58 healthy subjects, AD and MCI patients demonstrate, on average, 2.7 out of the 5 asocial signs during an interview (Sockell et al., 2006). The control group of healthy, elderly people scored 0.5 out of 5, on average. For the emotion test, in a group of 159 patients with AD and 58 healthy subjects, the average patient with AD scored between 4-6. For the Apathy Scale, in a group of 159 patients with AD and 58 healthy subjects, the average score of AD patients is 44 out of 72 points (Sockell et al., 2006). For the AD smell test, 147 patients with MCI and 100 AD were tested with the sniff test, and they got a 6.78 out of 15 on average (Table 3). For the Table below, the scores for the MCI, AD and healthy elders were converted to the scale of the LARS scale of the CDT. This was done in a traditional proportion table. The control group, with 58 people, on average, scored 1.53 out of 15. In the CDT, a higher score in the AD is correlated with a higher risk of AD. From this distribution of data, this is submitted into the CDT, which gives the average score

of both the control group and the experimental group (Tabert et al., 2005). In the normalized set of 1056 total data points, 784 being the experimental group with AD, and 272 being the control group of healthy elderly people, the average CDT score for the experimental group is 2.44 out of 11, which puts it in the range of being safe from the risk of AD for four years (Sinz et al., 2008). The average CDT score for the experimental group with DAT is 6.10/11. This puts it in the range of DAT, which all of the members of the experimental group had.

Table 3. Demographic and Clinical Data for the CDT. Data from “A 10-item smell identification scale related to risk for Alzheimer's disease.” by M. Tabert, 2005, Annals of Neurology.

Demographic Variable	Healthy Elderly (n = 63), Mean (SD)	MCI Patients (n = 147), Mean (SD)	AD Patients (n = 100), Mean (SD)	<i>p</i> ^a	MCI Nonconverters (n = 109), Mean (SD)	MCI Converters (n = 38), Mean (SD)	<i>p</i> ^b
Age (yr)	65.71 (9.38)	67.43 (9.85)	71.72 (9.54)	<0.001	65.59 (9.99)	72.71 (7.28)	<0.001
Education (yr)	16.68 (2.60)	14.96 (4.29)	13.09 (4.35)	<0.001	15.27 (4.19)	14.08 (4.49)	0.142
Sex (% female)	54.0	55.1	63.8	0.33	54.1	57.9	.417
Folstein	29.37 (0.768)	27.28 (3.23)	19.96 (5.96)	<0.001	27.68 (3.43)	26.13 (2.21)	0.01
MMSE score							
UPST score	34.86 (4.18)	31.22 (6.45)	23.72 (6.48)	<0.001	33.02 (4.68)	26.05 (7.96)	<0.001
B-SIT score	10.60 (1.53)	9.56 (2.21)	7.04 (2.62)	<0.001	10.12 (1.70)	7.95 (2.67)	<0.001
10-item Scale score	8.98 (1.24)	8.26 (1.66)	5.48 (1.71)	<0.001	8.75 (1.23)	6.84 (1.90)	<0.001

^aOne-way analysis of variance or Fisher's exact test (sex) were conducted to compare healthy elderly, MCI patients, and AD patients.
^b*t* tests or Fisher's exact test (sex) were conducted to compare nonconverters vs converters to AD on follow-up evaluation.

MCI = mild cognitive impairment; SD = standard deviation; AD = Alzheimer's disease; MMSE = 30-item Mini-Mental State Examination; UPST = University of Pennsylvania Smell Identification Test; B-SIT = Brief Smell Identification Test.

4.4 Limitations of Method

From the presented proposal, there are some minor problems that arise in its methodology, execution, and necessity. For the methodology, all of the data was derived from human studies. This allows human error or a misinterpretation of the presented results. Also, the statistical analysis done on the data may dilute the data or all for unintentional changes. For the data points, there are almost three times the number of experimental participants as there are control participants. For the test, AD develops differently in different people. This makes it so that it is nearly-impossible for there to be a single, universal test to diagnosis AD. This proposal was designed to be able to diagnose the vast majority of AD cases through the traditional and developmental patterns (Alzheimer's, 2015). However, not every case follows these patterns. For the execution, the time and cost saving analysis may change to new discoveries, poor implementation of the program and by individual doctors. This proposal requires there to be an interview setting, with a qualified interviewer trained in the medical or psychological sciences. This requirement may be a burden on some hospitals, decreasing its implementation. Also, the importance of this proposal may vary from doctor to doctor. Some doctors may feel that the proposal is too lenient and may not follow all of the suggestions. Out of necessity, proposal was designed to be a conglomeration of the most accurate and most detailed AD tests in current use. This proposal does not take into consideration upcoming or theoretical AD prediction models. These models may yield future results, but are not considered substantive enough to base a proposal on (DeFina et al., 2013).

4.5 Statistical Analysis

The data were examined for normal distribution (tested with the Kolmogorov-Smirnov test) and for a homogeneity of variance (tested with the Levene test) (Bayard et al., 2014). For the normally distributed data, parametric tests were used (Student's t-test for

independent samples, univariate analysis of covariance (ANCOVA), analysis of covariance with repeated measures, and Greenhouse-Geisser adjusted degrees of freedom (MANCOVA)). If there were significant deviations from the normal distribution, we used corresponding nonparametric methods (Mann Whitney, U test, chi-square test, and logistic regression) (Bayard et al., 2014). We calculated partial eta squared (η^2) and Cohen's d' as a measure of the effect size and designated the effect size as small ($\eta^2 = 0.01$; $d' = 0.2$), medium ($\eta^2 = 0.06$; $d' = 0.5$), or large ($\eta^2 = 0.14$; $d' = 0.8$). The level of significance was set at $P < 0.05$. All statistical analyses were carried out using the Statistical Package for the Social Sciences (SPSS) version 19 for Windows, and all of the analysis was carried out by the Bayard Laboratory (Bayard et al., 2014).

5 Conclusion

AD is the fastest-growing cause of dementia and mental illness for elders in the world (Alzheimer's, 2015). Because of this increase, the rush to diagnosis AD in elders is leading to an increasing number of false positives. More than one fifth of AD diagnosis are false positives. In the U.S, 18.2 billion hours of care and 230 billion dollars are given to the AD diagnosis process (Alzheimer's, 2015). This means that 3.6 billion hours and 46 billion dollars are spent on false positives. These resources could be better spent on more research. In the current AD diagnosis process, patients must go through physical and mental tests that are costly in terms of time and money. Patients have to pay around 5000 dollars per hour for an AD consultation (Sockell et al., 2006). This is a large amount of money for someone who is yet to be diagnosed. This current proposal is designed to save time and money for both patients and doctors. By holding off the more expensive parts of the AD diagnosis process until patients have reached a certain level of risk, this proposal can save billions of dollars and hours. From the discussed methods, the discussed AD diagnosis methods will be used as a ground work. Using their publicly available samples sizes sample variable, this proposal intended to use this as a base for future studies of the CDT. In the proposal, 1056 data points from 272 healthy elderly people (control) and 784 elderly people with DAT are used in order to develop a new way of diagnosing AD (BRF, 2017). In the CDT, those that score below a certain percentile do not have to worry about developing AD for the next four years. Those within a warning range should be monitored and should retake the test in six months to a year (Appendix). The CDT is a 5-stage questionnaire and interview test. The 5-stages covered are emotions, interview signs, apathy, sniff test, and the IGT (ambiguity task). The reason these specific variables are covered is that from the previous studies, it has been shown that emotions, apathy, and ambiguity are the largest influencers in the decision-making process. In the study with the CDT, the average score of the experimental group was 6.10/11, which is the range for DATA or mild AD. The average score of the control group was 2.4/11, which means that it is unlikely that they will develop AD for the next four years (Appendix). With this proposal, more than 46 billion dollars of AD funding can be diverted from false positives into more advanced testing. The U.S as a whole could save about 3.72 billion more hours from the diagnosis process to help patients with known cases of AD (Alzheimer's, 2015). For future research, the proposal suggests developing less behavior-focused methods of diagnosis. More affordable and time-effective physical diagnosis methods are needed. The correlation between hippocampal volume and cognitive abilities is applicable to this research proposal, as well as the development of Tau Tangles (Nathan et al., 2017). These physical diagnosis methods could lead to more affordable and reliable diagnosis methods in the future with a combination of behavioral and physical analysis.

