University of Nevada, Reno

A Behavior Analytic Account of Stereotype Threat

A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Psychology

by

Lauren Brown, M.A., BCBA

Bethany Contreras, Ph. D/Dissertation Advisor

August, 2022

THE GRADUATE SCHOOL

We recommend that the dissertation
prepared under our supervision by

Lauren Brown

entitled

A Behavior Analytic Account of Stereotype Threat

be accepted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

Bethany Contreras, Ph.D.
Advisor

W. Larry Williams, Ph.D.
Committee Member

.Steve Rock, Ph.D.
Committee Member

Matthew Lewon, Ph.D.
Committee Member

Markus Kemmelmeier, Ph.D.
Graduate School Representative

David W. Zeh, Ph.D., Dean
Graduate School

August, 2022

Abstract

Although behavior analysis has contributed substantially to the understanding and study of learning in humans, cultural influences are often either overlooked or not accounted for in how they impact individuals in their day-to-day lives. One example in which this has occurred is in accounting for stereotypes. The field of Social Psychology has contributed a significant body of research on stereotypes and discusses in detail the conditions under which individuals are likely to be impacted by stereotypes. One common finding, often referred to as stereotype threat (Steele & Aronson, 1995), refers to how stereotypes can negatively impact individual performances under certain testing conditions. While data on stereotype threat indicates a clear pattern of decreases in performance scores for individuals in the threatened group, studies on stereotype threat have not examined: 1) whether stereotype threat occurs when arbitrary, non-stereotyped tasks are presented, 2) trends in individual data, or 3) how each individual is impacted by threat, lift, and neutral statements across similar tests. In addition, although researchers have offered many assumptions why stereotype threat occurs, none have evaluated the function of language in stereotype threat (c.f., Relational Frame Theory; Hayes, Barnes-Holmes, & Roche, 2001). The current study aimed to examine whether stereotype threat and stereotype lift by group affiliation (i.e., gender) would occur on an arbitrary, computer-based memory test and if other test-taking behaviors were affected by performance differences across four studies. Results indicated overall patterns consistent with the research base. Typical stereotype threat and lift patterns emerged more frequently when longer scripts were provided to participants prior to testing.

Keywords: Stereotype Threat, Stereotype Lift, Relational Frame Theory (RFT)

Acknowledgment

Art Williams stated: “I am not telling you it is going to be easy – I am telling you it is going to be worth it.” This project would not have been worth the time, effort, and sleepless nights without the help of many individuals. First off, to my advisors. I would like to thank Dr. Bethany Contreras, not only for her help in advising me on this project, but for creating the opportunity for me to research an area of interest outside the field and being willing to join me on this journey. As well, I want to thank Dr. Larry Williams for his guidance over the years in early versions of this project and computer program. Without their guidance and support, this research would never have been possible.

I would also like to thank Dr. Scott Miller, who spent many hours creating complex formulas to enable easier data analysis across thousands of different data points for this research. As both my partner in geeking out on science and in life, I thank you for all the support and love you have always given me.

Additionally, I would like to thank my mom, sibling, and grandmother, who continually encouraged me in my educational endeavors and supported me thought years and years of schooling. Without such a wonderful, supportive family, I know I would not be where I am today. For that, I thank each and every one of you so much.

Lastly, I would like to thank my committee, lab members, friends, and research assistants who helped provide direction, problem solve, and with data collection. No project is successful without collective problem solving, and I greatly appreciate this support.

Thank you all for encouraging me throughout this journey and making it all worth it!

Table of Contents

List of Tables

List of Figures

A Behavior Analytic Account of Stereotype Threat

In a panel discussion at the 2014 Center for Inquiry Conference, an audience member asked astrophysicist Neil Degrasse Tyson about the low rates of women in science (STEM) fields, and whether this was related to genetic differences between males and females (Angel, 2014). Although he (admittedly) is not female, he likened the issue to similar discussions about genetic differences between people of color and Caucasians. He stated: "The fact that I wanted to be an […] astrophysicist was hands down the path of most resistance through the forces […] of society. I wanted to become something that was outside of the paradigms of expectation". Although Degrasse Tyson is now a well-known astrophysicist, he noted that all of the barriers he faced throughout his life have continued to keep out or limit people of color and women from careers in the sciences. He argued:

> My life experiences tell me that when you don't find Blacks in the sciences, you don't find women in the sciences… I know that these forces are real because I had to survive them in order to get to where I am today. So, before we start talking about genetic differences, you got to come up with a system where there is equal opportunity (Angel, 2014).

