AN APPLICATION OF MANY-FACET RASCH MODEL TO THE ANALYSIS OF AN IMPLICIT ASSOCIATE TEST: DISENTANGLING IMPLICIT GENDER-SCIENCE STEREOTYPE

By

SIYU WAN

(Under the Direction of Louis A. Castenell)

ABSTRACT

The article aims to measure implicit gender-science stereotype of female and male individuals. A Many-Facet Rasch Measurement analysis was used to disentangle the contribution of specific associations to the overall IAT measure. A preference for associating males with science and females with liberal arts is observed in both gender groups. Male participants show stronger stereotype than female participants, and this preference is driven primarily by associating males with science rather than females with liberal arts. Besides, some stimulus words played different roles to the overall IAT effects. This research supported that MFRM is a useful method for exploring IAT. As consequences, we argue that researchers should be more careful when choosing stimulus for IAT and interpreting IAT effects.

INDEX WORDS: Implicit Gender-Science Stereotype, Implicit Association Test, Many-Facet Rasch Model

AN APPLICATION OF MANY-FACET RASCH MODEL TO THE ANALYSIS OF AN IMPLICIT ASSOCIATE TEST: DISENTANGLING IMPLICIT GENDER-SCIENCE STEREOTYPE

By

SIYU WAN

B.S., Central China Normal University, China, 2015

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF ARTS

ATHENS, GEORGIA

2017

AN APPLICATION OF MANY-FACET RASCH MODEL TO THE ANALYSIS OF AN IMPLICIT ASSOCIATE TEST: DISENTANGLING IMPLICIT GENDER-SCIENCE STEREOTYPE

By

SIYU WAN

Major Professor: Louis A. Castenell

Committee: George Engelhard
Martha Carr

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2017

DEDICATION

I dedicate my thesis work to my family and many friends. A special feeling of gratitude to my loving parents, whose words of encouragement and push for tenacity ring in my ears.

I also dedicate this dissertation to my many friends who have supported me throughout the process. I will always appreciate all they have done.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the support and assistance of the members of her committee and colleagues during the preparation of this manuscript. Drs. Louis Castenell, George Engelhard, and Martha Carr served as invaluable sources of knowledge, guidance, and encouragement throughout this entire process. With the deepest respect and sincerity - thank you.

The author further acknowledges the assistance of her fellow graduate students in the development and analysis of the data used for this study. Jue Wang, M.A., and Yawei Shen, M.A. spent countless hours of their time involved in the development of data analysis, their administration, and numerous codes written procedures, to which the author is indebted.

Gratitude is also extended to the participants included in the study. Although they may not realize it, their involvement in this study and the findings revealed from analyzing their lives will have an impact on future generations.

Finally, the author wishes to thank her family for their support and encouragement in attaining his educational, professional, and personal aspirations.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Gender differences in mathematical performance and science, technology, engineering, mathematics (STEM) are well known, men tend to have better performances than women on many advanced mathematical ability tests, including Scholastic Aptitude Test (SAT) and the Graduate Record Exam (GRE) (Brown & Josephs, 1999; Hyde, Fennema, & Lamon, 1990, Walton & Spencer, 2009, Major & O'Brien, 2005, Nosek, & Smyth, 2011). Thus, there is a stereotype which proposes that women have less mathematical aptitude than men (Benbow, Lubinski, Shea, & Eftekhari-Sanjani, 2000; Nosek, & Smyth, 2011). Researchers found, through stereotype threat, women's performances and interests in related domains have been negatively influenced by this stereotype (LeFevre, Kulak, & Heymans, 1992). And this difference in performances may lead to different career trajectories: compared with women, men are more likely to major in mathematics, and pursue math-intensive careers, such as computer science and engineering (Davies, Spencer, Quinn, & Gerhardstein,2002; Jacobs & Eccles, 1992; Quinn & Spencer, 2001; Sekaquaptewa & Thompson, 2003; Shih, Pittinsky, & Ambady, 1999; Spencer, Steele, & Quinn, 1999).

Stereotype threat describes the situation in which there is a negative stereotype about a persons' group, and the person will concern about being confirmed, judged or treated negatively based on this stereotype (Spencer et al. 1999, Steele 1997, Steele & Aronson 1995). According to the stereotype threat theory, women's math performance is

lowered under threat situation not due to women's insufficient ability, but because women feel threatened by the possibility that their performance will confirm the negative stereotype associated with their social group. Under threat condition, women reported more negative domain-related thoughts (Cadinu, Maass, Rosabianca, & Kiesner, 2005) less entrepreneurial intentions (Gupta & Bhawe, 2007) and interests in attending a mathematics, science, and engineering conference (Murphy et al.,2007). Fogliati and Bussey (2013) created the stereotype threat condition by telling their participants that men outperform women on the test, then participants were required to complete a mathematic test. Compared with women who are under no-stereotype condition, women who received negative stereotype information showed an increase in self-esteem, but decreases in performances and motivation.

Although in advanced mathematics studies, science-related activities and careers, the participation of girls and women has increased over the years; and expression of gender- science stereotype publicly has been considered improperly (Hyde, Lindberg, Linn, Ellis, &Williams, 2008; Halpern et al., 2007). By using Implicit Association Test (IAT), some researchers found that implicit gender- science stereotype still existed (Greenwald & Farnham, 2000; Kiefer & Sekaquaptewa, 2007; Nosek, Banaji, & Greenwald, 2002). As a well-developed and widely used paradigm, IAT has been used on a large number of topics in different areas. During the last 20 years, researches using IAT to test attitudes towards target groups (e.g., African-Americans, homosexuals) have increased abundantly (Egloff & Schmukle, 2002; Greenwald & Farnham, 2000; Greenwald & Nosek, 2001; Lane, Banaji, Nosek, & Greenwald, 2007; Nosek et al., 2007; Rudman, Greenwald, Mellott, & Schwartz, 1999). Compared with explicit measures (e.g.,

questionnaires), which may be heavily affected by social pressure and strategies of impression management, IAT does not rely on self-reported procedures. It provides behavioral measures of association strengths among mental representation.

But there are still some criticisms against IAT on a measurement level, some psychologists point out that it does not allow the latent dimension on the individual to be measured. While IAT effects can be influenced by stimulus features (Brendl, Markman, & Messner, 2001; Mitchell, 2004), and category labels (DeHouwer, 2001; Fazio & Olson, 2003). Aiming at solving the limitation of the traditional algorithm of IAT, and decomposing the IAT effect, psychologists have come up with some new and advanced measurement methods. This research will adopt the Many-Facet Rasch Model (MFRM) of data analysis to explore individual's implicit gender- science stereotype, aiming at disentangling the contribution of specific associations, to overall IAT measurement. The following literature review was organized into four parts: Gender-Science Stereotype, Implicit Association Test, Many-Facet Rasch Measurement, and The Application of MFRM on IAT.

Gender-Science Stereotype

The terminology, stereotype, was firstly referenced in the psychological area in 1992. American journalist Walter Lippman used the word, "stereotype", in his work Public Opinion. From a social psychology perspective, the stereotype is a belief that can be adopted about specific types of individuals or certain ways of doing things, based on their sex, race, living area, and occupation (McGarty, Vincent, & Spears,2002). These thoughts or beliefs sometimes not reflect reality accurately. As a specific cognitive schema, the stereotype can help to simplify and systematize the procedure of information

processing. When stereotype was used, information is identified, recalled, predicted, and reacted to more easily. Stereotypes can also be considered as categories of objects or people. Between stereotypes, there are lots of difference between objects or people. But within stereotypes, objects or people are more likely sharing similarity with each other. The contents of these stereotypes reflect what qualities that individuals have been attributed. These stereotypes will guide our behavior, in turn, we will act toward the person like these assigned qualities were true. Although using stereotype can increase the speed of some cognitive processes, it would be problematic, if the stereotype contains too many affective and favorable components.

Through stereotype threat, the stereotype has been proved to cause some negative influences. Stereotype threat is a situational predicament in which people are or feel themselves to be at risk of confirming negative stereotypes about their social groups. Under these threat circumstances, fear of confirming the stereotype poses a threat to targets of the stereotype, thereby undermining their performance (Steele & Aronson, 1995). As soon as it was brought into the academic field, stereotype threat has become one of the most popular topics in the field of social and educational psychology. Aronson (1995) found that because there was the stereotype that compared with other groups, African Americans were less intelligent, this stereotype threat could lower the performance of African Americans on SAT test, which is used for college entrance in the United States. The range of impaired groups is wide. Stereotype threat can negatively affect financial decision making (Carr & Steele, 2010), golf putting (Stone et al. 1999), safe driving (Yeung & von Hippel, 2008), and memory performance among older adults (Mazerolle et al. 2012).

Gender stereotype is one of them, there is a stereotype involved gender and math ability that proposes women have less mathematical aptitude than men. Some researchers (e.g., Benbow, Lubinski, Shea, & Eftekhari-Sanjani, 2000) found that junior high school boys had better performances compared with girls on advanced quantitative assessments. This performance difference also existed in some tests of students' advanced mathematical ability, including Scholastic Aptitude Test (SAT) and the Graduate Record Exam (GRE; Brown & Josephs, 1999; Hyde, Fennema, & Lamon, 1990). More and more research evidence indicates that these gender-science stereotypes will influence women's interest and performance in the math domain by stereotype threat (Davies, Spencer, Quinn, & Gerhardstein,2002; Jacobs & Eccles, 1992; Quinn & Spencer, 2001; Sekaquaptewa & Thompson, 2003; Shih, Pittinsky, & Ambady, 1999; Spencer, Steele, & Quinn, 1999; Stoet & Geary, 2012; Tomasetto, Alparone, & Cadinu, 2011; Shapiro & Williams, 2012).