6 References

1. Alzheimer's, A. (2015). 2015 Alzheimer's disease facts and figures. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 11(3), 332.
2. Bayard, S., Jakus, J., Raffard, S. and Gely-Nargeot, M. (2014). Apathy and Emotion-Based Decision-Making in Amnesic Mild Cognitive Impairment and Alzheimer's Disease. *Behavioural Neurology*, 2014, 1-7. <http://dx.doi.org/10.1155/2014/231469>
3. Biomedical Research Forum. (n.d.). Retrieved November 09, 2017, from <http://www.alzforum.org/databases>
4. Boyle, P., Yu, L., Wilson, R., Gamble, K., Buchmann, A. and Bennett, D. (2012). Poor Decision Making Is a Consequence of Cognitive Decline among Older Persons without Alzheimer's Disease or Mild Cognitive Impairment. *PLoS ONE*, 7(8), 436-443. <https://doi.org/10.1371/journal.pone.0043647>
5. Brevers, D., Bechara, A., Cleeremans, A., & Noël, X. (2013). Iowa Gambling Task (IGT): twenty years after – gambling disorder and IGT. *Frontiers in Psychology*, 4, 665. <http://doi.org/10.3389/fpsyg.2013.00665>
6. Cosentino, S., Metcalfe, J., Cary, M. and Karlawish, J. (2017). Memory awareness influences everyday decision-making capacity in Alzheimer's disease. *Alzheimer's and Dementia*. <http://dx.doi.org/10.1016/j.jalz.2011.05.683>
7. DeFina, P., Moser, R., Glenn, M., Lichtenstein, J. and Fellus, J. (2013). Alzheimer's Disease Clinical and Research Update for Health Care Practitioners. *The Journal of Neuroscience*. <http://dx.doi.org/10.1155/2013/207178>
8. Delazer, M., Sinz, H., Zamarian, L. and Benke, T. (2007). Decision-making with explicit and stable rules in mild Alzheimer's disease. *Neuropsychologia*, 45(8), 1632-1641. <https://doi.org/10.1016/j.neuropsychologia.2007.01.006>
9. Duff, K., McCaffrey, R. and Salomon, G. (2002). The Pocket Smell Test. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 14(2), 197-201. <https://doi.org/10.1176/jnp.14.2.197>
10. Fischer, C., Qian, W., Schweizer, T., Ismail, Z., Smith, E., Millikin, C. and Munoz, D. (2017). Determining the impact of psychosis on rates of false-positive and false-negative diagnosis in Alzheimer's disease. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 3(3), 385-392. <https://doi.org/10.1016/j.trci.2017.06.001>
11. Gleichgerrcht, E., Ibáñez, A., Roca, M., Torralva, T. and Manes, F. (2010). Decision-making cognition in neurodegenerative diseases. *Nature Reviews Neurology*, 6(11), 611-623. <https://dx.doi.org/doi:10.1038/nrneuro.2010.148>
12. Hernández, F., de Barreda, E. G., Fuster-Matanzo, A., Lucas, J. J., & Avila, J. (2010). GSK3: a possible link between beta amyloid peptide and tau protein. *Experimental neurology*, 223(2), 322-325.
13. High, D. M. (1992), Research with Alzheimer's Disease Subjects: Informed Consent and Proxy Decision Making. *Journal of the American Geriatrics Society*, 40: 950–957. <http://dx.doi.org/10.1111/j.1532-5415.1992.tb01995.x>
14. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.
15. Hot, P., Ramdeen, K., Borg, C., Bellon, T. and Couturier, P. (2013). Impaired Decision Making in Alzheimer's Disease. *Clinical Psychological Science*, 2(3), 328-335. <http://dx.doi.org/10.1177/2167702613504094>

16. Karlawish J. (2008), Measuring Decision-Making Capacity in Cognitively Impaired Individuals. *Neuro-Signals*, 16(1): 91-98.
<http://doi.org/10.1159/000109763>
17. Kim, S., Cox, C. and Caine, E. (2002). Impaired Decision-Making Ability in Subjects with Alzheimer's Disease and Willingness to Participate in Research. *American Journal of Psychiatry*, 159(5), 797-802.
<https://doi.org/10.1176/appi.ajp.159.5.797>
18. Lane-Brown, A. and Tate, R. (2017). Measuring apathy after traumatic brain injury: Psychometric properties of the Apathy Evaluation Scale and the Frontal Systems Behavior Scale. *Taylor & Francis Online*, 13(4), 999-1007.
<http://dx.doi.org/10.3109/02699050903379347>
19. Li, X., Lu, Z.-L., D'Argembeau, A., Ng, M., & Bechara, A. (2010). The Iowa Gambling Task in fMRI Images. *Human Brain Mapping*, 31(3), 410–423.
<http://doi.org/10.1002/hbm.20875>
20. Mohandas, E., Rajmohan, V., & Raghunath, B. (2009). Neurobiology of Alzheimer's disease. *Indian Journal of Psychiatry*, 51(1), 55–61.
<http://doi.org/10.4103/0019-5545.44908>
21. Murdock, B. B. (1972). Short-term memory. *Psychology of learning and motivation*, 5, 67-127.
22. Nathan, P., Lim, Y., Abbott, R., Galluzzi, S., Mazzoni, M., Babiloni, C., Albani, D., Bartres-Faz, D., Didic, M., Feretti, L., Parnetti, L., Salvadori, N., Müller, B., Forlani, G., Girtler, N., Hensch, T., Jovicich, J., Leeuwis, A., Marra, C., Molinuevo, J., Nobili, F., Pariente, J., Payoux, P., Ranjeva, J., Rolandi, E., Rossini, P., Schönknecht, P., Soricelli, A., Tsolaki, M., Visser, P., Wiltfang, J., Richardson, J., Bordet, R., Blin, O. and Frisoni, G. (2017). Association between CSF biomarkers, hippocampal volume and cognitive function in patients with amnesic mild cognitive impairment (MCI). *Neurobiology of Aging*, 53, 1-10.
<http://doi.org/10.1016/j.neuroimage>
23. Paula, V., Guimarães, F., Diniz, B. and Forlenza, O. (2009). Neurobiological pathways to Alzheimer's disease: Amyloid-beta, TAU protein or both? *Dementia & Neuropsychologia*, 3(3), 188-194. <http://dx.doi.org/10.1590/S1980-57642009DN30300003>
24. Raskin, J., Cummings, J., Hardy, J., Schuh, K., & Dean, R. A. (2015). Neurobiology of Alzheimer's Disease: Integrated Molecular, Physiological, Anatomical, Biomarker, and Cognitive Dimensions. *Current Alzheimer Research*, 12(8), 712–722. <http://doi.org/10.2174/1567205012666150701103107>
25. Rocca, W. A., Petersen, R. C., Knopman, D. S., Hebert, L. E., Evans, D. A., Hall, K. S., White, L. R. (2011). Trends in the incidence and prevalence of Alzheimer's disease, dementia, and cognitive impairment in the United States. *Alzheimer's & Dementia, The Journal of the Alzheimer's Association*, 7(1), 80–93.
<http://doi.org/10.1016/j.jalz.2010.11.002>
26. Ross, M. (2017). Apathy: Concept, Syndrome, Neural Mechanisms, and Treatment. *Clinical Neuropsychiatry*, 1(4), 304-314. <https://doi.org/10.1053/SCNP00100304>
27. Sinz, H., Zamarian, L., Benke, T., Wenning, G. and Delazer, M. (2008). Impact of ambiguity and risk on decision making in mild Alzheimer's disease. *Neuropsychologia*, 46(7), 2043-2055.
<https://doi.org/10.1016/j.neuropsychologia.2008.02.002>
28. Smebye, K., Kirkevold, M. and Engedal, K. (2012). How do persons with dementia participate in decision making related to health and daily care? A multi-case study.

- BMC Health Services Research*, 12(1), 241. <https://doi.org/10.1186/1472-6963-12-241>
29. Sockeel, P., Dujardin, K., Devos, D., Denève, C., Destée, A., & Defebvre, L. (2006). The Lille apathy rating scale (LARS), a new instrument for detecting and quantifying apathy: validation in Parkinson's disease. *Journal of Neurology, Neurosurgery, and Psychiatry*, 77(5), 579–584. <http://doi.org/10.1136/jnnp.2005.075929>
30. Tabert, M., Liu, X., Doty, R., Serby, M., Zamora, D., Pelton, G., Marder, K., Albers, M., Stern, Y. and Devanand, D. (2005). A 10-item smell identification scale related to risk for Alzheimer's disease. *Annals of Neurology*, 58(1), 155-160. <http://doi.org/10.1002/ana.20533>
31. Uday, U. (2015). Neurobiology of Alzheimer's Disease. *European Psychiatry*, 30(7), 1438. [https://doi.org/10.1016/S0924-9338\(15\)31112-3](https://doi.org/10.1016/S0924-9338(15)31112-3)
32. Zamarian, L., Weiss, E. and Delazer, M. (2010). The Impact of Mild Cognitive Impairment on Decision Making in Two Gambling Tasks. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 66(1), 23-31. <https://doi.org/10.1093/geronb/gbq067>

7 Appendix

7.1 Emotions Test

The interview is structured, and the questions should be posed exactly as stated. To obtain the best validity, it is not advisable to change the vocabulary or to add additional comments to the questions. Before beginning the interview, the patient has to be instructed as follows: "I am going to ask you some questions about your daily life. It is important that you base your answers on your life over the last four weeks" If the patient evokes general events or any that predate the last month, he or she must be reminded that only the current situation must be referred to: "Please try to answer according to your current way of life, by referring to the last four weeks." A precise scoring mode is proposed for each reply and should be followed as closely as possible. When an item does not apply to the patient, it is scored "0," for non-applicable (NA). When the reply is not clear at all and cannot be classified, it is also scored "0." The range of the scale is from -8 to 8. Anywhere from a -2 to a +2 is a normal score.

When you watch a movie, do you become emotional?

- No, I do not experience any specific emotion. (+1)
- No response, N/A, unable to understand (0)
- Yes (-1)

When someone tells you a joke, or you see something funny on television, do you easily laugh?

- No, I do not experience any specific emotion. (+1)
- No response, N/A, unable to understand (0)
- Yes (-1)

Do you feel happy when somebody tells you good news?

- No, I do not experience any specific emotion. (+1)
- No response, N/A, unable to understand (0)
- Yes (-1)

Do you feel sad when you hear any bad news?

- No, I do not experience any specific emotion. (+1)
- No response, N/A, unable to understand (0)
- Yes (-1)

When you have a problem (you lost your keys), does it worry you?

- No (+1)
- No response, N/A, unable to understand (0)
- Yes (-1)

When something is not working, do you give up, or do you look for a solution?

- No, I give up (+1)
- N/A (0)
- Yes, I look for a solution (-1)

When you and your family have a minor problem (lost the car keys), does it worry you?

- No (+1)
- N/A (0)
- Yes (-1)

Do you like to ask your friends or family how they feel on a regular basis?

- No (+1)
- N/A (0)
- Yes (-1)

7.2 Interview Signs

During the Interview, if you notice any of these signs, make sure to make them off

- Poor Orientation (+1)
- Increased Forgetfulness (+1)
- Language Perseverations (+1)
- Change in Personality and Emotional Status (+1)
- Social Isolation (+1)

7.3 Sniff Test

Blindfold the participant. Then present them three strips. One strip will smell like coffee beans, one like mouthwash, and one like soap. For each of these, ask the participant to describe any memories associated with the smell. For each strip, you will look for the participant to describe the who, what, where, when, and why of the memory associated with that strip. If they are successful in describing these, they gain 0 points. If they describe all but one, then they have 1 point. A point is added for every description they are unable to give. This is on a scale from 0-15. 0-3 is the range of people without MCI. 15 is the maximum; the participant cannot associate any detail with memories.