Everyone's histories are riddled with gendered assumptions as to how people should behave, what careers are appropriate, what people should learn/pursue, and so on. Although choosing classes, majors, and careers to pursue may seem like non-gendered tasks, "forces of society", as Degrasse Tyson puts it, are consistently guiding everyone's paths toward or away from certain options whether people are aware of them or not. One area in particular where gendered assumptions frequently emerge is in the classroom, especially when tests are administered. Barriers to learning and how well one performs in this context are often unaccounted for, dismissed, or poorly understood, such as how verbal behavior can influence performance (for example, in the form of stereotypes). This study aims to define and discuss the concept of stereotype threat as it relates to the performances of different groups of people on various academic tests, how these stereotypes operate from the perspective of relational frame theory (RFT), evaluate the effects of stereotype threat and stereotype lift by providing contextual cues during memory testing across men and women, and to discuss the potential societal implications that stereotype threat and lift present.

Stereotype Threat

Social Psychologists have been addressing and researching the concept of stereotype threat since Steele and Aronson first coined the term in 1995. They defined stereotype threat at that time as "being at risk of confirming, as self-characteristic, a negative stereotype about one's own group" (Steele & Aronson, 1995, p. 797). Researchers have since expanded their definitions (Spencer et al., 1999; Steele, 1997), such as Spencer and colleagues (2016), who argue that stereotype threat occurs:

> When members of a stigmatized group find themselves in a situation where negative stereotypes provide a framework for interpreting their behavior, the risk of being judged in light of those stereotypes can elicit a disruptive state that undermines performance and aspirations in that domain. (p. 415)

Thus, in educational settings in particular, stereotype threat can potentially manifest as gaps in performances of people in stigmatized groups, despite how frequently certain content is taught to all students.

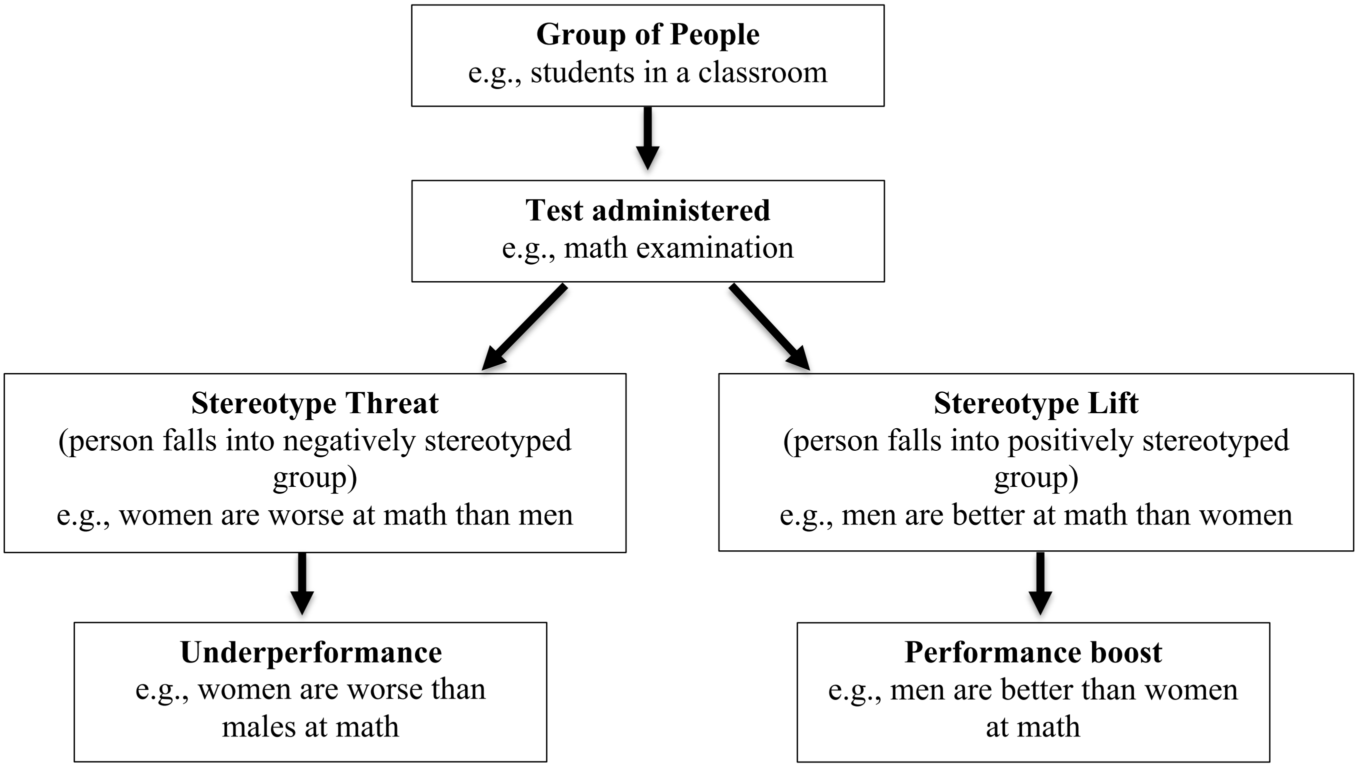
To illustrate, let's use a common stereotype in many cultures: men are better at math than women. During math tests, stereotype threat helps to explain how women might underperform on the same exam in comparison to their male counterparts, even when variables are controlled for such as the amount of effort put into preparing for the exam, note taking, and attendance at lectures. Walton and Cohen (2003) explain: "negative stereotypes about women's ability in math, and about racial minorities' intellectual abilities in general, are so ubiquitous that evaluative test can trigger among members of these groups the fear that […] they could confirm a negative stereotype about their [group]" (p. 456). The presence of stereotypes under such conditions could lead to changes in the stereotyped persons behavior, either overtly in the form of taking longer to answer questions, rushing through questions, or not answering some or all questions, and covertly in the form of second guessing one's answers, repetitive thoughts/evaluations of oneself, as well as the experience of anxiety or other negative thoughts/emotions. At the heart of stereotype threat, as many researchers have analyzed in some capacity through the lens of specific sociological or social psychological theories, is the changes in one's behavior in response to a stereotyped situation.