More specifically, under standard test-taking situations, in which math tests are perceived to be diagnostic of math ability, women typically experience stereotype threat and perform worse than men (Smith & White, 2002). However, when creating a stereotype threat-free environment by telling women that the math test is gender-fair (e.g., Schmader, 2002), or by instructing women that the test is not a diagnostic of their mathematical ability (e.g., Gonzales, Blanton, & Williams, 2002; Quinn & Spencer, 2001), there is no significant performance difference between women and men. Besides lowering female students' mathematical performances, stereotype threat can lead to many other negative influences. Fogliati and Bussey (2013) found the stereotype threat reduce their motivation to improve that led to the worse performances of women's math in the

future. And this negative influence is very strong and long-lasting, even for those women who have already in science, technology, engineering, and math (STEM) fields (Good, Aronson, & Harder, 2008).

Evidence showed that unconscious processing of stereotype information of tests may be sufficient to cause the negative influence in targets' performance under threat circumstances. Reports of explicit concerns about being stereotyped or stereotype-consistent performance, such as evaluation apprehension, anxiety, and performance expectations, do not reliably mediate stereotype threat effects (Bosson, Haymovitz, & Pinel, 2004; Wheeler & Petty, 2001; Johns, Schmader, & Martens, 2005). It is hard for target people to detect when their performance is under the effect of stereotypes accurately. Therefore, stereotype threat effects sometimes occur without conscious awareness. If the unconscious process of stereotype-relevant information can arouse stereotype threat, then implicit gender-science stereotypes, or non-conscious associations of men more than women with mathematics, may influence women's susceptibility to stereotype threat.

This opinion has been explored and proved by using a special experimental paradigm, Implicit Association Test (IAT). Researchers (Greenwald & Farnham, 2000; Lemm & Banaji, 1998) found men tend to associate their personal information, such as their names, with "typical" male traits, while women tend to implicitly associate their information with stereotypically female traits. These associations can also be observed in the academic domain. Both male and female participants implicitly prefer to associate men more with math and science, and women more with arts and humanities (Kiefer & Sekaquaptewa, 2006; Nosek, Banaji, & Greenwald, 2002). In other words, they both hold

implicit gender-science stereotypes. Moreover, these implicit stereotypes related to less explicit math identification, less positive attitudes to mathematics, and lower reported performance on math-related achievement tests for women (Kiefer & Sekaquaptewa,2007; Nosek et al., 2002).

Implicit Association Test

During the last decades, research on implicit methods has increased significantly. The birth of series of implicit techniques was not only due to the need of implicit social cognition research, but also the result of the application of response-time paradigm. The characteristic of implicit cognition is that past experience will affect individual performance, even these earlier experience is not stored consciously, that is, it is unavailable to self-report or introspection (Greenwald, 1990). As a popular research topic in social cognition field, implicit attitude is an automation process occurred in the unconscious situation (Greenwald, & Banaji, 1995). Because of their unconscious, automatic, it is hard to measure by traditional and self-reported methods, psychologists turned to implicit measurement to avoid the social desirability or impression management strategies.

There are different implicit (or indirect) techniques to test implicit attitudes, such as evaluative priming (EP; Fazio, Sanbonatsu, Powell, & Kardes, 1986), the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), which has played a predominant role since it came out. Many other tests were came up with rapidly following IAT, included the go/no-go association task (GNAT; Nosek & Banaji, 2001), the extrinsic affective Simon task (EAST; DeHouwer, 2003), the affect misattribution procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005), the single category– IAT

(SC–IAT; Karpinski & Steinmen, 2006), and the sorting paired features (SPF; Bar-Anan, Nosek, & Vianello, 2009). Most of these implicit methods based on the response-time paradigm.

The basic procedure is providing participants some stimulates and asking them to response as quickly and precisely as possible, and recording the time between the appearance of stimuli and response of participants, the latency is the response time. The length of latency reflects the complexity of inside processing procedure. When it comes to social cognition filed, stimulus always own special social meaning, which may arouse the corresponding response of participants. The stimuli can be consistent with their implicit attitude, or conflict with it, thus the complexity of processing procedure of participants is different. Under the quick reaction requirement, participants can hardly use conscious strategies to control their reaction, thus the result of participant's social cognition in this condition was considered as implicit. In order to measure individual differences in implicit social cognition on inaccessible levels, researchers need to use sensitive indirect measures (Greenwald, & Banaji, 1995). Greenward et.alt. (1998) based on a response-time paradigm, improved and developed traditional latency method, came up with the Implicit Association Tests (IAT).

From a physiological perspective, the neural network model is the basis of IAT. This model proposed that information is stored at a series of nodes of neural connection, which are organized by layered semantic relationship. Hence, by measuring the distance of two concepts on neural connection, we can test the relationship between two concepts (Farnham, Greenward, and Banaji,1999). By using a computerized classification task, IAT can measure the degree of automatic connection between two concepts (concept

words and attribute words), then it can achieve measurement of individual implicit attitude.

IAT also use reaction time as an index, the basic procedure of IAT is presenting a series of attribute words, participants are required to classify them (into concept words) and press related keys as quickly as possible, their reaction time was automatically recorded. Between the concept word (such as white, black) and attribute word (such as good, bad), there are two possible situations: compatible situation (such as white - smart, black - stupid) and incompatible situation (such as white - stupid, black - wise) or vice versa. The so-called compatible, which means the link between the two words is consistent with the participant's implicit attitude. In other words, the link between the two words is close and reasonable for the participant, otherwise, it is incompatible. When concept words and attribute words are compatible, under the requirement of quickly reaction, classifying procedure will use more automatic process, which is relatively easier and therefore the reaction time will be shorter; when they are incompatible, the classification task will need complex conscious processing procedure, which is relatively difficult, and therefore the reaction time will be longer. The difference of reaction time between the compatible condition and incompatible conditions is the indicator of implicit attitude.

Table 1.

Process of classical Implicit Association Test (IAT)

Stage	Left key assignment	Right Key assignment
1.Initial target-concept discrimination	Black People Faces	White People Faces
2. Associated attribute discrimination	Good	Bad
3. Initial combined task	Black People Faces + Good	White People Faces + Bad
4. Reversed initial target-concept discrimination	Bad	Good
5. Reversed combined task	Black People Faces + Bad	White People Faces + Good

IAT is generally conducted on the computer, Table 1 gives a classic IAT which assessed implicit attitudes toward Black and White people, as an example to explain its process. IAT is usually divided into five stages (or into seven stages when there are two trails for each combined task), each stage contains a discrimination task. In Stage 1, participants are asked to rapidly classify pictures into the categories Black People (by pressing the left computer key) and White People (by pressing the right computer key). Then the same task for stage 2 to classify categories good (by pressing the left computer key) and bad (by pressing the right computer key) in Stage 2. In Stage 3, the previous two tasks are combined. Participants are instructed to press the left key when any item belonged to category Black People or good appears on the screen, and press the right key when any item belonged to White People or bad appears on the screen. In the next stage,

the task in Stage 2 is reversed, attribute word bad is paired with the left key, while word good is paired with the right key. Similarly, Stages 5 reverses the earlier combined pairings task of Stages 3: Black People + bad now share the left response key, and White People + good share the right response key. The computer automatically records response latencies (in milliseconds) and error rates. After a series of statistical process, the data of combined tasks (stage 3 and 5) are used to obtain the IAT scores (IAT effects), scores are the mean response latencies difference between two combines tasks.

In 2003, based on analyses of larger sets from public website Project Implicit, Greenwald and colleagues (2003) made some modification to scoring algorithm, they used D score to replace the previously scoring method. D is computed as the difference in average response latency between two combined tasks, divided by its associated "inclusive" standard deviation. Compared with the initial algorithm, D score is an improved scoring algorithm in many aspects, such as internal consistency. There are still some rooms for improvement, the Many-Facet Rasch Measurement is one of methods to deal with problems of IAT.

Since the initials publication of IAT in 1998, lots of researches have conducted to provided many evidences concerning the psychometric properties of IAT measures (Egloff & Schmukle, 2002; Greenwald & Farnham, 2000; Greenwald & Nosek, 2001; Lane, Banaji, Nosek, & Greenwald, 2007; Nosek et al., 2007; Rudman, Greenwald, Mellott, & Schwartz, 1999). IAT measures had good internal consistency, the α coefficient of IAT was from 0.77-0.95 (Bosson, Swann, & Pennebaker, 2000; Dasgupta & Greenwald, 2001; Greenwald & Farnham, 2000; Greenwald & Nosek, 2001, Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005). Besides, IAT measures are not affected

by variations in subjects' familiarity with IAT stimuli (Dasgupta, McGhee, Greenwald, & Banaji, 2000; Ottaway, Hayden, & Oakes, 2001; Rudman et al., 1999); and IAT measures are relatively insensitive to procedural variations including the number of trials, the number of stimuli, and the interval between trials (Greenwald et al., 1998; Nosek, Greenwald, & Banaji, 2005). While the test-retest reliabilities of IATs had a median of .56 across some different tests (Nosek et al., 2006).