7.4 Apathy Test

For this test, you will ask these questions to the participant in an interview-like setting. For each question, you will record on a scale from 1-4 how strongly they say they exhibit this characteristic. 1 means that they strongly agree/exhibit these characteristics. 4 means they strongly disagree/do not exhibit these characteristics. The scale should be from

18-72. 72 means that they are extremely apathetic, while 18 means that they are extremely sympathetic.

1. You are interested in things
2. You get things done during the day.
3. Getting things started on your own is important to you
4. You are interested in having new experiences
5. You are interested in learning new things
6. You put little effort into anything.
7. You approach life with intensity
8. Seeing a job through to the end is important to you
9. You spend time doing things that interest her/him
10. Someone has to tell you what to do each day
11. You are less concerned about your problems than you should be
12. You have friends
13. Getting together with friends is important to you
14. When something good happens, you get excited
15. You have an accurate understanding of her/his problems
16. Getting things done during the day is important to you
17. You have initiative
18. You have motivation

7.5 Ambiguity Test through the Iowa Gambling Task (IGT)

In this part of the exam, you will take the Iowa Gambling Task. The rules are explained on the screen. After you have finished your exam, take your score, x , and plug it into the formula: $(2000-x)/(1000)$

7.6 Cumulative Formula for Overall Score

$$(E + I + A/12 + S/5 + (2000-IGT)/(1000))/(2)*$$

E = Emotion Test

I = Interview Signs

A = Apathy Test

S = Smell Test

IGT = Iowa Gambling Task

*Note: If the score you get is a decimal, round to the hundredth number.

7.7 Index for Results

Emotion Test:

Range: -8 to +8

Normal Range: -2 to +2

Interview Signs:

Range: 0 to +5

Normal Range: 0 to +1

Apathy Test:

Range: +18 to +72

Normal Range: +18 to +28

Smell Test:

Range: 0 to +15

Normal Range: 0 to +2

IGT:

Range: 0 to +4500

Normal Range: +2250 to +4500

Range: 0 to +11

Normal Range: 0 to +4

Warning Range for Testing: +4 to +6

Range for a Likely Case of Mild Dementia: +6 to +8

Requires More Physical Testing as it is likely a Severe Case: +8 to +11



The Neural Mechanism, Genetic Basis, and Possible Treatment of Procrastination

Yuyang Sun

Author Background: Yuyang Sun grew up in China and currently attends Shenzhen Middle School in Shenzhen, China. His Pioneer seminar topic was in the field of neuroscience and titled "The Decision-Making Brain."

1. Abstract

Procrastination is a prevalent problematic behavior that brings serious consequences to individuals who suffer from it²⁹. Although this phenomenon has received increasing attention from researchers, the underlying neural substrates of it are poorly studied. To examine the neural mechanism underlying procrastination, we analyze the experiments conducted by other researchers. The main results are the following: (1) Behavioral procrastination is positively correlated with the regional activity of the ventromedial prefrontal cortex (vmPFC) and the parahippocampal cortex (PHC), while negatively correlated with that of the medial prefrontal cortex (mPFC) and dorsal anterior cingulate cortex (ACC). (2) The functional connectivity between PHC and mPFC shows a positive association with procrastination. (3) Apolipoprotein E (APOE) might be an important genetic factor of procrastination. Based on the results, we also provide two possible ways to prevent behavioral procrastination, which are called reappraisal and task division. These two techniques could reduce the regional activity of PHC and thus reduce people's avoidance behavior.

2. Introduction

2.1 Organization of the Brain.

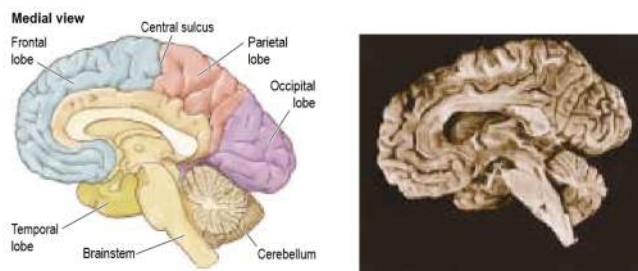


Fig 1. In the medial views of the human brain, the locations of the frontal, parietal, occipital, and temporal lobes of the cerebral hemispheres are shown, as are the cerebellum and the three major sulci (the central sulcus, lateral fissure, and longitudinal fissure) of the cerebral hemispheres.

The human brain is divided into 3 parts: cerebrum, cerebellum, and brainstem. The cerebrum is the largest part of the brain and is responsible for the voluntary action of our body. It can be further divided into lobes, which contains cortices that are responsible for different outside stimuli. In our study, we focus on some cortices that are associated with behavioral procrastination and investigate how they interact with each other.

In order for readers to better understand this paper, we provide a diagram explaining the anatomical directions relative to our head and brain. Since our paper includes terminology such as “dorsal anterior cingulate cortex” and “ventromedial prefrontal cortex,” this diagram could help our reader to have a clearer idea of where the cortices are located and the regional connection between these cortices.

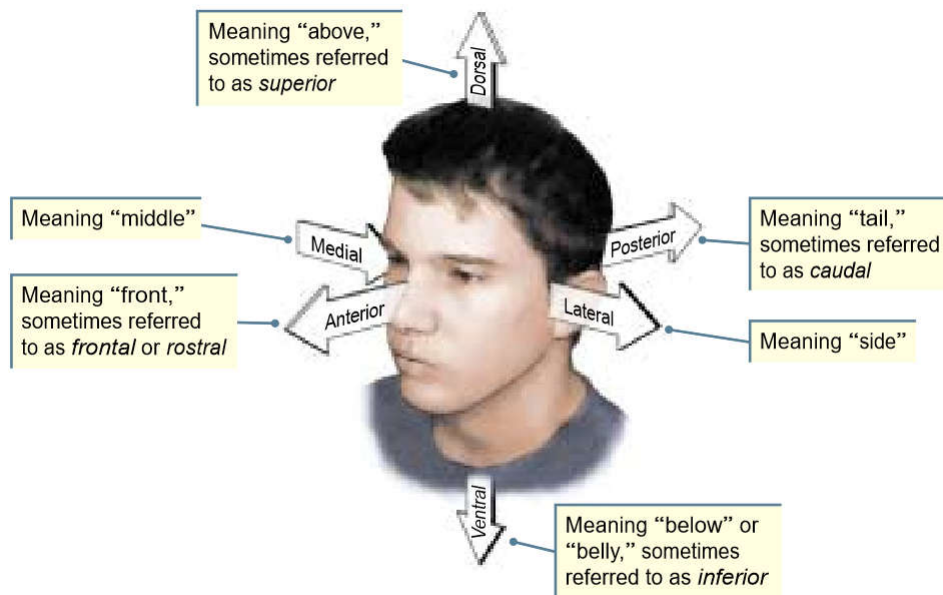


Fig 2. Anatomical directions relative to the head and brain.

2.2 Brief Introduction to Procrastination

Procrastination – the voluntary but irrational delay of an intended course of action – is a widespread phenomenon^{13,19,24}. It is harmful to our psychological, physical and financial well-being. Studies have shown that more than 75% of college students procrastinate, and more than 20% of populations are affected by chronic procrastination¹⁶. Some people even view procrastination as an unavoidable behavior. Existing evidence has documented that procrastination is associated with poor mental health and low life satisfaction⁶. The extensive literature concerning procrastination has attributed this phenomenon to cognitive and affective factors²⁵. Several models have been proposed to interpret the psychological mechanisms underlying procrastination, which predominantly characterize procrastination as a failure of temporal-based self-control that might be the result of problematic inhibition of the motive force of emotion¹⁵. Sirois proposed a cognitive escape hypothesis, which states that procrastinators manifest avoidant cognitive tendencies to promote immediate emotion regulation (an intuitive reaction to pacify their emotion when

people feel anxious) at the cost of long-term goals²⁹. However, despite the increasing attention from researchers on procrastination, the neural bases of this phenomenon have been understudied. Therefore, we decided to use a “behavior-brain-gene” mode to uncover the neural substrate and genetic bases underlying procrastination, which could provide some insight into how to avoid such behavior.

3. Hypothesis

I hypothesize that procrastination is correlated with the hyper-activity in the vmPFC and PHC through hampering the top-down control signals of the prefrontal cortex. Medial prefrontal cortex (mPFC), along with dorsal anterior cingulate cortex (ACC), is responsible for regulating such behavior.

Within the context of the program, it is not possible to conduct original experiments. Therefore, this paper sums up and analyzes the experiments conducted by other researchers, along with providing two possible ways to prevent procrastination

3.1 mPFC and ACC in Cognitive Control

Human neuroimaging studies suggest that the medial prefrontal cortex (mPFC), including the dorsal anterior cingulate cortex (ACC), along with other brain structures, are involved in differential processing of unfavorable outcomes (Fig. 3B). These include studies using monetary rewards and punishments and studies using abstract performance feedback^{17,22}. Similar parts of the mPFC are activated by primary reinforcers such as pain affect and pleasant tastes, suggesting that the mPFC plays a general role in coding the motivational value of external events.

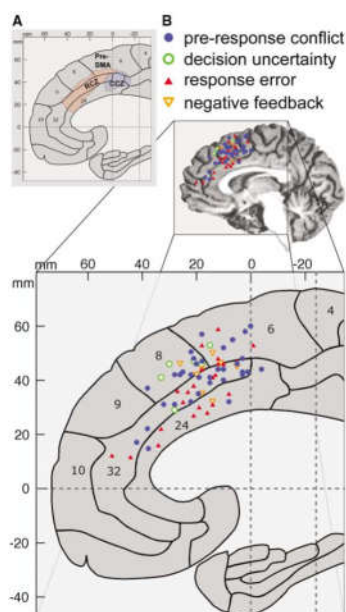
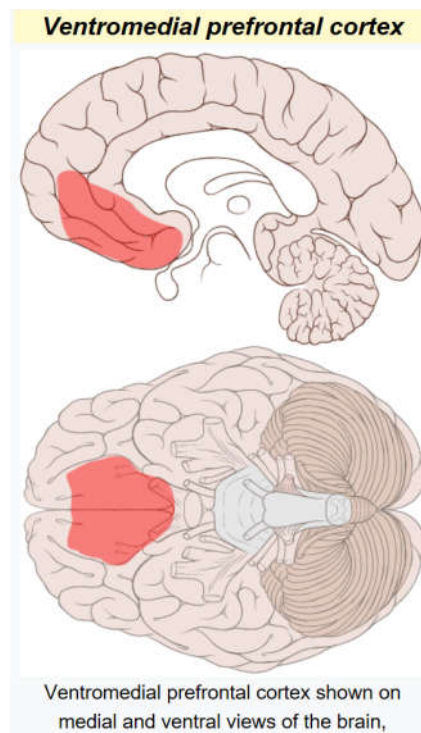


Fig. 3. Areas in the medial frontal cortex involved in performance monitoring.