Stereotype Lift

Stereotype threat is only a single domain of a larger stereotype context that can affect performance. The stereotype context can also include an important effect called stereotype lift. Walton and Cohen (2003) define stereotype lift as "the performance boost caused by the awareness that an outgroup is negatively stereotyped" (p. 456). Stereotype lift, also called "counterstereotype threat" (e.g., Forbes & Schmader, 2010) or "stereotype boost" (e.g., Crisp et al., 2009), describes the behavior pattern of the non-stereotyped group in a situation in which a stereotype threat is present. To return to the previous example, while the stereotype "men are better at math than women" negatively impacts women during math tasks, the inverse occurs for men. That is, men are likely to perform better than their women counterparts when the stereotype favors their own group. Figure 1 below depicts the concepts of stereotype threat and stereotype lift.

Figure 1

Depiction of the context in which stereotype threat/lift are likely to occur for males and females in the context of a math examination


Group of People
e.g., students in a classroom
Test administered
e.g., math examination
Stereotype Threat
(person falls into negatively stereotyped group)
e.g., women are worse at math than men
Stereotype Lift
(person falls into positively stereotyped group)
e.g., men are better at math than women
Underperformance
e.g., women are worse than males at math
Performance boost
e.g., men are better than women at math

To depict how stereotype threat/lift might occur, who stereotype threat/lift affects, and why it may affect these individuals, we will first delve into the extensive research on stereotype threat/lift, followed by a discussion on how a behavior analytic account of stereotype threat/lift might operate and the potential mechanisms by which such problematic stereotypes may continue to exist.

The Ins and Outs of Stereotype Threat

Research has shown that there are specific conditions that make stereotypes more likely to impact the performances of particular individuals. The following section will address the individuals or groups of individuals that tend to be the most impacted by stereotype threat.

Who Has Stereotype Threat Been Shown to Impact?

African Americans and minority groups. Steele and Aronson (1995) first discussed stereotype threat as it related to the differences they observed in the performances of two particular groups of individuals: White Americans versus African Americans. Although the achievement gap between Caucasian and African American students has been discussed frequently in research (Humphreys, 1975; Lawrence et al., 2010; Linn, 1973; Stanley, 1971), indicating that White students generally perform better than their African American counterparts, many factors contributing to this outcome are often unaccounted for or even misunderstood. For instance, Jensen (1980) found that among Black and White students who had similar schooling experiences and who achieved the exact same score on their Scholastic Aptitude Tests (SAT), the Black students were more likely to have poorer outcomes (e.g., lower GPA's, time until they graduated, retention rates). Steele and Aronson (1995) argued that this phenomenon could best be conceptualized as “the overprediction or underachievement phenomenon, because it indicates that, relative to Whites with the same score, standardized tests actually overpredict the achievement that Blacks will realize” (p. 798). They conclude that: “this evidence suggests that Black-White achievement gaps are not due solely to group differences in preparation” (p. 798), indicating the need for a better understanding of the situational and cultural factors that are constantly influencing students.

Other research has reported similar findings with respect to the impact of stereotype threat on the Black-White achievement gap (Alter et al., 2010; Lawrence et al., 2010; Mendes et al., 2002; Nadler & Clark, 2011; Nguyen & Ryan, 2008; Purdie-Vaughns et al., 2008; Walton et al., 2013; Walton & Spencer, 2009).

Additionally, similar results have also been seen with the Latinx population (Armenta, 2010; Gonzales et al., 2002). For instance, Armenta (2010) argued that for students whose social identities were strongly identified with the Latinx (Hispanic) culture, performance on a math test decreased when compared with another minority group (i.e., Asian students). Data also indicate that stereotype threat impacts other minority groups, such as French-Arabs (Chateignier et al., 2009) and Ugandans (Picho & Schmader, 2018), to name a few.