The validity of IAT was less satisfactory when compared with reliability. Correlations between IATs and other implicit measures are relatively weak (Bosson et al., 2000). Banse (1999) proposed that is due to unsatisfactory reliabilities of other implicit measures like priming procedures. Besides, two meta-analyses that cover many domains (including attitude, self-concept, and stereotype IATs) found that average correlations between IATs and explicit measurement were 0.24 (Hofmann, Gawronski et al., 2005) and 0.37 (Nosek, 2005). When it comes to predictive validity for behavioral measures of IATs, results were across domains. There are correlations between IAT and other behavioral measures, such as math SAT scores (Nosek, Banaji, & Greenwald, 2002b), anxious behaviors (Asendorpf et al., 2002), and alcohol consumption of a month (Wiers, van Woerden, Smulders, & de Jong, 2002). But there are still some areas, IAT measures did not make an expected prediction (e.g., food choice in Karpinski & Hilton, 2001).

Many-Facet Rasch Measurement

Many-facet Rasch measurement model belongs to a big family of models with roots in the dichotomous Rasch Model, which developed by Danish mathematician and statistician Georg Rasch (1960/1980). The Rasch model based on objective measurements in the natural sciences sets a set of objective criteria for measurement in

the social sciences to ensure that the information provided by the measurements is more objective and reliable (Bond & Fox, 2007). After half a century of development, Rasch model has been widely used in the field of psychology.

By observing the individual performance of subjects (usually expressed as the raw score), the Rasch Model can measure latent variables, which are not directly observable. According to the principle of Rasch model, the probability of getting a specified response (e.g. right/wrong answer) can be estimated as a function of person and item parameters, which contains the individual's ability and the difficulty of one item. Whether an individual can answer a topic correctly or not entirely depends on the comparison between individual ability and difficulty of the item. Unlike general Item Response Theory (IRT) that adopts a "the model fits data" opinions and uses different parameters to adjust the data set, the Rasch model is an idealized mathematical model, which requires that "data fits the model" to achieve objective measurement (Andrich, 2004). Wright and Stone (1979) pointed out that, there were two basic requirements of Rasch model: (1) For any item, individuals owned higher ability are more likely to make the correct answer than individuals owned lower ability; (2) For any individual, they always have better performances on easy items than on difficult items.

The most fundamental model was developed by Rasch, this model has been variously referred to as the Rash model (Wright & Stone, 1979), or the simple logistic model (SLM, Andrich, 1988), or the one-parameter logistic (1PL) model (Yen & Fitzpatrick, 2006). This model only has two facets: individual ability and item difficulty. Equation 1 defines the dichotomous Rasch model, given a response X_{ni} to a test, which is equal to one if the answer is correct, zero if the answer is wrong; β_n represents the ability

of the individual n, and δ_i represents the difficulty of item i, the following mathematical form presents the dichotomous Rasch model in its exponential form.

$$P(X_{ni} = x_{ni}|\beta_n, \delta_i) = \frac{\exp[x_{ni}(\beta_n - \delta_i)]}{1+\exp(\beta_n - \delta_i)}. \quad (1)$$

By using Equation 1, we can calculate the probability of getting a correct response or an incorrect response, respectively.

$$P(X_{ni} = 1|\beta_n, \delta_i) = \frac{exp(\beta_n - \delta_i)}{1+exp(\beta_n - \delta_i)} \quad (2)$$

and

$$P(X_{ni} = 0|\beta_n, \delta_i) = \frac{1}{1+exp(\beta_n - \delta_i)}. \quad (3)$$

Then by using logarithm, we can obtain

$$ln\frac{P(X_{ni}=1|\beta_n,\delta_i)}{P(X_{ni}=0|\beta_n,\delta_i)} = ln\frac{exp(\beta_n - \delta_i)/[1+exp(\beta_n - \delta_i)]}{1/[1+exp(\beta_n - \delta_i)]} = \beta_n - \delta_i. \quad (4)$$

Based on this model, the probability that individual n answers item i correctly, that is P (*Xni=1*), depends on the difference between the ability of the individual (β_n) and the difficulty of item i (δ_i). The more (or less) able the individual is, and the easier (or more difficult) the item is, the more (or less) probable that a correct response will be obtained. If an individual ability equals to an item's difficulty, $\beta_n - \delta_i = 0$, with *exp*(0) = 1, the individual will have a .05 chance of getting a correct response.

The most important advantage of Rasch model is measurement invariance when compared with item response theory (IRT). When a given set of data fits the Rasch model, individual measures are invariant across different sets of items, and item measures are invariant across different individuals. Measurement invariance has a crucial implication: Response scores are sufficient statistic for the estimation of individual ability, which means the number-correct score contains all the information required for the estimation

of an individual's measurement from a given set of observations. In other words, how many items the individual responded correctly is important instead of which item the individual answered correctly. The same for the estimation of item difficulty, the estimation of an item depends on how many people answered that item correctly.

Many-facet Rasch measurement (MFRM) is a model that can suit analysis requirements of multiple variables having impacts on assessment outcomes. MFRM model incorporates more variables or facets compared with the classic testing situation which only includes two of them. Since its first comprehensive theoretical statement by Linacre in 1989, MFRM has been adopted rapidly in highly varied disciplines and fields of research. In education assessment field, MFRM can be used to evaluate alignment of items to contented standards (Anderson, Ervin, Alonzo & Tindal, 2015), or to evaluate APELC performance assessment (Engelhard & Myford, 2003); In language assessment area, MFRM can be used to measure speaking ability through group oral testing (Bonk & Ockey, 2003), or analyze rater effects in Japanese writing assessment language testing (Kondo-Brown, 2002); In medical education field, MFRM can be adopted to measure physician-patient communication skills (Harasym, Woloschuk, & Cunning, 2008), or evaluate student selection using the MMI (Till, Myford & Dowell, 2013). Besides above-mentioned application, MFRM can also be used to study implicit attitudes towards specific groups or objects, which is the key topic of this research.

The Application of Many-facet Rasch Model on Implicit Associate Test

A three-facets Rasch model by introducing a new facet accounting for the associative condition the item is presented in to analyze the IAT data. The new model takes on the following form:

$$ln\frac{P_{nijk}}{P_{nij(k-1)}} = \beta_n - \delta_i - \gamma_j \quad (5)$$

Where P_{nijk} (resp. $P_{nij(k-1)}$) is the probability that respondent n would respond to stimulus i in condition j; β_n represents the ability (speed) of respondent n; δ_i represents the difficulty (speed of categorization) of stimulus i; γ_j represents the ease of condition j. Respondents, stimuli, and conditions are three facets. Based on response to an IAT, the model shows that the probability that a respondent n gives a response stimulus i on condition j depends on the additive effects of the speed of the respondent (β_n), the speed of categorization of the stimulus (δ_i), the ease of the condition (γ_j).

When compared with application of MFRM in other fields, such as measuring rater effects, there are far fewer researches focused on IAT, almost all of them were conducted by psychologists in University of Padua, Italy. All their research findings (Vianello & Robusto, 2010; Anselmi, Vianello, & Robusto, 2011, 2013; Anselmi, Vianello, Voci, & Robusto, 2013) can be summarized into following aspects: (1) they applied MFRM in different forms of implicit tests, including Go/Not to Go association task (GNAT, Nosek & Banaji, 2001) and IAT, and proved the applicability and superiority of the MFRM compared to traditional measure methods; (2) they successfully verified participants' implicit attitudes towards specific targets, such as Black people, Homosexuals, and overweight people; (3) They explored the relationship between participants' groups and characteristics of attribute stimulus. The preferences of different group participants were affected by the different stimulus (positive or negative).

Anselmi, Vianello, and Robusto (2011) conducted a study using MFRM to explore the positive association's primacy in IAT, which is the first research adopting

MFRM to evaluate independently the different strengths of the positive and negative associations that are present in a standard IAT. The "positive associations primacy" effect represents the superiority of the associations of concepts with attribute "Good", compared to the associations with "Bad". In study 1, they found t that white participants implicitly preferred whites to blacks, and this preference would be better interpreted as implicit in-group favoritism rather than implicit out-group prejudice. In study 2, they used Weight IAT to test people (who were divided into 2 groups based on their own Body Mass Index) implicit attitudes towards obese people. The result indicated this "positive associations primacy" did not depend on respondents' membership of one of the IAT target categories, nor did it depend on implicit preference for the in-group. In line with Study 1, their preference for thin people was mainly due to the set of positive words, while negative words somehow lowered the IAT effect.

In the next few years, Anselmi, Vianello, and Robusto (2013) continued focusing on the application of MFRM in IAT. In order to explore the interaction between group membership and contribution of positive and negative association to overall measure, they (Anselmi, Vianello, Voci, & Robusto, 2013) tested implicit sexual attitudes of heterosexual, homosexual and bisexual people. Unlike the former research, there was a preference difference due to participants' own sexual orientation: both heterosexual and homosexual participants preferred people who had the same sexual orientation to them; whereas, bisexual participants implicitly preferred heterosexuals to homosexuals. For heterosexual participants, the preference for their own sexual orientation was mostly driven by the attributed positive words to heterosexuals, instead of negative words to homosexuals. However, in the homosexual group, the positive and negative attribute had

a similar effect on their weaker preference for homosexuals. Last but not least, bisexual participants' preference for heterosexuals came from the association of negative words to homosexuals, rather than positive traits to heterosexuals.

By using a series of IAT, such as implicit academic identity (associations between math/arts and self/other), implicit attitudes (associations between math/arts and good/bad), implicit stereotyping (associations between math/arts and male/female), and implicit math anxiety (associations between math/arts and anxious/confident), researchers (Nosek, Smyth, et al., 2007; Nosek and Smyth, 2011) found women showed weaker implicit positivity toward math than did men, weaker implicit identification with science, and stronger implicit math anxiety. Also, both men and women hold strong implicit gender stereotypes associating science with male. But there is no study adopting MFRM to test implicit gender- science stereotype.