Electrophysiological recordings in humans have identified the purported event-related brain potential associated with the mPFC response to unfavorable outcomes: the

feedback-related error-related negativity (or “feedback ERN”)²⁶. This negative-polarity voltage deflection peaks at approximately 250 to 300 ms after a stimulus indicating the outcome, and has greater amplitude for negative performance feedback and outcomes that indicate monetary losses than for positive feedback and monetary gain⁹. The timing of this brain potential suggests that the mPFC computes or has access to a rapid evaluation of the outcome stimulus. Furthermore, previous studies indicate that the amplitude of the feedback ERN shows a graded sensitivity to the value of outcome stimuli that is normalized with respect to the subjectively expected outcome value (mean) and experienced range of outcome values (variance)⁸.

3.2 PHC and vmPFC in Episodic Propection and Decision-Making.



The ventromedial prefrontal cortex is a part of the prefrontal cortex in the mammalian brain. It plays an important role in the process of decision making. Studies have shown that patients with bilateral lesions of the vmPFC develop severe impairments in personal and social decision-making even though most of their intellectual ability is preserved⁷. For instance, they have difficulty choosing between options with uncertain outcomes, regardless of whether the uncertainty is in the form of a risk or of an ambiguity. After developing a lesion, these patients have difficulty learning from their mistakes, making the same decisions again and again even though they lead to negative consequences. These patients choose alternatives that give immediate rewards, but they often ignore the future consequences of their actions⁵. Therefore, vmPFC is proposed to be responsible for determining the possible value of an action and for choosing the behavior that gives the greatest reward.

Previous studies have demonstrated the engagement of the parahippocampal cortex (PHC) in the processing of episodic memory and emotional stimuli. Aminoff et al. proposed an integrative account, which states that the PHC is sensitive to contextual association and reflects activation of the relevant stored contextual representation¹. According to this account, the PHC could function to incorporate the current context with long-term associations of the context built up in memory. This is consistent with the findings that the PHC is a key region responsible for episodic or semantic prospection¹. Intriguingly, existing evidence has showed that episodic future thinking is associated with procrastination, and imaging negative future events make individuals inclined to choose immediate rewards²¹. Combined, the positive correlation between the regional activity of the PHC and procrastination implies that procrastination is accompanied by an increased negative episodic prospection encoded in the PHC.

3.3 Activity in PHC and vmPFC inhibits the top-down control of the mPFC and ACC

In a study by Chenyan Zhang, Yan Ni and Tingyong Feng from Southwest University in China, 66 right-handed college students were recruited in order to identify the neural mechanism responsible for behavioral procrastination¹⁰. All participants completed the resting state fMRI scan before behavioral measures, which contained the General Procrastination Scale. The General Procrastination Scale (GPS) (see appendix) designed by Lay Clarry was used to measure trait-procrastination. GPS contains 20 items involving learning activities and daily life behavior (e.g., “I often find myself performing tasks that I had intended to do days before”; “I do not do assignments until just before they are to be handed in,” etc.). All items were rated on a 5-point Likert-type scale with the response alternatives anchored at the ends with 1 (extremely uncharacteristic) to 5 (extremely characteristic). The aggregate scores were computed by summing across responses to each item where higher scores indicate greater procrastination. Previous studies have demonstrated that the GPS has satisfactory internal consistencies²⁰.

The resting-state functional connectivity result of their study shows that procrastination can be predicted by the functional connectivity that uses vmPFC and PHC as seeds. Functional connectivity results show that procrastination scores are positively correlated with functional connectivity between mPFC and PHC, suggesting that excessive episodic thinking, especially negative episodic thinking, can inhibit the cognitive monitoring of mPFC over goal-direct behaviors.

Intriguingly, the PHC and the vmPFC are core regions involved in decision making that interact with each other to facilitate decision-making based on mnemonic scene construction. In procrastination, greater activity of the PHC resulted from negative future thinking might bias the decision-making processing in the vmPFC towards immediate satisfaction. This process would further affect the activity of several other regions. In particular, greater negative vmPFC-seed connectivity with the dorsal medial prefrontal cortex (dmPFC), the inferior frontal gyrus (IFG) and some others is associated with an increase in procrastination tendency²⁹. Empirical evidence has demonstrated that the dmPFC and the IFG are core regions involved in cognitive control. Specifically, the dmPFC is engaged in the conceptualization and evaluation of self-referential stimuli. Recent studies have proposed that the dmPFC plays a critical role in intentional control, supporting top-down emotion regulation and voluntarily refraining from planned behaviors. These findings suggest that the dmPFC is a key region that exerts internal cognitive control during the decision-making process. In addition, some pooled studies have documented the involvement of the bilateral IFG in inhibitory and attention control. The present findings -- that increased negative correlation between the vmPFC and these regions gave rise to

procrastination -- indicated that regional activity of the vmPFC could hamper the activity of the prefrontal cortex, and thus override the top-down control signals to focus on short-term satisfaction (e.g., mood repair)². This is consistent with the argument that in behaviors such as drug addiction, strong bottom-up signals triggered by the impulsive system could hijack the goal-driven cognitive signal and result in a failure of self-control².

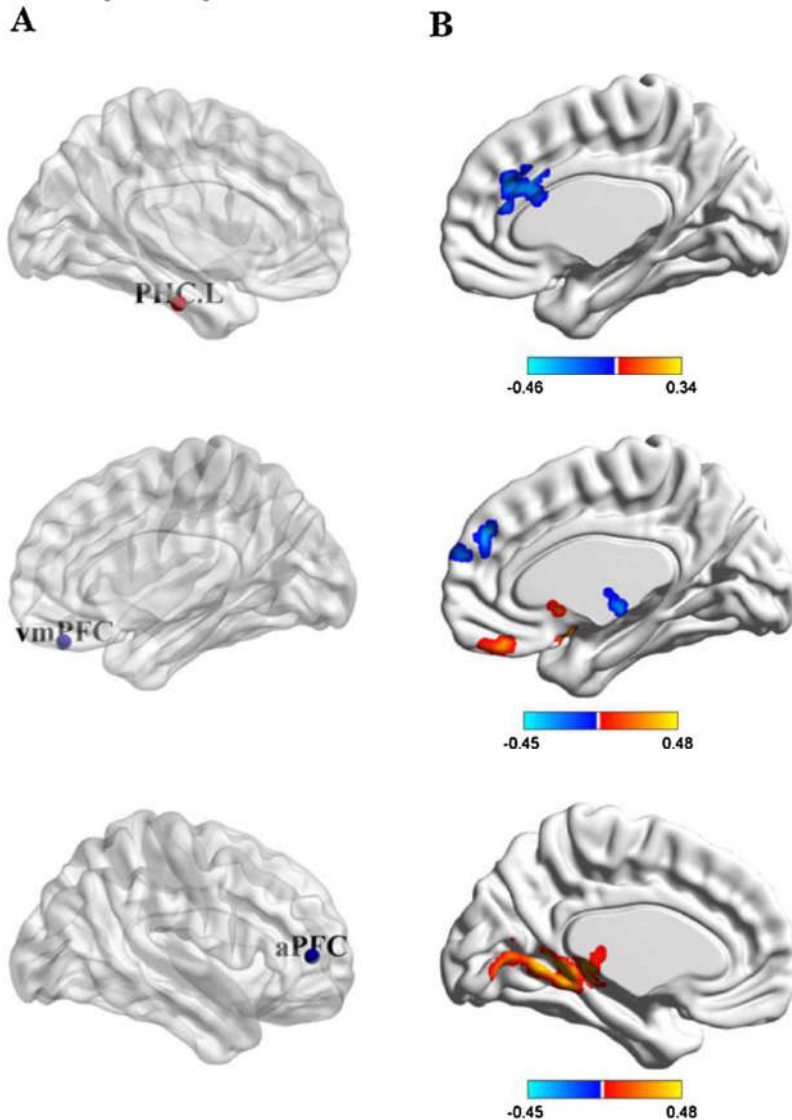


Fig 4. Resting-state functional connectivity results.

A: defined spherical seed regions of interest (ROI) (diameter = 6 mm);

B: functional connectivity between seed ROIs and other brain regions was significantly correlated with procrastination ($p < 0.05$; Alphasim corrected, cluster size > 85).

3.4 Genetic Basis of Procrastination

In order to discover the relationship between procrastination behavior and a genetic mechanism, Arvey, Rotundo and Johnson investigated male twins that are raised in the same environment (including 118 monozygotic twins and 93 dizygotic twins)⁴. The result suggests a similarity of procrastination behavior among monozygotic twins and dizygotic twins, which implies that there are some genetic bases underlying procrastination. Recently, Junlin Shen and Wen Qin investigate Modulation of the Apolipoprotein E (APOE) and sortilin-related receptor (SORL1) genes on hippocampal functional connectivity in healthy young adults¹⁸. In their studies, a total of 287 healthy, young, right-handed subjects (134 males and 153 females; mean age: 22.7 ± 2.4 years, ranging from 18 to 29 years) were selected from 323 subjects who participated in their study¹⁸. Their genomic DNA was extracted from 3000 μ l of the whole blood using the EZgeneTM Blood gDNA Miniprep Kit (BiomigaInc, San Diego, CA, USA). The standard protocols were used to genotype SORL1 rs2070045 and APOE, and Resting-state fMRI data were obtained using Gradient- Echo Single-Shot Echo-Planar Imaging sequence¹⁸.