Women and gender minorities. Another commonly researched demographic discussed in stereotype threat research is that of biological sex and sometimes gender. In general, research will often report on the score differences on various tests between biological males in comparison to the scores of biological females. Similar to the reported Black-White achievement gap, the average female versus male score on any given academic task have historically and are still sometimes interpreted to indicate some innate, biological difference between males and females with respect to that particular academic task. For instance, Benbow and Stanley (1980) concluded from their data on students who were identified to enter the Study of Mathematically Precocious Youth (SMPY) program, that males were naturally better at math because their data showed males outperforming females on a difficult math test. One critical issue with Benbow and Stanley's research, however, was that the male participants vastly outnumbered the female participants (at about a 10 to 1 ratio). Still, in summarizing their results, the authors state: "we favor the hypothesis that sex differences in achievement in and attitude toward mathematics result from superior male mathematical ability, which may be in turn related to greater male ability in spatial tasks. This male superiority is probably an expression of endogenous and exogenous variables" (p. 1264).

Many researchers, however, have been skeptical of explanations of learned behaviors (i.e., those behaviors that require teaching in order to be able to do effectively) as solely resulting from unlearned, biological features. For instance, Spencer and colleagues (1999) discussed this concept further by examining male and female performance on mathematical questions under conditions where participants were reminded of the stereotype that men were better prior to completing the tests (gender difference) and where no gender difference was indicated. Data showed that, under conditions where no gender difference was salient or indicated, the gap between the scores of the male and female participants diminished, both amongst groups of men and women who were well-versed in math and those who were less well-versed in math. The authors argue that "being the potential target of a negative group stereotype […] creates a specific predicament: in any situation where the stereotype applies, behaviors and features of the individual that fit the stereotype make it plausible as an explanation of one's performance" (p. 21).

Although trends have been observed that show the gap between men and women's performance closing slowly over time (Cole, 1997; Feingold, 1988), mainly where it relates to grades in classes, researchers have argued that "despite the gains, women still underperform on some standardized tests and still are less likely than men to major in math and science or enter careers that demand these skills" (Quinn & Spencer, 2001, p. 68). Much research contends that stereotype threat continues to not only be observed, but also remains a problem (e.g., Hyde et al., 1990; Quinn & Spencer, 2001; Spencer et al., 1999, 2016; Walton & Spencer, 2009).

Schmader (2002) attempted to further examine the role of gender in stereotype threat by looking at his participant's gender identification and comparing those self-identifications with their respective performances on a mathematics test. They found that, for women, when gender identity was "an important part of their self-definition" (p. 199), women were more likely to perform worse on a math test than women who didn't see their gender identity as an important part of how they defined themselves. The reverse occurred for the men sampled: those who identified more as "men" performed better than those who didn't see being a "man" as an important part of their self-identity. Thus, women who identified more prominently as "women" performed worse than their male counterparts, which depicts and is directly in line with the research on stereotype threat, while men who identified as "men" performed better than their female counterparts, depicting and in line with research on stereotype lift. Schmader concludes from their research that: “although there was a hint of such gender differences in performance […], the [data] suggest that additional variance in performance can be explained by examining the moderating effects of situational cues to […] individual differences in gender identification” (p. 199).

Other researchers have reported similar findings to Schmader (2002) (who collected data in the U.S.) in both Germany (Keller & Dauenheimer, 2003) and the Netherlands (Marx et al., 2005). However, Swedish samples showed a different pattern. Eriksson and Lindholm (2007) found that the women in their samples who had higher self-identifications with being women in a situation where negative stereotypes of women’s performance were present (i.e., mathematics exam) performed better than those who didn’t identify as strongly with being a woman. This is in direct contrast to what Schmader (2002), Keller and Dauenheimer (2003) and Marx et al. (2005) observed. Eriksson and Lindholm argue that cultural differences could play a factor in their data, as Sweden tends to both promote and support more notions of gender equality in the workplace, classroom, and home.