Current Study

The current study seeks to further our understanding of implicit gender-science stereotype. Specifically, this study wants to validate that both female and male tended to associate male with science subjects, and female with liberal arts subjects. Besides, this research tends to find whether this preference has differences between women and men. Finally, this research aims to test 14 stimuli words which Gender-Science IAT used to represent concepts, science and liberal arts. Four research questions were investigated:

1. Whether both female and male participants hold male-science/female-liberate arts stereotype?
2. Can this gender-science stereotype be considered more as a male-science bias, rather

than female-liberate arts bias?

3. Do male participants endorse stronger male-science/female-liberate arts stereotype when compared to female participants.
4. Do all 14 stimulus words can represent concept science, and liberate arts equally across two different conditions.

CHAPTER 2

METHODS

Ethics Statements

The whole process of the study was conducted online. Visitors to the Project Implicit Website (https://implicit.harvard.edu/implicit/) self-selected to participate in the ‘‘Gender-Science IAT’’ task. Participants voluntarily searched for and accessed the Project. Before the study begins, they were informed that the study might detect associations that they were not consciously aware of and with which they might even explicitly disagree. And participants needed to give their consent by clicking on the following sentence: ‘‘I am aware of the possibility of encountering interpretations of my IAT test performance with which I may not agree. Knowing this, I wish to proceed’’. Participants were able to drop out of the study at any time without any consequences. The study required 10–15 minutes to complete and participants would receive feedbacks about their IAT performance at the end.

Participants

Based on the data reduction criteria of IAT, respondents whose average latencies for critical combined task block were over 1,800 ms were removed from the analysis. Respondents who made in excess of 25% errors in any critical block were removed from the analysis.

100 respondents (50% female, 50% male) who finished all 200 trails of Gender-Science IAT in 2015 have been randomly chosen. Their mean age was 26.23 (SD =9.127;

range from 16 to 53). The sample is heterogeneous, they were various of age, educational level, and career areas, thus, not representative of any specific population.

Materials and Procedure

The gender- science stereotype IAT used category labels Science, Liberal Arts, Male, and Female. There were 8 words to represent concept Male (Man, Son, Father, Boy, Uncle, Grandpa, Husband, and Male), and Female (Mother, Wife, Aunt, Woman, Girl, Female, Grandma, and Daughter) respectively.14 different subjects were used to represent concept Science (Astronomy, Math, Chemistry, Physics, Biology, Geology, Engineering) and Liberal Arts (History, Arts, Humanities, English, Philosophy, Music, Literature).

Table 2.

Process of Gender-Science Stereotype Implicit Association Test

Block	Number of Trails	Left Response (E)	Right Response (I)
1	20	Female	Male
2	20	Liberal Arts	Science
3	20	Female +Liberal Arts	Male + Science
4	40	Female +Liberal Arts	Male + Science
5	40	Male	Female
6	20	Male +Liberal Arts	Female + Science
7	40	Male +Liberal Arts	Female + Science

Note: Block 3 and 4 were counterbalanced across participants with block 6 and 7.

The classical seven-block IAT was used for the procedure (Greenwald, Nosek, & Banaji, 2003). Stimuli were presented in the center of the computer screen, and

respondents had to categorize them by pressing, as quickly and accurately as possible, the response key E or I. A red cross appeared when respondents made a mistake. The procedure had seven blocks. Three practice blocks involved the categorization of stimuli that represented either the gender stimuli or the subject stimuli. Four critical blocks required the simultaneous categorization of stimuli representing the four categories with two response options. In one option, Male and Science shared a response key, and Female and Liberal Arts shared the other. In the other condition, Male and Liberal Arts shared a response key, and Female and Science shared the other. The order of the two conditions was counterbalanced across all respondents.

Data Analysis

The IAT data (available at http://www.opensciencframework.org/) were analyzed through MFRM, a model belonging to the family of Rasch models. The model allows the investigation of the contribution of individual stimuli to the overall IAT measure. The MFRM also considers any source of systematic variability (facet) which might be useful for explaining the result. In the present study, facets are (a) persons, (b) gender of respondents, (c) associative condition, and (d) attribute stimuli.

Responses smaller than 300 ms and greater than 10,000 ms were deleted from the analysis, and response times were discretized according to percentiles computed on the 100 (number of respondents) ×2 (number of associative conditions) ×14 (number of attribute stimuli) complete data matrix. For our dependent variable, response time, the values 5, 4, 3, 2 and 1 identify very fast, fast, medium, slow and very slow responses, respectively.

The MFRM analysis was performed using the computer program FACETS 3.65.0 (Linacre, 2009a). A parameter α was estimated for each respondent indicating his/her speed in completing the IAT, a parameter β for female and male respondents indicating their speed, a parameter γ for each attribute stimulus indicating its speed of categorization, and a parameter ε for each associative condition indicating the ease of the condition. All the estimates are interval measures. Higher values indicate higher response speed, and they should be interpreted as higher respondents' speed in completing the IAT, higher speed of categorization of the stimuli, and greater ease of the associative conditions. Estimates of gender, associative conditions, and attribute stimuli were constrained to have a mean element estimate of zero.

The MFRM analysis has following indices:

1. The Infit and Outfit statistics evaluate the fit of the data to the model. If in the range from 0.5 to 2, it represents a good fit (Linacre, 2009a).

2. The Separation Ratio (G) represents a measure of the spread of the estimates relative to their precision. It ranges from 1 to infinity. G = 2, for instance, means that the dispersion in the measures of the elements in the facet is two times greater than the imprecision in their estimates (Wright, 1996).

3. The Separation Reliability (R) shows how reproducibly different the measures are. It ranges between 0 and 1. If R is close to 1, there is a high probability that the elements of the facet with high measure estimates actually have higher measures than those with low measure estimates (Linacre, 2009a). If $R < 0.5$, it is likely that the value of G is completely due to measurement error.

4. The Fixed (all-same) chi-square tests the hypothesis that the elements of a facet have the same logit in relation to the measurement error (SE).

The MFRM also allows the analysis of the interactions between elements of different facets. The interaction between the facets associative condition and participants' gender allowed us to investigate whether the ease of the associative conditions changes under different gender groups. Moreover, the interaction between the facets attribute stimuli, associative condition, and gender allowed us to investigate whether the speed of categorization of the stimuli changes according to the gender of respondents and the associative condition they are presented in. This interaction analysis also provides us with the contribution of each individual stimulus to the overall IAT measure (Anselmi, Vianell, Voci, & Robusto, 2013).

CHAPTER 3

RESULTS

Wright Map

Firstly, 35.10% variance can be explained by the Rasch measurement, which supports unidimensioinality (Bond, T., & Fox, C. M., 2015). This analysis provided, for each facet, its location on the latent trait" Response Speed" (in logits), and a series of statistical indices useful to the facet and its components. Figure 1 shows a graphical representation of how all these elements of the four facets are displayed on the latent trait "Response Speed", which locates the elements according to stimuli's "reorganizability", participants' ability, participants' gender and ease of the task in the case of conditions (compatible and incompatible blocks). Because the logit scale is an interval-level

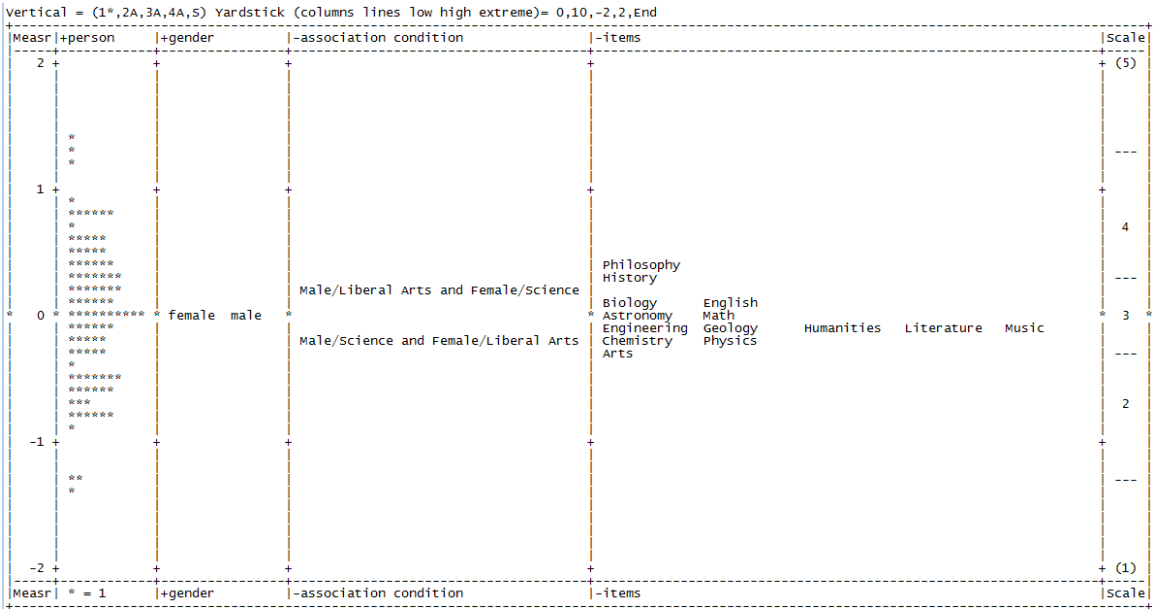

Figure 1. Wright Map: location of different facets on the latent trait "Response Speed"

measurement scale, the equal distances at any point on the vertical scale are of equal size, they represent equal amounts of response time.