Using parametric ANCOVA, the main effect of APOE was found in positive right hippocampal rsFC with the posterior cingulate cortex (PCC) (Fig. 5a), prefrontal cortex (Fig. 5b) and bilateral sensorimotor cortices (SMC) (Fig. 5c, d), and left hippocampal rsFC with the dorsal anterior cingulate cortex (dACC). Since the connectivity between PHC (part of the hippocampal cortex) and PFC are positively correlated to procrastination, and the connectivity between PHC and dACC is negatively related to procrastination, this finding suggests that the Apolipoprotein E (APOE) might have some influence on one's procrastination behavior.

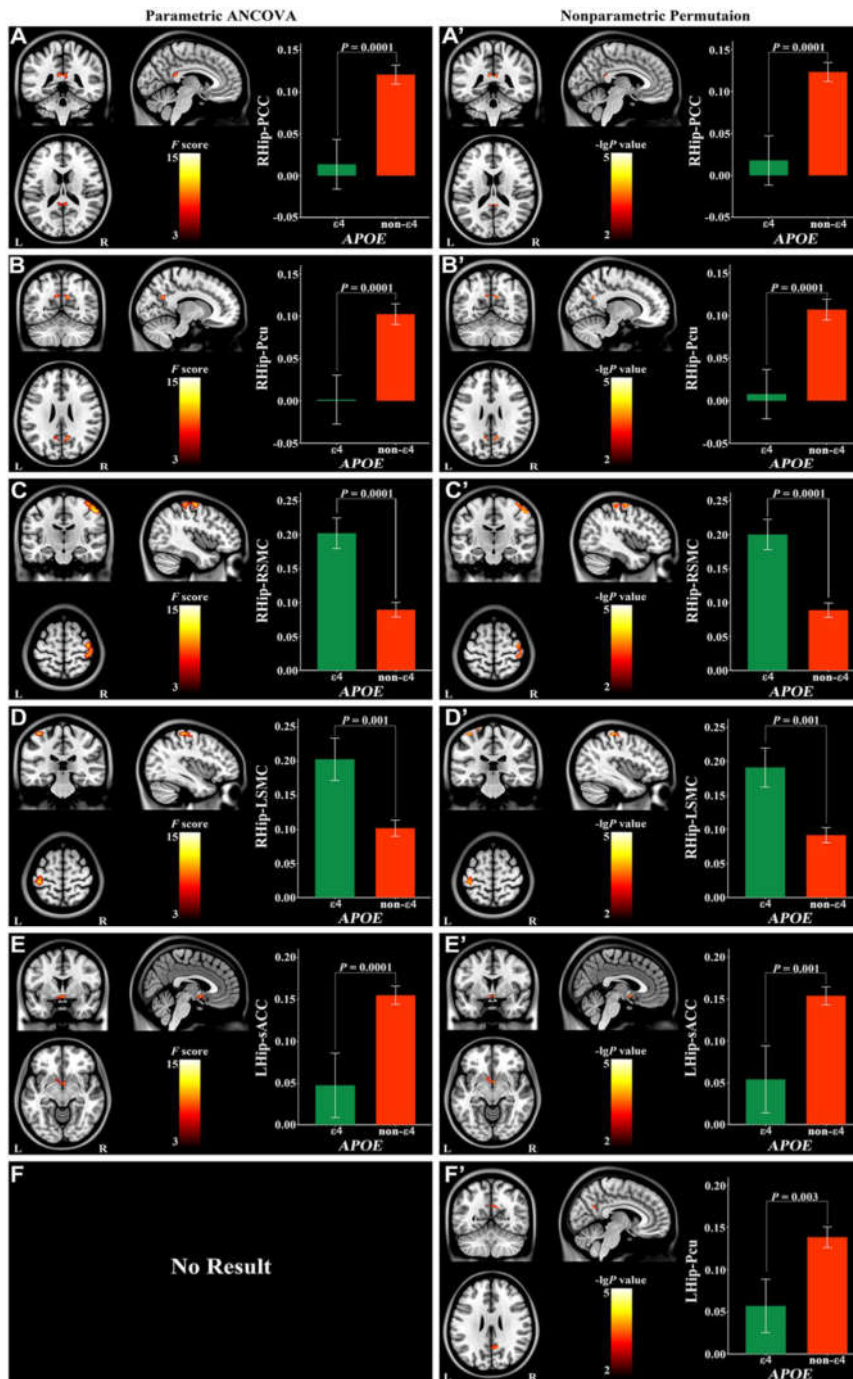


Fig. 5 Main effect of APOE on hippocampal positive connectivity. Hip hippocampus, L left, PCC posterior cingulate cortex, PFC prefrontal cortex, R right, dACC dorsal anterior cingulate cortex, SMC sensorimotor cortex.

4. Discussion

Combining those findings, we suggest that cognitive control failures correlated with procrastination, reflected in the overwhelming priority of the PHC and vmPFC activity over inhibitory control exerted by the prefrontal cortex, and reflected in the absence of the mPFC control, are the main reasons for procrastination. This is consistent with the balance model of self-control, which suggests that self-control fails once impulses overwhelm prefrontal control or PFC function is impaired. Note that the fMRI results in the experiment performed by Zhang et al also showed a negative correlation between the regional activity of the ACC and procrastination. Considering the engagement of the ACC in reward and value evaluation, it might be that greater procrastination is associated with lower personal value. Whilst the ACC is of heterogeneity, different parts of the ACC function differ according to task demands. Therefore, further evidence is needed to interpret the role of the ACC in procrastination. Their resting-state fMRI results demonstrated that the key process underpinning procrastination is the trade-off between cognitive control and affective processing. Procrastination occurs while the mPFC loses control over the PHC, and hyperactivity of the PHC and the vmPFC biases one's attention towards immediate/short-term satisfaction and overrides the activity of brain regions responsible for inhibiting internal/external distracting stimuli¹⁰. In this case, our findings reveal the neural substrates underlying procrastination and provide new perspectives to understand the mechanisms behind this phenomenon.

5. Possible Remedies for Procrastination

5.1 Technique of Reappraisal

One possible way to deal with procrastination is by using the "reappraisal" technique. It is a well-known emotion regulation strategy to help people think more realistically and thereby reduce people's anxiety about future events. In a study performed by Shinpei Yoshimura, 15 different participants were divided into 3 groups. All of them were told to view a countdown from 10 to 1, but only one group was told to use the technique of reappraisal. All the participants were asked to report their subjective fear rating, and they used functional magnetic resonance imaging (fMRI) to investigate neural activity associated with the process. The results showed that the participants who used the technique of reappraisal experienced less anxiety, since the brain region responsible for anxiety was not much involved³⁰. In addition, recent neuroimaging studies have revealed that prefrontal regions, including the medial, orbital, and lateral prefrontal cortices as well as the anterior cingulate region, are involved in the regulation of negative emotion¹⁴. In this experiment, these regions were all involved in the process of reappraisal, suggesting that reappraisal is a useful emotional regulation strategy. The fMRI results also revealed decreased thalamus, insula, and amygdala activity in the group who used the technique of reappraisal. These regions have been implicated in prior studies of anticipatory emotion, including anticipatory anxiety and negative episodic future thinking, which arises when one experiences a noxious stimulus²⁷. The finding suggests that reappraisal could be a useful way to reduce people's anxiety state about future events and therefore reduce people's avoidance behavior, since behavioral procrastination is positively correlated to negative episodic future thinking.

5.2 Task Division

Another possible way to deal with procrastination is by dividing a long-term task into some short-term tasks. As mentioned before, procrastination occurs while the regional

activity of the PHC and the vmPFC biases one's attention towards immediate/short-term satisfaction and overrides the activity of brain regions responsible for inhibiting internal/external distracting stimuli¹⁰. Facing a long-term task might lead to negative episodic future thinking and thus inhibit the top-down control from medial prefrontal cortex. Also, when people face difficult, long-term tasks that require a great amount of time to accomplish, ACC might underestimate the value of performing the task. In contrast, when people divide a task into small goals, the prefrontal cortex and dorsal anterior cingulate cortex can better determine the value and reward of finishing the task. Therefore, it could be easier for the PFC and ACC to exert cognitive control and to monitor our performance. Also, this technique could help us focus on the task we are working on and thus prevent us from negative future episodic thinking. In short, our brain is more efficient in regulating a short-term task than a long-term task.

6. Conclusion

We have provided an overview of the evidence suggesting vmPFC and PHC are the key regions involved in behavioral procrastination. Also, through the understanding of the neural mechanism underlying procrastination, we provide two possible ways to reduce our tendency to procrastinate. By analyzing the functional connectivity between PHC and mPFC, we hypothesize that Apolipoprotein E (APOE) might be an important genetic factor of procrastination.

Future research should provide a more detailed analysis of how the brain works when people are trying to regulate their procrastination behavior. This is very important since it could provide us some insight into what kind of stimulus is efficient in regulating such behavior. Also, the gender difference should be studied in order to provide us with a more thorough understanding of procrastination.