Although Eriksson and Lindholm’s (2007) findings are interesting, the vast majority of research depicts stereotype threat as a common phenomenon that is experienced by many different groups of people (e.g., Aronson et al., 1999; Ben-Zeev et al., 2005; Davies et al., 2002; Keller & Dauenheimer, 2003; O’Brien & Crandall, 2003a; Pronin et al., 2004; Sekaquaptewa & Thompson, 2003; Spencer et al., 2016)

The Consequences of Stereotype Threat

Thus far, some of the research on stereotype threat has shown overall that during stereotyped contexts, we are likely to observe a number of different behaviors and short/long term consequences. In their review of research to date on stereotype threat, Spencer and colleagues (2016) argue that this “finding holds across diverse stereotype-threat manipulations, test types, and targeted groups (e.g., African Americans, Latino Americans, Turkish Germans, and women)” (p. 422). While short term consequences, such as an immediate test results being lower than one had hoped for, can lead to statements or thoughts of “I’m not good enough” or “I can’t do this”, long term effects of continued thoughts and beliefs about one’s ability can lead to many different outcomes, such as lower rates of appropriate placements in classes, advanced placements, college or university admittance, as well as an increased likelihood of dropping classes, changing majors or professional interests, withdrawing entirely from college (Osborne, 2001), fewer women and minorities in academia ,leadership roles and self-esteem regarding those leadership roles (Burnette et al., 2010; Hoyt et al., 2010), career satisfaction, attitudes, and intentions in relation to careers chosen (von Hippel et al., 2011). Real world data unfortunately confirm these likelihoods: although women’s participation in STEM fields has increased over time, women still remain underrepresented and are paid less than men. In a U.S. Census Bureau report, Martinez and Christnaucht (2021) state: “Women made gains – from 8% of STEM workers in 1970 to 27% in 2019 – but men still dominated the field. Men made up 52% of all U.S. workers but 73% of all STEM workers.” While stereotype threat alone will not explain the complexities involved in women being underrepresented in STEM fields or the wage-gap between men and women in general, it does offer an opportunity to consider how individual learning histories and language around gender performance differences can impact people on a much larger scale.

Research has shown the impact of stereotype threat on individuals' behaviors during, prior to, and after testing. McGlone and Aronson (2006) argue that "the predicament of being stereotype-threatened can overwhelm factors such as skill, preparedness, and cultural background, factors to which academic performance is customarily attributed" (p. 487). For instance, participants might self-handicap (Keller, 2002), when a threat is present by spending less time studying or preparing for the test (Stone, 2002), or by attempting fewer questions during testing (Davies et al., 2002; Steele & Aronson, 1995). Participants might also lower their expectations (Cadinu et al., 2003), or become more vigilant for signs or behaviors associated with failure (Schmader et al., 2009). Even more alarmingly, conservative estimates from a meta-analysis by (Walton et al., 2013) suggest that, on the SAT-Math test specifically, stereotype threat can account for: 57-94% of the gender gap, 23-39% of the White-Latino gap, 17-28% of the Black-White gap. Even in situations where threats might not lead to decreased performance (as noted with easier tasks or when asking test takers to perform tasks with which they are fluent), the fact that there is some type of change in performance due to these statements is worth noting and examining further. Hence, it is vitally important for us to not only understand the contexts in which stereotype threat/lift is likely to occur, but also how we can arrange conditions under which stereotype threat is less likely to have an impact on both short- and long-term behaviors of individuals who are affected by stereotype threat.

Is Stereotype Threat Consistently Observed?

We would be remiss if we didn't also discuss studies which have not found a basis for the observation of stereotype threat/lift, but who have instead argued for the null hypothesis, or when studies reporting stereotype threat have supposedly inflated their findings due to publication or experimenter bias (Flore & Wicherts, 2015; Pennington et al., 2016, 2019). Rather, they report that when different statistics are run on reported data from previous research, the differences between participant scores in the control versus threat conditions "range from negligible to small" (Shewach et al., 2019, p. 1529). Additionally, proponents of the null hypothesis in stereotype threat research argue that certain experimental arrangements can also influence whether or not stereotype threat is observed, such as the gender of the experimenter (Pennington et al., 2016; Stone & McWhinnie, 2008).

Although the debate of whether stereotype threat can explain gaps in performance between different groups is important to our discussion, we will elaborate later on the impact of both positions in favor of and against stereotype threat in relation to our data.

How Does Stereotype Threat Operate and When Does it Occur?

Many researchers have attempted to discuss why stereotype threat might occur. Spencer and colleagues (1999) argued that

> Possibly because communicative processes play such a central role in the acquisition of stereotypes…- that is, public and private discourse, the media, school curricula, artistic canons, and the like- knowledge of them is widely disseminated throughout a society, even among those who do not find them believable. This means that people who are the targets of these stereotypes are likely to know them too. And herein lies the threat. In situations where the stereotype applies, they face the implication that anything they do or any feature they have that fits the stereotype makes it more plausible that they will be evaluated based on the stereotype. (pp. 5-6)

As has been discussed previously, whenever a threat is present (whether it be blatant, subtle, explicitly stated, etc.), it is likely that individual performance will be impacted if the person falls into the stereotyped group. Further details regarding how this might happen are discussed hereafter.