Model Fit Statistic

From the overall perspective, both infit and outfit statistic were excellent for the all these four facets person (infit=1.00, outfit=1.01), gender (infit=1.00, outfit=1.01), attribute stimuli (infit=1.00, outfit=1.01), and associative condition (infit = 1.00, outfit=1.01,). There were no misfit items, and only 4 out of the 100 respondents (4%) whose infit or outfit are outside the recommended range (0.5-2).

Respondents completed the IAT with different speeds (range from -1.40 to 1.40; R = 0.91, G =3.13; $\chi 2$ (99) = 852.7, $p < 0.001$). The stimuli were also categorized with different speeds (range from -0.26 to 0.42; R =0.88, G =2.71; $\chi 2$ (13) = 113.8, $p < 0.001$). The condition one Male/Science and Female/Liberal Arts was easier than the condition two Male/Liberal Arts and Female/Science; γ_1 =- 0.24; γ_2 = 0.24; R = 0.99, G =10.23; $\chi 2$ (1) = 211.1, $p < 0.001$). The size of the IAT effect was reflected from the distance between the two conditions Δ=0.48, which is significantly different from 0 ($Z=24.00$, $SE=0.02$, $p<0.01$). That is to say, taken all together, respondents implicitly preferred to associate male with science, and women with liberal arts than male with liberal arts, and female with science. While Female participants and male participants have the same measurement logistic value in general ($\beta_{female} = \beta_{male}$ =0; R =1.00, G =0.00; $\chi 2$ (1) =0, $p = 1$). Table 3 provided all this data-model fit information.

Table 3.

Fit Summary Statistic

		Person	Gender	Association Condition	Stimulus
Measure	M	0.00	0.00	0.00	0.00
	SD	0.17	0.02	0.24	0.18
Infit	M	1.00	1.00	1.00	1.00
	SD	0.29	0.06	0.05	0.10
Outfit	M	1.01	1.01	1.01	1.01
	SD	0.33	0.06	0.07	0.12
Separation Reliability (R)		0.91	1.00	0.99	0.88
Separation Ratio (G)		4.50	0.00	10.23	2.71
χ^2		852.7^{***}	0	211.1^{***}	113.8^{***}
df		99	1	1	13

$^{***}p < .001$

Interaction Analysis between Condition and Gender

Then we tended to the analysis of the interactions between elements of different facets to investigate this implicit Gender-Science stereotype further. The first interaction between the facets associative condition and participants' gender showed that both female and male respondents hold implicit male-science/ female-liberal arts stereotype, they significantly preferred associating male with science, and female with liberal arts to associating male with liberal arts, and female with science. ($\Delta_{female} = 0.44$, $Z_{female} = 14.6$, $SE=0.03$, $p<0.001$; $\Delta_{male} = 0.52$, $Z_{male} = 17.33$, $SE=0.03$, $p < 0.001$), but male respondents

held a stronger preference than female respondents ($Z_{male\text{-}female}$ =1.6, *SE=0.05,* $p < 0.05$). In order to compare results with classical analysis, we also calculated the traditional D score, which is in the same pattern of MFRM results: mean D_{femle} *=0.30,* SD_{female} *= 0.42; mean* D_{male} *= 0.39,* SD_{male} *=0.34.*

Table 4

Interaction between Gender and Association Condition

		1. Male/Science and Female/Liberal Arts	2.Male/Liberal Arts and Female/Science	Contrast
Female	ORS	2329	1907	422
	Measure	-0.22	0.22	0.44
	SE	0.03	0.03	0.05
Male	ORS	2317	1849	468
	Measure	-0.26	0.26	0.52
	SE	0.03	0.03	0.05

Note: ORS = observed raw scores.

Interaction Analyses between Condition, Gender, and Item

In order to figure out whether this difference truly reflected participants' implicit attitude, or was due to stimulus words the Gender-Science IAT used. We conducted interaction analyses between the facets attribute stimuli, genders, and associative condition. Firstly, we investigated the interaction between stimulus group level, science and liberal arts. For male respondents, the IAT effect computed on science stimuli is significantly stronger than that computed on liberal arts stimuli ($\Delta_{science}$ *=0.57,* $\Delta_{liberal\ art}$ *=0.44,* $Z_{science\text{-}liberal\ arts}$ *= 2.6, SE=0.05,* $p < 0.005$). Thus, their implicit preference for

science relative to art is driven most by associating science stimuli with male people, rather than art stimuli with female people. For female respondents, on the contrary, liberal arts stimuli contributed more to IAT effect than science stimulus, but this difference is not significant ($\Delta_{science} = 0.4$, $\Delta_{liberal\ art} = 0.45$, $Z_{science\text{-}arts} = -0.05$, $SE = 0.05$, $p = 0.158$).

Table 5

Interaction between Gender, Association Condition, and Categorical Stimulus

			1.Male/Science and Female/Liberal Arts	2.Male/Liberal Arts and Female/Science	Contrast
Male	Science	ORS	1197	935	262
		Measure	-0.34	0.23	0.57
		SE	0.05	0.05	0.05
	Art	ORS	1120	914	206
		Measure	-0.17	0.27	0.44
		SE	0.05	0.05	0.05
Female	Science	ORS	1184	985	199
		Measure	-0.25	0.15	0.40
		SE	0.04	0.04	0.04
	Art	ORS	1145	922	223
		Measure	-0.17	0.28	0.45
		SE	0.05	0.05	0.05

Note: ORS = observed raw scores.

Then we investigated the interaction on single item level, Table 6 and Table 7 provided information concerning the differential stimulus functioning, separately for the female and male respondents. For each individual stimulus, it is shown whether its overall speed of categorization (i.e., estimated across the two associative conditions) changes according to the specific associative condition the stimulus is presented in. This interaction analysis allowed us to investigate the contribution of each individual stimulus to the overall implicit measure.

The speed with which female respondents categorized the Liberate Arts word, history, decreased in the condition Male-Science/Female-Liberal Arts, and increased in the condition Male-Liberal Arts/Female-Science. This was the stimuli word that female respondents tended to associate more closely with male. Therefore, history, was the stimuli that most contributed to decreasing the implicit male-science/female-liberal arts stereotype observed in female respondents, $t\ (98) = -2.29,\ p<0.05$.

For male respondents, stimulus Math ($t\ (98) = 3.66,\ p<0.01$) and Music ($t\ (98) = 2.12,\ p<0.05$) own differential stimulus functioning. The speed of categorization of these stimuli increased in the condition Male-Science/Female-Liberal Arts and decreased in the condition Male-Liberal Arts/Female-Science. Male respondents associated math more easily with male, and music more easily with female. Thus, Math and Music were the stimuli that had most contribution most to increasing implicit male-science/female-liberal arts stereotype observed in male respondents

Table 6

Speed of Categorization of Stimuli in Two Associative Condition for Female Respondents

Stimulus	1.Male/Science and Female/Liberal Arts			2.Male/Liberal Arts and Female/Science					
	ORS	MSR	SE	ORS	MSR	SE	t	*df*	Cohen's d
Literature	175	0.11	0.12	126	-0.13	0.12	1.35	98	.27
Humanities	185	0.27	0.13	131	-0.05	0.12	1.80	98	.36
Arts	193	0.40	0.13	149	0.21	0.12	1.08	98	.22
Engineering	167	-0.01	0.12	132	-0.04	0.12	0.15	98	.03
Music	169	0.02	0.12	135	0.01	0.12	0.07	98	.01
Physics	177	0.14	0.12	145	0.15	0.12	-0.07	98	-.01
Chemistry	174	0.09	0.12	145	0.15	0.12	-0.34	98	-.07
Astronomy	163	-0.07	0.12	135	0.01	0.12	-0.45	98	-.09
Math	170	0.03	0.12	143	0.12	0.12	-0.53	98	-.11
Geology	172	0.06	0.12	146	0.17	0.12	-0.60	98	-.12
Philosophy	136	-0.46	0.12	112	-0.35	0.13	-0.61	98	-.12
Biology	161	-0.10	0.12	139	0.06	0.12	-0.97	98	-.20
English	150	-0.26	0.12	138	0.05	0.12	-1.81	98	-.37
History	137	-0.45	0.12	131	-0.05	0.12	-.2.29*	98	-.46

Note: ORS = observed raw scores; MSR = measure. The t values test the hypothesis that the difference between the measure is equal to zero. Cohen's $d = \frac{2t}{\sqrt{df}}$.

*$P < 0.05$

Table 7

Speed of Categorization of Stimuli in Two Associative Condition for Male Respondents

Stimulus	1.Male/Science and Female/Liberal Arts			2.Male/Liberal Arts and Female/Science					
	ORS	MSR	SE	ORS	MSR	SE	t	df	Cohen's d
Math	181	0.27	0.13	108	-0.42	0.14	3.66**	98	.74
Music	183	0.31	0.13	128	-0.08	0.13	2.12*	98	.43
Geology	176	0.19	0.13	130	-0.05	0.13	1.31	98	.26
Philosophy	143	-0.33	0.12	100	-0.57	0.14	1.30	98	.26
Engineering	181	0.27	0.13	138	0.08	0.12	1.08	98	.22
Chemistry	180	0.25	0.13	144	0.17	0.12	0.48	98	.10
English	158	-0.10	0.12	125	-0.13	0.13	0.16	98	.03
Physics	179	0.24	0.13	149	0.24	0.12	-0.04	98	-.01
Biology	144	-0.31	0.12	120	-0.21	0.13	-0.58	98	-.12
Arts	173	0.14	0.13	152	0.29	0.12	-0.86	98	-.17
Literature	167	0.04	0.13	146	0.20	0.12	-0.89	98	-.18
Humanities	161	-0.05	0.12	142	0.14	0.12	-1.08	98	-.22
History	135	-0.45	0.12	121	-0.19	0.13	-1.45	98	-.29
Astronomy	156	-0.13	0.12	146	0.20	0.12	-1.88	98	-.38

Note: ORS = observed raw scores; MSR = measure. The t values test the hypothesis that the difference between the measure is equal to zero. Cohen's $d = \frac{2t}{\sqrt{df}}$.