7. References

1. Aminoff, E. M., Kveraga, K. & Bar, M. The role of the parahippocampal cortex in cognition. 2013, *Trends Cogn. Sci.* 17, 379–390, doi:10.1016/j.tics.2013.06.009.
2. Antoine, B. Decision making, impulse control and loss of willpower to resist drugs: a neurocognitive perspective. 2005, *Nat. Neurosci.* 8, 1458–1463.
3. Ariely, D. & Wertenbroch, K. Procrastination, deadlines, and performance: Self-control by precommitment. 2002, *Psychol. Sci.* 13, 219–224.
4. Arvey, R. D., Rotundo, M., Johnson, W., & McGue, M. The determinants of leadership: The role of genetics and personality. 2003, April. *Industrial and Organizational Psychology*, Orlando, FL.
5. Bechara, A; Tranel, D; Damasio, H. "Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions". 2000, *Brain.* 123 (11): 2189–202. PMID 11050020. doi:10.1093/brain/123.11.2189.
6. Caldwell, L. & Mowrer, R. The link between procrastination, delay of gratification, and life satisfaction: A preliminary analysis. 1998, *Psi Chi Journal of Undergraduate Research* 3, 145–150.
7. Carlson, Neil R. *Physiology of Behavior*. 2013, 11th ed. Boston: Pearson.
8. C. B. Holroyd, J. T. Larsen, J. D. Cohen, *Psychophysiology.* 41, 245 (2004).
9. C. B. Holroyd, M. G. H. Coles, *Psychol. Rev.* 109, 679 (2002).
10. Chenyan Z, Yan N, Tingyong F. The effect of regulatory mode on procrastination: Bi-stable parahippocampus connectivity with dorsal anterior cingulate and medial prefrontal cortex. *Behavioural Brain Research* 329 (2017) 51–57.
11. Davachi, L., Mitchell, J. P. & Wagner, A. D. Multiple routes to memory: distinct

- medial temporal lobe processes build item and source memories. 2003, *Proc. Natl. Acad. Sci. USA* 100, 2157–2162, doi: 10.1073/pnas.0337195100.
12. Díaz-Morales, J. F. & Ferrari, J. R. In *Time Perspective Theory*. 2015, Review, Research and Application 305–321.
 13. D.E. Gustavson, et al., Genetic relations among procrastination: impulsivity, and goal-management ability implications for the evolutionary origin of procrastination, *Psychol. Sci.* 25 (6) (2014) 1178–1188.
 14. Golkar A, Lonsdorf TB, Olsson A, Lindstrom KM, Berrebi J, et al. 2012, Distinct contributions of the dorsolateral prefrontal and orbitofrontal cortex during emotion regulation.
 15. Gifford, A. Emotion and self-control. 2002, *Journal of Economic Behavior & Organization* 49, 113–130.
 16. Harriott, J. & Ferrari, J. R. Prevalence of procrastination among samples of adults. 1996, *Psychol. Rep.* 78, 611–616.
 17. J. O'Doherty, M. L. Kringelbach, E. T. Rolls, J. Hornak, C. Andrews. *Nature Neurosci.* 4, 95 (2001) J.R. Ferrari, *Still Procrastinating: The No Regrets Guide to Getting It Done*, John Wiley & Sons, 2017.
 18. Junlin S, Wen Q, Qiang X, Lixue X, Jiayuan X, Peng Zhang, Huaigui L, Bing L, Tianzi Jiang, Chunshui Y. Modulation of APOE and SORL1 genes on hippocampal functional connectivity in healthy young adults. *Brain Struct Funct* (2017) 222:2877–2889.
 19. J.R. Ferrari, *Still Procrastinating: The No Regrets Guide to Getting It Done*, John Wiley & Sons, 2010, 2017.
 20. Lay, C. At last, my research article on procrastination. 1986, *Journal of Research in Personality*, 20, 474–495.
 21. Liu, L., Feng, T., Chen, J. & Li, H. The value of emotion: how does episodic prospection modulate delay discounting? 2013, *PloS One* 8, e81717.
 22. M. Ullsperger, D. Y. von Cramon, *J. Neurosci.* 23, 4308 (2003).
 23. Piers S. *The Nature of Procrastination: A Meta-Analytic and Theoretical Review of Quintessential Self-Regulatory Failure*. American Psychological Association. 2007, Vol. 133, No. 1, 65–94.
 24. P. Steel, *The Procrastination Equation: How to Stop Putting Things off and Start Getting Stuff Done*, Random House Canada, 2010.
 25. Rebetz, M. M. L., Rochat, L. & Van der Linden, M. Cognitive, emotional, and motivational factors related to procrastination: A cluster analytic approach. 2015, *Pers. Individ. Dif.* 76, 1–6.
 26. Richard R, Markus U, Eveline A, Sander Nieuwenhuis. The Role of the Medial Frontal Cortex in Cognitive Control. 2004, *Science*.
 27. Sarinopoulos I, Grupe DW, Mackiewicz KL, Herrington JD, Lor M, et al. 2010. Uncertainty during anticipation modulates neural responses to aversion in human insula and amygdala.
 28. Sirois, F. M. Absorbed in the moment? An investigation of procrastination, absorption and cognitive failures. 2014, *Pers. Individ. Dif.* 71, 30–34.
 29. Wenwen Z, Xiangpeng W, Tingyong F. Identifying the Neural Substrates of Procrastination: a Resting-State fMRI Study. 2016, *Nature. Scientific Reports* 6:33203 DOI: 10.1038/srep33203
 30. Yoshimura S, Okamoto Y, Yoshino A, Kobayakawa M, Machino A, et al. Neural Basis of Anticipatory Anxiety Reappraisals. 2014, *PLoS ONE* 9(7): e102836.

The Power and Significance of Agency in *The Iliad*

Yumeng Li

Author Background: Yumeng Li grew up in China and currently attends Beijing National Day School in Beijing, China. Her Pioneer seminar topic was in the field of literature and was titled "The Hero in Literature and Culture."

Introduction

The whole narrative of Homer's *The Iliad* appears to be dominated by the idea of fate. Even the gods, enjoying the privilege of being immortal, sometimes surrender to the fated events delivered in prophecies, or attempt to use their own will to control events. Heroes, particularly the literary characters with exceptional qualities in this epic, appear to be more restricted in actions under the control of fate.

The poem opens with the most famous lines:

Rage—Goddess, sing the rage of Peleus's son Achilles,
murderous, doomed, that cost the Achaeans countless losses,
hurling down to the House of Death so many sturdy souls,
great fighters' souls, but made their bodies carrion,
feasts for the dog and birds,
and the will of Zeus was moving towards its end (Homer 77).

The whole narrative is built on "the will of Zeus," suggesting that the gods' will is deeply rooted in the Trojan War and every event is inescapable. A number of previous scholars, including Voigt and Snell, have claimed that heroes are deprived of full autonomy under the dominant power of gods, even if they do act independently sometimes (qtd. in Gaskin 7).

Fate is a mysterious protocol that determines a character's outcome in advance of the character's final day, usually regarded as unalterable by people. The idea of fate frequently appears in literary works, especially those with divinities. As a reader, I wonder why the characters exist and interact with other characters when their fate is already decided at the moment of their birth. In my understanding, characters who can think, talk, act, and react should be more than puppets of other factors in the literature, especially puppets of immortal beings and fate. Rather, poets or writers are humans who create work that reflects their sense of humanity and more or less opens up human possibility. In Homer's *The Iliad*, the heroes can also reflect humanity and agency.

The heroes, or agents, have some degree of agency, which is to have "a self which determines, rather than is determined, to action," and "the self arrives at this determination by considering available reasons for action in the light of its overall purposes" (Gaskin 1). Richard Gaskin's interpretation of agency is a highly refined and accurate interpretation of the Homeric heroes in *The Iliad*. The ability to decide for themselves is the most important and seemingly hidden characteristic of heroes. Some of them may decide to surrender to

fate, while others embrace it. Confrontations with fate also allow heroes to prove their existence as true heroes by displaying their heroism—an individual's commitment to a noble purpose while willing to accept the consequences of achieving that purpose (Jayawickreme and Di Stefano 165). Ultimately, heroic actions stand from the heroes' ability to choose from available options and define heroes of themselves in this way, despite the control of the gods' will and the presence of fate.

I.

Filtering out the complicated factors of fate and the gods' will, the heroes consciously face choices and choose for themselves. On several occasions, there are several choices that lie in front of them, so the heroes are able to weigh the benefit and cost of them to choose wisely and choose the one that conforms to their personal values.

Achilles, as one of the major heroes in the poem, exhibits his excessive rage from the beginning to the very end of the work. While his rage keeps him from participating in the battles and helping the Achaeans during hard times, it also enhances Achilles' agency. Achilles, as informed by his mother, Thetis, has two ways to live his life when Agamemnon asks him for help: he can remain in Troy, so "[his] journey home is gone, but [his] glory never dies," or he can go back home and give up his pride and glory, where "the stroke of death will not come on [him] so quickly" (Homer 265). Driven by his desire for life over glory and his unquenchable anger, Achilles chooses to turn Agamemnon's invitation down and witness the massive deaths of the Achaeans. The anger of Achilles functions as a crucial motivation of his decision that cannot be intervened by other forces, as this rage originates from himself. He has complete autonomy to decide which action he is about to take, and the incentive to act is from his own human instinct.

Intriguingly, Achilles' human emotions are more detectable when he is not so heroic, which further enhances his agency. His pettiness shows much of his human side. When he is enraged by Agamemnon, he exhibits no sign of tolerance to the loss of Briseis, which equals to loss of his own honor. Achilles' reactions are not so surprising, as the desire for honor is deeply rooted in ancient Greek culture and can be seen in many literary works. As seen in Sophocles' play *Antigone*, Antigone also shares a similar point of view claiming that "the worst can befall is but to die an honorable death" (Sophocles 312). The rage is a heroic exhibition of Achilles, but he later beseeches his mother Thetis to "go to Olympus, plead with Zeus, if [she] ever warmed his heart with a word or any action" (Homer 90). At this point, Achilles' personal values drive him to regain his lost honor, but he chooses a more emotional way to achieve the goal. Being petty and begging his mother for help, Achilles displays the opposite side of his heroism according to his preference.

Achilles' rage even drives him to attempt the murder of Agamemnon after the great king takes away his war prize, Briseis. Furious and uncontrollable, Achilles draws out his spear and points it to Agamemnon, only to be prevented by a force exerted by the goddess Athena that holds back his hair. Persuaded by Athena's words, Achilles complies and holds back his weapon. Although it may seem that Achilles is under the thumb of Athena and the will of Mount Olympus, he is not compelled to obey their rules. Obeying the gods' order is a reasonable decision Achilles makes to compromise in the situation. As pointed out by previous scholars like Willcock and Jones, "the goddess can advise but she does not compel: the decision and responsibility remain with Achilles" (qtd. in Jones 110). Despite the fact that the gods' will cannot be compared to the same level as fate, agency is not taken away from the heroes. The power of free choice is as strong as it is in any other occasions.