Some research in social psychology has argued that stereotype threat is not pervasive, but rather that it occurs primarily under specific conditions. For example, in order to observe the effects of stereotype threat, “it assumes that the test taker construes the test as a fairly valid assessment of […] ability, that they still care about this ability at least somewhat, and that the test be difficult” (Spencer, Steele, & Quinn, 1999, p. 25). In general, researchers have concluded that in order for stereotype threat to be observed, the test taker must be challenged in some way by the task at hand (perhaps because the task itself is something the test taker is less fluent in or familiar with), and that the task itself might evoke frustration from the test taker (Steele, 1997). In a review of research conducted on stereotype threat, Spencer and colleagues (2016) argued that there are three proposed mechanisms that can explain why stereotype threat might occur**.**

Mechanism 1: Underperformance due to extra pressure to succeed. The first proposed mechanism, “underperformance due to extra pressure to succeed” (Spencer et al., 2016, p. 420) refers to how pressure to do well may lead to underperformance at a specific task. Spencer and colleagues break this down even further into the following categories: “mere effort, working memory depletion, and conscious attention to automated processes” (p. 420).

Mere effort. To address the concept of mere effort, which the authors define as when “people experiencing stereotype threat are motivated to perform well in order to disconfirm the stereotype” (Spencer et al., 2016, p. 420), there are two important points to note. First, data has shown that stereotype threat is less likely to be observed when the learner is engaging in tasks which are considered easy or which the learner is fluent at. For instance, Ben-Zeev and colleagues (2005) showed that when asking participants to complete tasks that are considered "easier" to do than more difficult tasks under threat conditions, opposite patterns were observed: rather than threat hindering performance, participants completing easier tasks actually did better under threat conditions than their control group counterparts. The authors argue that their results are consistent with social facilitation theory, which posits that under conditions where individuals are asked to complete easier tasks, the physiological arousal one experiences during such tasks can lead to increased performance scores, while more difficult tasks might produce decreased performance scores, again due to arousal (Zajonc, 1965). They state: "when arousal is attributed to the threat, it may trigger strategies designed to suppress thoughts about the threatening stereotype, cause lowered expectations, lead to self-handicapping, reduce attentional focus, or engage [other…] mechanisms" (p. 179). Other research has also noted similar effects (O'Brien & Crandall, 2003b; Seibt & Förster, 2004). Similarly, the mere effort theory also posits that "individuals experiencing stereotype threat will take steps to compensate for their performance if they recognize that they have made an incorrect response and are provided the opportunity to correct it" (Pennington et al., 2019, pp. 718-719).

On the other hand, stereotype threat is more likely to be observed when tasks which are difficult or which the learner is not fluent are required (Jamison & Harkins, 2007). Much of the research on stereotype threat asks test takers to engage in tasks that are considered "difficult." In fact, two meta-analyses on stereotype threat research have found that a vast majority of research utilizes difficult or moderately difficult task requests (Nguyen & Ryan, 2008; Walton & Cohen, 2003).

Underperformance due to working memory depletion. This can be said to occur when a negative stereotype is present and relevant to the learner's performance and it "triggers a physiological stress response and a monitoring process to detect self-relevant information and signs of failure [and]… efforts to suppress negative thoughts and feelings that result from these two processes" (Spencer et al, 2016, p. 421). When these mechanisms are in place, much of the learner's working memory is used up, thereby making successful performance less likely. Those who spend a lot of time on things other than memorizing or studying the content, or in thinking about stress, anxiety, or other things that might hinder performance, the learner ends up using a good deal of valuable time and energy. Many researchers have argued for the working memory depletion account in explaining why stereotype threat occurs (Rydell et al., 2009; Schmader et al., 2009).

Conscious attention to automated processes. The final concept of how underperformance can be observed is in conscious attention to automated processes, of which Schmader and colleagues (2008) state "the monitoring process triggered by stereotype threat undermines such automatic behaviors by making individuals more conscious of their performance and more vigilant for signs of failure, leading to a controlled rather than automated form of regulation" (p. 421). From this perspective, even behaviors considered fluent could be hindered as a result of being aware of these behaviors according to Schmader and colleagues.

Mechanism 2: Threats to self-integrity and belonging. According to Spencer et al., (2016), the second mechanism, threats to self-integrity and belonging, reflects the behaviors that the learner might engage in that directly place them at risk of confirming the stereotype. For example, self-handicapping can occur by failing to practice or study (Stone, 2002), the learner might report stress or related factors that contribute to underperformance (Keller, 2002; Steele & Aronson, 1995), they might also attempt fewer test questions or skip questions (Davies et al., 2002; Steele & Aronson, 1995), or perhaps even lower their own expectations that they have for themselves (Cadinu et al., 2003). Another way in which researchers have conceptualized how people might attempt to either lessen or remove threats to self-integrity and belonging is through "disassociation" with characteristics of the stereotyped group. In a study looking at the conditions under which women were likely to perform well at math-related tasks, Pronin et al. (2004) argued that when math identified female participants were confronted with threatening gender stereotypes about math, the women engaged in disidentification by disavowing specific in-group features thought to increase one's risk of being negatively judged in a valued domain" (p. 165). The authors argue that this is one way people might cope with stereotype threat when in a threat situation. Hence, individuals might highlight or stress those characteristics which are more likely to be associated with success in the task at hand. The current research can examine if failing to adequately practice, stress/confidence, and number of questions answered within the time frame are factors that lead to changes in performance from group to group.