$^{*}P < 0.05$, $^{**}P < 0.01$

CHAPTER 4

DISCUSSION

This article investigated the implicit "Gender-Science" stereotype of female and male individuals. Consistently with the former researches (Chambers, 1983; Benbow, Lubinski, Shea, & Eftekhari-Sanjani, 2000; Farland-Smith, 2009; Steffens, Jelenec, & Noack, 2010), we found that both male and female respondents preferred to connecting male with science, and female with liberal arts, and this preference was stronger for the male participants than for the female participants. Investigating at the contribution of every single stimulus, we found that the strong preference of associated male with science and female with liberal arts observed in male participants was mostly driven by the attribution of associating male with science rather than associating female with liberal arts. Differently, the weaker preference of female respondents was not particularly driven by either associating male with science or associating female with liberal arts. Although insignificantly, liberal arts stimuli contributed more to IAT effect than science stimulus.

This research also found that stimulus used in "Gender-Science" IAT had different IAT effects to the overall test. For male participants, word "math" and "music" increases the IAT effect, which means compared to other stimuli, "math" is a more typical word to represent the concept "science", and "music" is a more typical word to represent the concept "liberal arts". For female participants, there was no stimulus increased IAT effect, but word "history" tended to decrease it. In other words, female

respondents did not consider history as a typical “liberal arts” subject. They tended to associate history with male rather than female.

Theoretical Implication

Inconsistent with some earlier research (Nosek, Banaji, & Greenwald, 2002; Ma, & Liang, 2008), this research found there was a gender difference in implicit male-science/ female-liberal arts stereotype between female and male participants. While Nosek, et al., (2002) found no gender difference in the magnitude of this effect was obtained; both men and women showed implicit math–gender stereotypes equally. While Ma, et al (2008) used Chinese students as participants found that the male-math stereotype IAT effects of female students were significantly larger than male students.

This result may be due to reduced gender stereotype intervention methods that were widely used in education area recently. These popular methods were aiming at reducing negative stereotype influences to women, which may pay less attention to men’s bias. One of the most popular and efficient methods is to providing role models to women. Marx and Roman (2002) used a competent female experimenter to administer a mathematical test, female subjects performed better than the test was conducted by a male experimenter. And when the female experimenters were in the high math competence level, the subject reported higher self-appraised math ability. McIntyre, Paulson and Lord (2003) just provided subjects some papers which talked about the success of women in “untraditional fields”, such as architecture, law, medicine and invention, they performed significantly better on mathematical tests. The usefulness of this method got proven in subsequent researches (McIntyre et al., 2005).

Stout, Dasgupta, Hunsinger, & McManus (2011) found contacting with same-sex experts in academic environments in STEM, not only improved subjects' math performances, but also enhanced women's self-concepts in STEM, attitudes toward STEM and motivation to pursue STEM careers. This strategy is effective even in cross-cultural conditions, Song and Liu (2014) used 150 female high school students as subjects in China, the experimental group used two female models as experimenters who were outstanding college students majored in math. The experimental group performed much better on math test and working memory test.

Using same experimental paradigm, IAT, to test similar topic, these studies got different results on gender effects. Which may due to the procedure of these researches and IAT itself. And the sample of this research is relatively small when compared with former researches. That is another possible reason for a different result.

Although this research did not get a similar result with former researches which used the same Gender-Science IAT, its finding still provides corroborative evidence to the existence of gender-science stereotype. In particular, with a large, heterogeneous sample, researchers (Nosek, & Smyth, 2011) observed that women were less implicitly favorable toward math than were men. Further, these implicit measures own cognitive consistency: women who associated math with male more strongly had more negative attitudes to math than women whose math-male associations were weaker.

Besides, this study found that stereotype of male participants was mostly driven by the attribution of associating male with science rather than associating female with liberal arts, and liberal arts stimuli contributed more to IAT effect than science stimulus to female participants. This result supported the viewpoint that gender-science

stereotypes would likely vary in nuanced ways across fields of study (Miller, Eagly, & Linn, 2015). For example, a large correlational study (n > 100,000) found that biological science majors reported weaker explicit gender-science stereotypes than physical science majors, but they still implicitly associated science with men at the same level (Smyth & Nosek, 2013).

Methodological Implication

This research also has methodological implications. First of all, researchers should be more careful about choosing stimulus of IAT. In the "Gender-Science" IAT, words, "history, math, music" have different contribution when compared with the left stimulus. Former researcher also found that in Race IAT, words "laughter, pleasure, glory, despicable, failure, agony" have different contribution to IAT effect (Anselmi, Vianello, & Robusto, 2011); in Sexuality IAT, words "pleasure, marvelous, lovely, tragic, humiliate, horrible" were different (Anselmi, Vianello, Voci, & Robusto, 2013); in Weight IAT, words "happy, pleasure, joy, evil, failure" had different contribution to overall effect (Anselmi, Vianello, & Robusto, 2013).

Secondly, the results of this study supported the viewpoint that IAT effect can be meaningfully decomposed. Former researchers which used positive and negative words as attribute stimulus found that there is a positive association primacy in the IAT (Anselmi, Vianello, & Robusto, 2011, Anselmi, Vianello, Voci, & Robusto, 2013, Anselmi, Vianello, & Robusto, 2013). Positive association primacy means that responses to negative words decreased the IAT effects, whereas responses to positive words increased it. For this study, the "Gender-Science" IAT did not use positive and negative words as its stimulus, but it found that the traditional gender-science stereotype was

divided into male-science stereotype and female-liberal arts stereotype. And the stereotype of male participants was attributed to associating male with science rather than female with liberal arts.

These findings remind researchers to be more careful when interpreting IAT results. Previous work used the IAT effect as a measure of implicit attitude and prejudice toward different social groups (Hugenberg & Bodenhausen, 2003; Rudman, Ashmore, & Gary, 2001; Rudman, Greenwald, Mellott, & Schwartz, 1999), and evidence also suggested that the measure of implicit prejudice has good convergent and discriminant validity (Gawronski, 2002). However, these results limit the generalizability of this interpretation (van der Maas, & Wagenmakers, 2010). Having a clear and deep understanding of targeted attitudes or prejudice is necessary. Otherwise, the IAT effect of a specific social group may be misleadingly interpreted. For instance, this study showed "Gender-Science" IAT, which is typically used to measure individual implicit stereotype, provided a measure of implicit ingroup-science favoritism for male participants.

These results also contribute to the debate on the nature of the IAT effect. Usually, the IAT is considered as providing an implicit measure of an object (e.g., black people) compared with another object (e.g., white people). Although IAT has been accepted as a useful paradigm and applied to many domains to test people implicit attitudes. It still has a limitation, the experimental design of IAT procedure cannot promise completely independent responses to one of the four categories of stimuli involved in the process. As a result, many other implicit measurements have been developed to overcome this limitation. Some of them are variants of the IAT, such as the

Single Category IAT (Karpinski & Steinman, 2006), the Single Target IAT (Bluemke & Friese, 2008). Whereas others are completely different, like the Go/No-Go Association Task (Nosek & Banaji, 2001), and the SPF task (Bar-Anan et al., 2009).

Finally, the article proposes MFRM that well suits the analysis of the IAT. The MFRM provides many advantages over traditional scoring procedures. When the data fit the MFRM produces requirement, it can provide detailed fit indexes of each element (e.g., stimuli, respondents, and conditions of association), of their spread along the continuum of possible scores (G), and of the reproducibility of their rank ordering (R). By interaction analysis, this model also allows us to test differences of effects on group and interindividual level, which is significant for us to find out the contribution of each stimulus to the overall IAT measure.

Limitation and Future Research

Although this research provided some evidence that the size of the IAT effect can be divided, there are still some limitation of present study. Firstly, the number of participants was rather small. The joint maximum likelihood estimates of model parameters, which are computed in a MFRM analysis, will produce some estimation bias (i.e., the departure of estimates from their ‘‘true’’ values) in small samples (Linacre, 2009b). Secondly, data were collected through Internet, the physical absence of the experimenter may increase the difficulty of ensuring that procedural instructions are clear to all respondents and that the task is performed in the proper way. Thirdly, because of the method of data collection, respondents to the Gender-Science IAT were self-selected and are not representative of any particular population. The second methodological limit of this research is. Therefore, future research should adopt more participants to test and

verify our results. Besides, other aspects of measurement might be taken into account in the future, such as validity, reliability, and predictive validity. It is worth investigating whether the relationship between valence, processing speed, and accuracy is due to individual effects of words in the IAT.

In summary, results of our studies suggest that researchers should be careful when choosing words as stimulus in IAT, and interpreting the IAT effects as if they were equally influenced by different associations. IAT effects should not be interpreted as unambiguous measures of implicit prejudice and that associations involved in the IAT effects might be effectively decomposed.