However, Achilles decides to engage in the Trojan War at last to revenge his best friend, Patroclus. At this point, no one threatens him to fight for the Achaeans or restricts

him from any kinds of actions. He has a clear view that he is destined to live a short life “with heartbreak” (Homer 91). When the death of Patroclus penetrates Achilles’ heart like the sharpest arrow, Achilles knows that he can no longer stay behind the safe walls and keep himself from the battlefield and that he must stand up as a hero and fulfill his responsibilities. He decides to stop raging and “let bygones be bygones” (Homer 490). Abandoning the previous feud and all the pettiness, Achilles’ figure grows lofty and grand: he holds back his rage and beats down his anguish, hoping to do what is best for the Achaeans, rather than his own benefit.

Not all heroes are making decisions with the desire to contribute positively to society. Some of them make readers question their decisions and values, but they are also making their own choice instead of being dominated by others. Paris, who is famous for his judgment of the fairest goddess and the resulting Trojan War, chooses to cowardly hide within the walls of Troy. Because of this, Paris is often referred to as “a pretty boy” by people, since he does not fulfill his duty as the prince of Troy and makes others bear the responsibilities of fighting a war caused by his behavior. Homer makes Paris’ heroism ambiguous to readers. When Menelaus invites him to “fight it out for Helen and all her wealth in a single combat,” (Homer 131) hoping to end the years-long agonies of war, Paris is reluctant to meet his duty. Paris also has to bear the scorn of his wife, Helen and his brother, Hector. Hector points out that the Trojan people are “dying around the city, the steep walls, dying in arms—and all for [Paris]” (Homer 206). Paris is finally convinced by Helen and returns to the battlefield. During the duel with Menelaus, however, something noticeable happens: Aphrodite, who is on the side of the Trojans, draws Paris back from the battle and makes him magically disappear in front of all the Achaeans and Trojans. His potential fate of death is disrupted by a goddess, so it undermines his courage in a way that also weakens his agency. Although Paris seems less motivated than Achilles in terms of choice, he indeed is endowed with the power of agency. Despite his involvement in the “judgment of Paris,” the divine powers do not have any significant role in these decisions of Paris. He is motivated to his actions, no matter how cowardly or unheroic they are.

Readers sometimes take it for granted that the narrator has already designed every scenario happening in the plot, but all of the characters are not just plot tools with different names and a series of actions. Literary works are intended to show readers something, preferably something that initiates reflection and thought. Therefore, the characters have to possess the basic qualities of human beings, including autonomy. The heroes in this epic poem, besides Achilles and Hector, all make choices at some point during their lives. It may be choosing to fight in the war, though Agamemnon is not of a higher rank than other leaders and they voluntarily fight for the sake of the Achaeans; it may be choosing whether to retreat from the war because of fear, as death is sweeping the battlefield and the desire for life against glory can determine one’s choice; it may be choosing whether to kill someone they capture, questioning the meaning of life and the use of war.

The list of heroes’ full autonomy goes on, and people may argue that those are only the few times fate fails to dominate. Indeed, when fate is clearly present, the heroes’ actions and behaviors can be analyzed in a more critical way.

II.

Fate is “the idea that what happens in some sense has to happen” (Solomon 435). As its intrinsic meaning implies, fate may not be prevented or changed. With the power of fate so strong and dominant in *The Iliad*, how can the heroes possibly make decisions for themselves? The heroes do not necessarily craft their lives in the way they want, but in the presence of fate, they have the free will to choose between believing in and surrendering to

fate or acting against it. In other words, they make choices with regard to the knowledge of fate, rather than making no choices because of fate.

One of the sad things about the Trojan War is that one of the sides must win and the other must lose. The heroes are young lives fighting for their side, and the cruel reality is that most of them will die during battle. Especially when the gods have already chosen a side to favor, or the prophecy has already announced the winner, the other one is then already doomed, despite the fact that the day has not yet arrived. It would seem to the heroes that no matter what they do, they will lose their lives and their city will fall. The paths seem so clear that people may fail to see roads in between. The doomsday is only a result, while the paths to reach the results are entirely dependent on the agents themselves. Will they die in a heroic way refusing to comply with fate, or will they sit on their fate and wait for their final day? The question is to be answered by the heroes themselves.

Hector is a flawed hero who is frequently criticized for committing wrong deeds and making bad decisions. One of the deadliest mistakes he makes is killing Patroclus and stripping off his armor, which belongs to Achilles. With the aide of Apollo, Hector stabs Patroclus and takes the young hero's life from his once living body. He then grows overwhelmingly proud and claims to strike Achilles by his spear first and kill him (Homer 440). In the later books, however, we know that Achilles is able to give Hector the deadly shot through the weak spot on his familiar armor. Moreover, his cowardice makes him flee from Ajax in Book 17 twice, and his foolish decision of ordering the Trojan army to camp outside the city causes the irreversible downfall of Troy. However, Hector should receive some merit for his understanding of fate and how he chooses to react to it, which is genuinely admirable.

Hector, the leader and soul of the Trojans, is a hero that does not surrender to fate. He acknowledges the existence of fate, but he does not feel frightened, scared, or "oppressed by that knowledge" (Jones 115). When Hector goes back home to Troy in Book 6, his wife, Andromache, displays great concern because of Hector's determination to run towards his fate. She sincerely pleads with Hector to think about the family, especially their "helpless son" (Homer 209). Andromache's speech to Hector is one of those few moments when marginal characters get the chance to express their opinions. She warns Hector that his "furious courage will destroy [him]" and the fate will weigh her down (Homer 209). Hector, who has a deep love for his family, refutes his wife's speech directly without a second thought:

No man will hurl me down to Death, against my fate.
And fate? No one alive has ever escaped it,
Neither brave man nor coward, I tell you—
it's born with us the day that we are born. (Homer 212)

Hector's attitude towards fate is mostly positive, and he accepts fate as a basic attribute of life. He understands that any attempt to alter what is supposed to happen is vain and futile, while decisively running towards fate is a better choice—also the choice Hector makes.

No doubt, Hector is one of the greatest warriors in the battle, and he takes numerous lives. Sadly, Hector is also fated to die during his battle with Achilles, but his calm perception of fate has enabled him to gain the power of real heroes. Although his mother and father have tried multiple times to prevent him from confronting Achilles, Hector grows to be less critical about the loss of life, and he values more his own glory and the representation of Troy. With Achilles roaring outside the walls of Troy, Hector cannot

remain inside the wall for safety and longevity—things a hero cannot dream of when they stand on a high moral point. Fighting for Troy is his responsibility, and facing Achilles is the challenge he must accept given his personal integrity. Hector is determined, and no one can “shake the fixed resolve of Hector” (Homer 545).

The death of Hector is completely determined when he faces Achilles, and the appearance of Athena only adds to the dramatic effect. Hector embraces his fate in a heroic way that shows he clearly knows the outcome (Homer 553), but still stands up and faces fate directly. His heroism would be remembered by every Trojan citizen because he is the man that dies during a combat defending Troy while facing such a fierce enemy like Achilles.

III.

Discussing the power of agency in *The Iliad* is not merely showing and digging through the heroes’ actions to argue against fate. Instead, the extent of agency displays ideas that are of more importance for interpreting the epic from a different perspective. Heroes are mortals, even if some of them may have a god or goddess parent. They are human beings with emotions, motivations, and minds. In a way, they are the same as we are; however, on a more precise scale, they hold their own uniqueness because of the actions they choose to take despite those fated outcomes. Holding the power of agency, heroes exhibit the beauty of humanity and heroism, and therefore define themselves and stand up as real heroes.

Heroism is the quality heroes possess, but each hero may have a different interpretation of how they want to display heroism, or generally how they behave heroically. Heroism, by Stevanović’s interpretation, is the pursuit of eternal fame and glory through brave death (13). There is no hero that can possibly be a hero without dying bravely in a war, and it is quite true in an ancient society where war is the only opportunity for heroes to exhibit their bravery and unique combat skills.

Achilles exemplifies that path. He cannot resist the desire to win fame and glory, so he returns to the war and revenge for Patroclus. Once he decides to fight, he is fated to die young. However, he still chooses to do so because of both his instinct and his rationale. He is born a noble person, and the blood flowing in his body is filled with the heroic bravery to fight for glory and show his masculinity. More importantly, his rationale is to revenge Patroclus, because his best friend died fighting in his place under the weapon of Hector. His sense of friendship urges him to discard any previous rage against Agamemnon and focus on the deep grief and resentment created by his human emotion. No individual person or factor can change the decision of Achilles, let alone fate, because his instincts make him think and act heroically.

During wartime, with violence and cruelty filling every corner of the battlefield, the exhibition of friendship as a sign of humanity is one of the most important contributions of Achilles’ agency. After Patroclus’ death, Achilles is so grieved that he refuses to have any food and even hopes “with all his heart that [he] alone would die far from the stallion land of Argos” so that Patroclus can be safely back home. This is one of those moments when we are pulled out of cruel battle scenes and step into Achilles’ shoes to comprehend his grief. The way Achilles chooses to react to the death of Patroclus demonstrates the profound emotion generated by friendship. In fact, the force of friendship is perhaps the most significant motivation for heroes like Achilles during combat, for his ability to manipulate and react to his emotions is the result of agency and humanity. Accordingly, one can confidently admire Achilles as the best of heroes considering how much effort he takes to suppress his grief and anger, pick up his weapon, and rush to the battlefield.

Hector, as another major character in the poem, embraces his fate by facing his destined death and the fall of Troy while seeking glory and brave actions. As I mentioned

earlier, Hector is not a flawless hero, and in fact, he has made a number of mistakes and is not always brave and heroic. On certain occasions, he is tempted to run away from fate. However, Hector still manipulates his choices and proves his worthiness as a hero in the end. The huge contrast of Hector's different decisions emphasizes the significance of his agency when he confronts Achilles for the very last time in his life.