Mechanism 3: Priming of stereotypes. The third and final mechanism proposed is priming of stereotypes. According to research out of the ideomotor paradigm on stereotype-threat, "behavior can be a consequence of priming effects[…] when a stereotype becomes activated, stereotype consistent behavior may follow automatically from that activation" (Spencer et al., 2016, p. 422). Thus, just being aware of the stereotype can lead to the learner engaging in those same behaviors expectant of the stereotype. Picho and Schmader (2018), in a study on whether stereotype threat occurs on a math test with young adolescents in Uganda, argue that not only do stereotypes need to be primed in order to be effective, but also that those stereotypes must be known amongst the population being studied. They argue that not all people may be affected by similar stereotypes due to a potential lack of awareness of "board cultural stereotypes and [also] that stereotype awareness could vary based on one's exposure (or lack thereof) of these stereotypes at the micro-cultural level (i.e., peers and family)" (p. 303). This speaks to the importance of one's history with respect to stereotypes relevant to that person, their culture, learning history, and so forth. The current research may also be able to address whether or not this explanation for when stereotype threat is likely to occur through the data collected across groups and participants if differences in performance (e.g., time spent on questions and studying, answers) are found to be statistically significant.

The Role of Confidence in Stereotype Threat/Lift. As aforementioned, stereotype threat and lift are more likely to be observed under conditions where extra pressure to succeed, threats to self-integrity and belonging, and/or priming of the stereotype have occurred. Due to the subjective nature of the thought process that individuals experience as they are studying, answering questions, etc., adding in a tool to help elucidate this process can aide in our discussion of which (if any) of these factors might be influencing performance.

In a review of literature on subjective confidence, Koriat (2012) stated that "assessments of subjective confidence in one's own knowledge and judgments have been used and investigated in a wide range of domains" (p. 80). Such domains include social cognition, perception, eyewitness testimony, memory and metacognition, decision making and choice, as well as others (Dunlosky & Metcalfe, 2009). Although behavior analysts have generally shied away from the use of more subjective topics such as confidence, having an appropriate method of utilizing confidence research to help study stereotype threat can also enable us to speak to the legitimacy of the previous research in this area from a behavioral perspective.

To discuss how previous research has conceptualized confidence, Koriat (2012) noted that "the main cue for confidence is self-consistency, that is, the extent to which the choice reached is supported across the representations sampled from memory" (p. 102). In other words, to be self-consistent means that a person has answered similarly or the same in questions of a similar nature in the past as they are currently. In the current research, this can be analyzed by comparing each participant's answers during the tests to their reported level of confidence and see if they have answered consistently across each question and have given similar answers as to how confident they are in each answer.

Although confidence in answers throughout testing hasn't been directly measured (to our knowledge) in previous studies on stereotype threat, some researchers have utilized some components of "confidence" in their analyses. For instance, Schmader and colleagues (2009) had participants rate their level of anxiety between tests, and primed them with statements of confidence or doubt regarding their future performance. This highlighted what they considered "the importance of the cognitive interpretation of arousal as a powerful and reliable predictor of cognitive performance" (p. 594). Additionally, Hoyt and colleagues (2010) measured female participants self-appraisals in their leadership abilities before testing to see if those self-appraisals led to increases or decreases in performance on the subsequent tests. They found that boosts in performance were observed for participants when blatant threats were present, but that when combined with situations in which threat is present, previous self-appraisals and performance were undermined.

Koriat (2012) points to important factors that influence a persons' confidence, which rely on "specifying the relationships between choice, confidence, response latency, and accuracy in terms of within-person dynamics" (p. 83). How quick or delayed a person responds to each question, how accurate their responses are, and how they rate their confidence for each question can all help to provide validity to the previous research in confidence, as well as to provide us with a good look at moment-to-moment behavior changes that are occurring during testing. Assessing confidence on tasks under lift and threat conditions may help to further elucidate the relationship between threat/lift statements and under- or -over performance.