References

Ambady, N., Parker, S. K., Steele, J., Owen-Smith, A., & Mitchell, J. P. (2004). Deflecting negative self-relevant stereotype activation: The effects of individuation. Journal of Experimental Social Psychology, 40, 401–408.

Anderson, D., Irvin, S., Alonzo, J., & Tindal, G. A. (2015). Gauging item alignment through online systems while controlling for rater effects. *Educational Measurement*: Issues and Practice, 34(1), 22-33.

Andrich, D. (1988). A general form of Rasch's extended logistic model for partial credit scoring. *Applied Measurement in Education*, 1(4), 363-378.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, 42(1), 7.

Anselmi, P., Vianello, M., & Robusto, E. (2011). Positive associations primacy in the IAT A many-facet Rasch measurement analysis. *Experimental Psychology*, 58(5), 376-384.

Anselmi, P., Vianello, M., & Robusto, E. (2013). Preferring thin people does not imply derogating fat people. A Rasch analysis of the implicit weight attitude. *Obesity*, 21(2), 261-265.

Anselmi, P., Vianello, M., Voci, A., & Robusto, E. (2013). Implicit sexual attitude of heterosexual, gay and bisexual individuals: Disentangling the contribution of specific associations to the overall measure. *Plos One*, 8(11).

Asendorpf, J. B., Banse, R., & Mücke, D. (2002). Double dissociation between implicit and explicit personality self-concept: the case of shy behavior. *Journal of personality and social psychology*, 83(2), 380.

Ashburn-Nardo, L., Knowles, M. L., & Monteith, M. J. (2003). Black Americans' implicit racial associations and their implications for intergroup judgment. Social Cognition, 21(1), 61-87.

Bar-Anan, Y., Nosek, B. A., & Vianello, M. (2009). The sorting paired features task: A measure of association strengths. *Experimental Psychology*, 56(5), 329-343.

Baron, A. S., & Banaji, M. R. (2006). The development of implicit attitudes evidence of race evaluations from ages 6 and 10 and adulthood. *Psychological Science*, 17(1), 53-58.

Banse, R. (1999). Automatic evaluation of self and significant others: Affective priming in close relationships. Journal of Social and Personal Relationships, 16(6), 803-821.

Benbow, C. P., Lubinski, D., Shea, D. L., & Eftekhari-Sanjani, H. (2000). Sex differences in mathematical reasoning ability at age 13: their status 20 years later. *Psychological Science*, 11, 474–480.

Bluemke, M., & Friese, M. (2008). Reliability and validity of the Single‐Target IAT (ST‐IAT): assessing automatic affect towards multiple attitude objects. European Journal of Social Psychology, 38(6), 977-997.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20(1), 89-110.

Bosson, J. K., Swann Jr, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology*, 79(4), 631.

Bond, T., & Fox, C. M. (2015). Applying the Rasch model: Fundamental measurement in the human sciences. Routledge.

Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. *Journal of personality and social psychology*, 81(5), 760.

Brown, R. P., & Josephs, R. A. (1999). A burden of proof: Stereotype relevance and gender differences in math performance. *Journal of Personality and Social Psychology*, 76, 246–257.

Brown, R. P., & Pinel, E. C. (2003). Stigma on my mind: individual differences in the experience of stereotype threat. *Journal of Experimental Social Psychology*, 39(6), 626–633.

Cadinu, M., Maass, A., Rosabianca, A., & Kiesner, J. (2005). Why do women underperform under stereotype threat? Evidence for the role of negative thinking. *Psychological science*, 16(7), 572-578.

Carr, P. B., & Steele, C. M. (2010). Stereotype threat affects financial decision making. Psychological Science.

Chambers, D. W. (1983). Stereotypic images of the scientist: The Draw‐a‐Scientist Test. Science education, 67(2), 255-265.

Cunningham, W. A., Johnson, M. K., Raye, C. L., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2004). Separable neural components in the processing of black and white faces. *Psychological Science*, 15(12), 806-813.

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: combating automatic prejudice with images of admired and disliked individuals. *Journal of personality and social psychology*, 81(5), 800.

Dasgupta, N., McGhee, D. E., Greenwald, A. G., & Banaji, M. R. (2000). Automatic preference for white Americans: Eliminating the familiarity explanation. *Journal of Experimental Social Psychology*, 36(3), 316-328.

Davies, P. G., Spencer, S. J., Quinn, D. M., & Gerhardstein, R. (2002). Consuming images: how television commercials that elicit stereotype threat can restrain women academically and professionally. *Personality and Social Psychology Bulletin*, 28(12), 1615–1628.

De Houwer, J. (2003). The extrinsic affective Simon task. *Experimental psychology*, 50(2), 77.

Egloff, B., & Schmukle, S. C. (2002). Predictive validity of an implicit association test for assessing anxiety. Journal of personality and social psychology, 83(6), 1441.

Engelhard, G., & Myford, C. M. (2003). Monitoring faculty consultant performance in the advanced placement English literature and composition program with a many□faceted Rasch model. *ETS Research Report Series*, 2003(1), 60.

Farland‐Smith, D. (2009). Exploring middle school girls' science identities: Examining attitudes and perceptions of scientists when working "side‐by‐side" with scientists. School Science and Mathematics, 109(7), 415-427.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, 54(1), 297-327.

Foogliati, J, V., & Bussey, K. (2013). Stereotype threat reduces motivation to improve: effects of stereotype threat and feedback on women's intentions to improve mathematical ability. *Psychology of Women Quarterly*, 37(3), 310-324.

Franck, E., De Raedt, R., & De Houwer, J. (2007). Implicit but not explicit self-esteem predicts future depressive symptomatology. *Behavior Research and Therapy*, 45(10), 2448-2455.

Gawronski, B. (2002). What does the implicit association test measure? A test of the convergent and discriminant validity of prejudice-related IATs. *Experimental psychology,* 49(3), 171.

Good, C., Aronson, J., & Harder, J. A. (2008). Problems in the pipeline: Stereotype threat and women's achievement in high-level math courses. *Journal of Applied Developmental Psychology*, 29(1), 17-28.

Greenwald, A. G. (1990). What cognitive representations underlie social attitudes? *Bulletin of the Psychologic Society*, 28(3), 254-260.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1).

Greenwald, A. G., & Farnham, S. D. (2000). Using the implicit association test to measure self-esteem and self-concept. *Journal of personality and social psychology*, 79(6), 1022.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6).

Greenwald, A. G., & Nosek, B. A. (2001). Health of the implicit association test at age 3. *Experimental Psychologies*, 48(2), 85-93.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. an improved scoring algorithm. *Journal of Personality and Social Psychology*, 85(2), 197.

Greenwald, A. G., Nosek, B. A., Banaji, M. R., & Klauer, K. C. (2005). Validity of the salience asymmetry interpretation of the implicit association test: Comment on rothermund and wentura (2004).

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the implicit association test: III. meta-analysis of predictive validity. *Journal of Personality and Social Psychology,* 97(1), 17.

Gupta, V. K., & Bhawe, N. M. (2007). The influence of proactive personality and stereotype threat on women's entrepreneurial intentions. *Journal of Leadership & Organizational Studies*, 13(4), 73-85.

Halpern, D. F., Aronson, J., Reimer, N., Simpkins, S., Star, J. R., & Wentzel, K. (2007). Encouraging girls in math and science.

Harasym, P. H., Woloschuk, W., & Cunning, L. (2008). Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. *Advances in Health Sciences Education*, 13(5), 617-632.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin*, 31(10), 1369-1385.

Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice implicit prejudice and the perception of facial threat. *Psychological Science*, 14(6), 640-643.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: a meta-analysis. *Psychological Bulletin*, 107, 139–155.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science*, 321(5888), 494-495.

Jacobs, J. E., & Eccles, J. S. (1992). The impact of mothers' gender-role stereotypic beliefs on mothers' and children's ability perceptions. *Journal of Personality and Social Psychology*, 63(6), 932–944.

Johns, M., Schmader, T., & Martens, A. (2005). Knowing is half the battle: teaching stereotype threat as a means of improving women's math performance. Psychological Science, 16, 175–179.

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology*, 81, 774–788.

Karpinski, A., & Steinman, R. B. (2006). The single category implicit association test as a measure of implicit social cognition. *Journal of Personality and Social Psychology*, 91(1), 16.

Kiefer, A. K., & Sekaquaptewa, D. (in press). Implicit stereotypes, gender identification, and math performance: a prospective study of female math students. *Psychological Science*.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*,19(1), 3-31.

Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding and using the implicit association test: IV. Implicit measures of attitudes, 59-102.

LeFevre, J. A., Kulak, A. G., & Heymans, S. L. (1992). Factors influencing the selection of university majors varying in mathematical content. *Canadian Journal of Behavioural Sciencet*, *24*(3), 276.

Lemm, K., & Banaji, M. R. (1998). Implicit and explicit gender identity and attitudes toward gender. Paper presented at the seventieth annual meeting of the Midwestern Psychological Association, Chicago, IL.

Linacre, J. M. (1989). Multi-faceted Rasch measurement.

Linacre, J. M. (2009). Facets Rasch measurement computer program (version 3.65. 0). Chicago: Winsteps.Com,

Ma, L., & Liang, N. (2008). A Study on Implicit Mathematics Gender Stereotype by IAT. Psychological Science, 31(1) : 35- 39.

Maison, D., Greenwald, A. G., & Bruin, R. (2001). The implicit association test as a measure of implicit consumer attitudes.

Marini, M., Sriram, N., Schnabel, K., Maliszewski, N., Devos, T., Ekehammar, B., . . . Nosekl, B. A. (2013). Overweight people have low levels of implicit weight bias, but overweight nations have high levels of implicit weight bias. *Plos One*, 8(12).