On first sight of Achilles, Hector is "unshakable, furious to fight Achilles to death" (Homer 542). Again, he stands in great contrast to his father Priam, who is deeply worried that Hector is going to face his doomsday fighting Achilles, the "man who robbed [him] of many sons [and] brave boys" (Homer 543). Being reasonably scared when he does face Achilles, the greatest Achaean warrior, Hector tries to negotiate with Achilles and runs around the city of Troy three times, having Achilles chasing closely after him. These actions are honestly not heroic, but they are the proper reaction of anyone who has to face such a fearsome enemy as Achilles. In the face of his fated doom, Hector is tempted to run away because of his fear of death and his longing for life. At this moment, he is surely manipulated by fate. Yet the more important moment to consider is when Hector finally stops running and turns to confront Achilles. The mind of Hector is finally resolved and determined, choosing not to be governed by fear, but to be led by heroism:

Well let me die—
but not without struggle, not without glory, no,
in some great clash of arms that even men to come
will hear of down the years! (Homer 551)

He still holds his power of agency even at the last second of his life, and he uses it to choose a death with glory, rather than death as the result of fate. His final actions are the most touching and sorrow-provoking scene in *The Iliad*. I am deeply moved by his bravery under such pressure, which ensures Hector of the highest level of glory and respect. He displays his best power as a strong, brilliant hero, like "a soaring eagle launching down from the dark clouds to earth to snatch some helpless lamb or trembling hare" (Homer 551).

Hector's death is a truly heroic death and is different from other mortals' deaths. As a hero who dies within the dictates of fate but acts with admirable moral height and bravery, he stands and defines himself as an irreplaceable hero. His death is "a death that means, not just a simple factual event" (Nikolopoulou 178).

Being always right and heroic is an impossible thing for any human being, so instead, it is the way heroes choose to arrive at their final decision that truly matters. The heroes' free will contributes to making the decisions meaningful in terms of heroism. Their agency in the presence of fate allows them to define themselves as real heroes and show the world how much admiration they deserve for using their agency to the greatest extent even when it cannot save them from death.

IV.

In a number of readers' opinion, gods are interfering heroes' fate and depriving them of their autonomy. It is not surprising for readers to think that way, because as Gaskin points out in his paper, several previous scholars also believe that "Homeric heroes do not make proper decisions because their decisions are made for them...by gods" (Gaskin 6). The power of gods is not to be confused with fate, but these two elements have a complex relationship, and they both contribute to the agency of heroes. On one hand, if we look closer to the actions in the text, the gods are incapable of altering fate and taking away heroes' right to choose. Instead, the gods work within the dictates of fate and act with their

supernatural power to either help the hero or watch him reach his final destination. On the other hand, heroes can decide to exert some effect on the will of the gods by seeking their aid through prayers. The agency of heroes is only revealed in the smaller actions and nuances.

During the intense fight between Achilles and Hector, Zeus tries to wage in and “pluck [Hector] from death and save his life” or “strike him down at last at Achilles’ hands for all his fighting heart” (Homer 547). In Zeus’s mind, he may be able to save Hector from the furious Achilles and keep the great warrior alive longer than fate desires. Immediately, the goddess of wisdom, Athena, appeals to Zeus strongly and warns him of the nature of fate:

A man, a mere mortal,
his doom sealed long ago? You’d set him free
from all the pains of death?
Do as you please—
but none of the deathless gods will ever praise you. (Homer 547)

Athena’s warning gives Zeus the clear indication that gods do not have control over the settled fate. In addition, meddling with fate will not give Zeus credit for saving Hector’s life, but the interference could bring him trouble and criticism from other gods.

Afterwards, Zeus weighs the fate of Hector and Achilles on his golden scale, “and down [goes] Hector’s day of doom” (Homer 547). The power of fate is so remarkable that the power of the gods seems helplessly weak. At some point in this unpredictable world, a person’s fate is already decided upon, and no one, not even the gods, has the ability to alter fate. Like heroes, the power of the gods also works within the dictates of fate, which should not be a blockage on the course of life for neither heroes nor gods. It can be interpreted as a fixed scheme of outcome with flexible pathways to reach it, but the ways are dominated by agency that defines the heroes, as mentioned in the previous section.

In those possible ways, the heroes’ agency is still under the clouds of fate such that it may require careful interpretation to detect the nuances within. Inarguably, the gods intervene for the heroes by their wish almost all the time, and it is surprising to see how “this occurs more frequently than heroes’ prayers to the gods” (Jones 113). However, heroes may choose to ask for help from the gods, especially when life or honor is at stake.

Some of the gods’ intervention comes from Achilles’ search for help when his honor is undermined by Agamemnon’s insult. He weeps and prays to his sea goddess mother Thetis for help, and begs her to talk with Zeus:

Remind him of that,
now, go and sit beside him, grasp his knees
persuade him, somehow, to help the Trojan cause,
to pin the Achaeans back against their ships,
trap them round the bay and mow them down.
So all can reap the benefits of their king—
so even mighty Atrides can see how mad [Agamemnon] was
to disgrace Achilles, the best of the Achaeans! (Homer 91)

As mentioned earlier, Achilles is being petty and a little childish at this point: in order to solve the trouble in his heart and win back his glory, Achilles pleads with his mother to go to Zeus and take revenge against Agamemnon by letting the Achaeans suffer.

While this request can be viewed as Achilles' choice not to be heroic, it is also a choice to take advantage of the power of the gods. In the end, Achilles does get what he deserves—eternal fame and glory, although at a heavy cost. How can gods make decisions for Achilles, when Achilles can decide to ask gods for help to fulfill his desires?

We may sometimes be confused by obvious facts and fail to understand the hidden truth. For heroes, asking aid and guidance from gods is an option to consider, though whether gods grant their wish is beyond the heroes' control. The gods have a complex relationship with fate and heroes, but they behave within the dictates of fate and have no power over heroes' choice nonetheless. Heroes have even more flexibility to make decisions than gods sometimes, because while the gods are required to mind the consequences in both the divine world and the mortal world at the same time, heroes are driven by their instinct, humanity, and heroism to make good use of their agency.

Conclusion

The dominance of fate and the gods' will is definitely extensive in the great heroic epic, *The Iliad*. This is shown by the depictions of how fate determines characters' outcomes, or how the gods are even fighting amongst themselves to award either the Achaeans or the Trojans victory through trickery and cunning. I am not trying to pretend that these facts do not exist and heroes are free agents handling every situation. In a world where fate determines most outcomes and the gods interfere with the affairs of mortals, the heroes still possess the rights to make decisions for themselves. Homer's epic is remarkable in this aspect, that no matter how difficult, behaving in accordance to personal values is still possible. The outcome is not conflictive, because it is already fixed in a way; the way to reach the outcome is the problem I am more concerned with. Carefully discerning these nuances in the process provides new insights into this long-debated problem—whether the heroes are completely puppets of gods.

From the prophecies, we know that Troy will fall in the end. However, *The Iliad* depicts the last year of the ten-year Trojan War and gives us some understanding of the grandeur of heroes and their actions under the dictates of uncontrollable forces. Achilles and Hector, each the greatest warrior from the Greek army and Trojan army, have flexibility to make the optimal decision within the range of fate. Paris, with qualities contrasted to other heroes' bravery, also faces choices, although in a more ambiguous way. In the process of discovering the nuances of their decisions, the beauty of humanity and heroism can be seen besides their agency. These qualities interact with each other, work together to shape the heroes, and provide readers insights into their efforts to define themselves.

Under the influence of fate, especially fated deaths, the heroes can falter at first and run away or surrender to it, but later, they embrace it as an essential part of their lives and run towards it to show their courage, bravery, agency and above all, the desire for glory. The societal norms of the ancient Greeks bound heroes in a situation where war is the sole way of proving oneself. By engaging in duels, heroes gain military honors and respect from others, as victory in battles implies their braveness and exceptional abilities to defeat their enemies for a noble purpose, fitting the definition for heroism.

Discerning critically through the heroes' seemingly unimportant decisions is not a way to debate the terminologies. Instead, it provides readers a new way to examine the heroes' qualities and abilities. They are more than brave warriors, open to multiple interpretations under a vast canvas of fate. They can paint on it and make it more colorful than the life of a fully free human. Their stories echo across thousands of years of history to us to prove their worthiness. They show that however strong the forces of fate and the gods' will may be, they can still do better than being manipulated by the force. They distinguish

their heroic nature from among a vast array of literary characters. I believe that the heroes in *The Iliad* are powerful agents with some degree of agency, and their ability to make self-determined decisions enhances their heroism, the presence of both obliged purposes and respectful consequences, and their humanity, the capacity to represent and inspire humankind.

Works Cited

- Homer. *The Iliad*. Trans. Robert Fagles. New York: Penguin Books, 2001. Print.
- Gaskin, Richard. "Do Homeric Heroes Make Real Decisions?" *The Classical Quarterly* 40.1 (1990): 1-15. *JSTOR*. Web. 16 Aug. 2017.
- Jayawickreme, Eranda, and Di Stefano, Paul. "How Can We Study Heroism? Integrating Persons, Situations and Communities." *Political Psychology* 33.1 (2012): 165-178. *Wiley Online Library*. Web. 22 Nov. 2017.
- Jones, P. V. "The Independent Heroes of *The Iliad*." *The Journal of Hellenic Studies* 116 (1996): 108-18. *JSTOR*. Web. 15 Sept. 2017.
- Morrison, J. V. "Kerostasia, The Dictates of Fate and the Will of Zeus in *The Iliad*." *Arethusa* 30.2 (1997): 276-296. *Project MUSE*. Web. 3 Sept. 2017.
- Nikolopoulou, Kalliopi. "Feet, Fate, and Finitude: On Standing and Inertia in *The Iliad*." *College Literature* 34.2 (2007): 174-193. *JSTOR*. Web. 12 Aug. 2017.
- Solomon, Robert. "On Fate and Fatalism." *Philosophy East and West* 53.4 (2003): 435-54. *Project MUSE*. Web. 21 Aug. 2017.
- Sophocles. *The Oedipus Trilogy*. Trans. F. Storr. Cambridge: Harvard University Press, 1912. E-book.
- Stevanovic, Lada. "Human or superhuman: The concept of hero in ancient Greek religion and/in politics." *Glasnik Etnografskog Instituta* 56.2 (2008): 7-23. *DOAJ*. Web. 3 Sept. 2017.