Marx, D.M., & Roman, J. S. (2002). Female role models: Protecting women's math test performance. *Personality and Social Psychology Bulletin*, 28(9), 1183-1193.

Mazerolle, M., Régner, I., Morisset, P., Rigalleau, F., & Huguet, P. (2012). Stereotype threat strengthens automatic recall and undermines controlled processes in older adults. *Psychological Science*, 0956797612437607.

McGarty, C., Yzerbyt, V., & Spears, R. (2002). Social, cultural and cognitive factors in stereotype formation. Stereotypes as explanations: The formation of meaningful beliefs about social groups. Cambridge: Cambridge University Press. pp. 1–15.

McIntyre, R. B., Lord, C. G., Gresky, D. M., Ten Eyck, L. L., Frye, G. D. J., & Bond Jr C. F. (2005). A social impact trend in the effects of role models on alleviating women's mathematics stereotype threat. *Current Research in Social Psychology*, 10(9), 116-136.

McIntyre, R. B., Paulson, R. M., & Lord, C. G. (2003). Alleviating women's mathematics stereotype threat through salience of group achievements. *Journal of Experimental Social Psychology*, 39 (1), 83-90.

Mierke, J., & Klauer, K. C. (2003). Method-specific variance in the implicit association test. *Journal of Personality and Social Psychology*, 85(6), 1180.

Miller, D. I., Eagly, A. H., & Linn, M. C. (2015). Women's representation in science predicts national gender-science stereotypes: Evidence from 66 nations. *Journal of Educational Psychology*, 107(3), 631.

Mitchell, C. J. (2004). Mere acceptance produces apparent attitude in the Implicit Association Test. *Journal of Experimental Social Psychology*, 40(3), 366-373.

Murphy, M. C., Steele, C. M., & Gross, J. J. (2007). Signaling threat: How situational cues affect women in math, science, and engineering settings. *Psychological Science*, 18(10), 879-885.

Nosek, B. A., & Smyth, F. L. (2011). Implicit social cognitions predict sex differences in math engagement and achievement. American Educational Research Journal, 48(5), 1125-1156.

Nosek, B. A., Sutin, E., Hansen, J. J., Wu, L., Sriram, N., Smyth, F. L., & Greenwald, A. G. (2006). Project Implicit: Advancing theory and evidence with technical and methodological innovation. Unpublished manuscript: University of Virginia.

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. Social Cognition, 19(6), 625-666.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math=D male, me = female, therefore math < me. *Journal of Personality and Social Psychology*, 83, 44–59.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the implicit association test: II. method variables and construct validity. *Personality and Social Psychology Bulletin*, 31(2), 166–180.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007a). The implicit association test at age 7: A methodological and conceptual review. *Automatic Processes in Social Thinking and Behavior*, 265-292.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G., & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, 18, 36-88.

Ottaway, S. A., Hayden, D. C., & Oakes, M. A. (2001). Implicit attitudes and racism: Effects of word familiarity and frequency on the implicit association test. *Social Cognition*, 19(2), 97-144.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as an implicit measurement. *Journal of Personality and Social Psychology*, 89(3), 277.

Quinn, D. M., & Spencer, S. J. (2001). The interference of stereotype threat with women's generation of mathematical problem-solving strategies. *Journal of Social Issues*, 57, 55–71.

Rothermund, K., & Wentura, D. (2004). Underlying processes in the implicit association test: Dissociating salience from associations. *Journal of Experimental Psychology*: General, 133(2), 139.

Rudman, L. A., Ashmore, R. D., & Gary, M. L. (2001). " Unlearning" automatic biases: the malleability of implicit prejudice and stereotypes. *Journal of personality and social psychology*, 81(5), 856.

Rudman, L. A., Greenwald, A. G., Mellott, D. S., & Schwartz, J. L. (1999). Measuring the automatic components of prejudice: Flexibility and generality of the Implicit Association Test. *Social Cognition*, 17(4), 437-465.

Sekaquaptewa, D., & Thompson, M. (2003). Solo status, stereotype threat, and performance expectancies: their effects on women's performance. Journal of Experimental Social Psychology, 39, 68–74.

Shapiro, J. R., & Williams, A. M. (2012). The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields. *Sex Roles*, 66(3-4), 175-183.

Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: identity salience and shifts in quantitative performance. *Psychological Science*, 10, 80–83.

Smith, J. L., & White, P. H. (2002). An examination of implicitly activated, explicitly activated, and nullified stereotypes on mathematical performance: it's not just a women's issue. *Sex Roles*, 47, 179–191.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*,35, 4–28.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African-Americans. *Journal of Personality and Social Psychology*, 69, 797–811.

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. *In Advances in Experimental Social* Psychology (pp. 379–440). San Diego, CA: Academic Press.

Steffens, M. C., Jelenec, P., & Noack, P. (2010). On the leaky math pipeline: Comparing implicit math-gender stereotypes and math withdrawal in female and male children and adolescents. *Journal of Educational Psychology*, 102(4), 947.

Stoet, G., & Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement?

Stout, J. G., Dasgupta, N., Hunsinger, M., & McManus, M. A. (2011). Steming the tide: Using ingroup experts to inoculate women's self -concept in science, technology, engineering, and mathematics (stem). *Journal of Personality and Social Psychology*, 100(2), 255-270.

Stone, D. (1999). Learning lessons and transferring policy across time, space and disciplines. *Politics*, 19(1), 51-59.

Song, S., & Liu, H. (2014). Effect of Counter-stereotype Information to Reduce the Effects of Stereotype Threat. *Chinese Journal of Clinical Psychology*, 22 (4), 386-389.

Sriram, N., & Greenwald, A. G. (2009). The brief implicit association test. *Experimental Psychology*, 56(4), 283-294.

Teachman, B. A., Gapinski, K. D., Brownell, K. D., Rawlins, M., & Jeyaram, S. (2003). Demonstrations of implicit anti-fat bias: The impact of providing causal information and evoking empathy. *Health Psychology*, 22(1), 68.

Teachman, B. A., Gregg, A. P., & Woody, S. R. (2001). Implicit associations for fear-relevant stimuli among individuals with snake and spider fears. *Journal of Abnormal Psychology*, 110(2), 226.

Till, H., Myford, C., & Dowell, J. (2013). Improving student selection using multiple mini-interviews with multifaceted Rasch modeling. *Academic Medicine*, 88(2), 216-223.

Tomasetto, C., Alparone, F. R., & Cadinu, M. (2011). Girls' math performance under stereotype threat: The moderating role of mothers' gender stereotypes. *Developmental psychology*, 47(4), 943.

van Ravenzwaaij, D., van der Maas, H. L., & Wagenmakers, E. J. (2010). Does the name-race implicit association test measure racial prejudice?. *Experimental psychology*.

Vianello, M., Anselmi, P., & Robusto, E. (2009). Analysis of evaluative attributes in a race IAT. *Organization Special*, 257(56), 43-52.

Vianello, M., & Robusto, E. (2010). The many-facet Rasch model in the analysis of the go/no-go association task. Behavior Research Methods,42(4), 944-956.

Walton, G. M., & Spencer, S. J. (2009). Latent ability grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science*, 20(9), 1132-1139.

Wheeler, S. C., & Petty, R. E. (2001). The effects of stereotype activation on behavior: a review of possible mechanisms. *Psychological Bulletin*, 127,797–826.

Wiers, R. W., Van Woerden, N., Smulders, F. T., & De Jong, P. J. (2002). Implicit and explicit alcohol-related cognitions in heavy and light drinkers. *Journal of abnormal psychology*, 111(4), 648.

Wright, B., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. Educational and *Psychological measurement*, 29(1), 23-48.

Wright, B. D., & Stone, M. H. (1979). Best test design. Rasch measurement.

Xu, K., Lofaro, N., Nosek, B. A., & Greenwald, A. G. (2017). Gender-Science IAT 2003-2015.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. *Educational Measurement*, 4, 111-153.

Yeung, N. C. J., & von Hippel, C. (2008). Stereotype threat increases the likelihood that female drivers in a simulator run over jaywalkers. *Accident Analysis & Prevention*, 40(2), 667-674.

APPENDIX A

FACET CODE FOR GENDER-SCIENCE STEREOTYPE IAT

```
Title = IAT in MFRM-long format
Arrange = M                ; arrange output tables in Measure ascending order
Facets = 4                 ; four facets: person, gender, condition, items
Umean = 0, 1               ; user-scaling = 0 +logit*1
positive=1,2,              ; person, gender, condition, item greater score mean greater
                             measure，which means they response quick
Gstats=Yes
Model=
?,?B,?B,?,R5               ; observations of items are rating in range 1-5
                           ; look for interaction/bias between association condition and
                             participants' gender
?,?B,?B,?B,R5              ; look for interaction/bias between stimuli words, condition, and
                             gender.
Subset dection = Report ; there should be 2 subsets, detect subsets. Show rulers and list in
                          Table 6. Subset numbers in Table 7
Lables=
1, person
1-100                      ; 100 participants were selected
*
```

2, gender, G

1, female, 0, 1

2, male, 0, 2

*

3, association condition

1, Male/Science and Female/Liberal Arts

2, Male/Liberal Arts and Female/Science

*

4, items

1, Astronomy

2, Math

3, Chemistry

4, Physics

5, Biology

6, Geology

7, Engineering

8, History

9, Arts

10, Humanities

11, English

12, Philosophy

13, Music

14, Literature

*

Data =