A Behavior Analytic Account of Stereotype Threat

The proposed mechanisms in previous research give us a comprehensive look at many of the ways in which we might conceptualize stereotype threat, as well as an opportunity to speak to how, when, and under what conditions stereotype threat/lift occurs or doesn't occur in our data from a behavioral perspective. However, a number of limitations with the current literature also exist. First, to our knowledge, the current literature has used only group design research, making group statistical comparisons easy but limiting our ability to better understand potential differences at the individual level. Per definition and the research base, stereotype threat exists at the individual level: it is how individuals behave with respect to different environmental contexts (e.g., presence of written statements on gender differences, presence of opposite gender experimenters, taking the test in the presence of mainly people of the opposite gender). While previous research has attempted to analyze different contextual factors, such as if individual participants “endorse” group stereotypes or not (e.g., Pennington et al., 2019) or how long each participant spent studying for the tests, data is consistently analyzed at the group level only.

Another limitation, which is also often a product of the group design, is that participants in the current research are only subject to one condition (control or no threat present, threat present, lift present, etc.). Thus, participants interactions with different types of statements are limited to the one condition in which they were assigned, reducing the ability to compare how each individual’s performance might change across various conditions (if at all).

As well, the current research tends to stem from a more cognitive lens. While this in itself is not necessarily a limitation, a behavioral account of stereotype threat could enhance and inform much of the discussion on stereotype threat, furthering our understanding of this line of research.

Given the reliance on language and the use of different statements or environmental contexts which can alter individual performances, a behavior analytic account of stereotype threat could enhance our understanding of the topic. As well, the ability to analyze individual differences within the group context will hopefully better provide some explanations of how stereotype operates at the individual level and what contexts might make stereotype threat more or less likely.

The Role of Language in Stereotypes

While the reasons as to how and why stereotype threat occurs (or whether it occurs at all) are frequently debated (e.g., arousal theory versus working memory depletion, arousal theory versus mere effort theory), these current theories and explanations to stereotype threat all suggest that language is an integral component of how stereotype threat works; that verbal behavior is the process by which stereotype threat is disseminated and affects attitudes, beliefs, and performances. Quinn and Spencer (Quinn & Spencer, 2001) state that "one of the most basic ways in which stereotypes […] are promulgated is through parents' and teachers' expectations" (p. 56). For instance, Jacobs and Eccles (1992) examined stereotypic beliefs that mothers held of both young boys and girls and found that these beliefs correlated directly with their children's actual mathematical, social, and sports abilities. In other words, mothers who held more stereotyped views of female and male capabilities (e.g., females are worse at math, but are better than males in the social domain) significantly influenced the likelihood that their children would essentially "live up to" those expectations. Similarly, research has shown that parents tend to overestimate their son's math abilities while underestimating how well their daughters will perform in math (Frome & Eccles, 1998). This subsequently affects the likelihood that their daughters will hold favorable opinions of math in general (Hyde et al., 1990), and hence the likelihood that those female children will enroll in more math classes (Eccles & Jacobs, 1986) and will pursue careers in areas related to mathematics.

What makes stereotyped thoughts and beliefs particularly insidious is that they can often be implicit or unknown to the person holding them. Parents might attempt to raise their male and female children similarly in effort to prevent their children from experiencing stereotypic patterns. But even well-intentioned parents can engage in behaviors that actually reinforce stereotypic patterns. For instance, Crowley et al. (2001) observed parents with their children in a naturalistic setting (i.e., museum) and found that parents spent more time explaining and discussing the science behind exhibits to their male children than their female children. Even such unintentional acts can influence the likelihood that male children will be more interested in and seek out careers in science.

The influence that language has on people encompasses not just one or two moments where a person might encounter stereotypic behaviors: these are present throughout the lifetime of individuals in different forms, frequencies, and levels of subtle or overt cues. Although accounting for a person's individual history with respect to stereotype threat is difficult, the use of Relational Frame Theory could help us to better account for how individual histories have been impacted by stereotypic statements, thoughts, and behaviors. Thus, Relational Frame Theory can help orient us better to these individual histories.

Relational Frame Theory (RFT)

Hayes and colleagues (2001) described relational frames as an elaboration on understanding language. Relational Frame Theory (RFT) is defined as “an explicitly psychological account of human language and cognition […that] approaches verbal events as activities” (Hayes et al., 2001, p. 22). RFT aims to provide an explanation as to how language enables humans to engage in processes such as comparing, contrasting, transforming stimulus function and making temporal, spatial, and deictic relations across the spectrum of human experiences. Hayes (2016) argues that “what brings these situations together is not their formal properties in a simple sense, but the verbal/cognitive activities that relate these events" (p. 875). According to RFT, language consists of verbal events or activities that demonstrate certain properties: mutual entailment, combinatorial entailment, and transformation of stimulus function.

Mutual entailment. To define, the terms mutual (two or more events) and entailment (to cause or involve) are chosen to describe the process whereby two events relate to one another. Hayes and colleagues (2001) state that "arbitrary stimulus relations are always mutual: If A is related to B, then B is related to A" (p. 29). For instance, if a person learns that something that is "big" is also "large", they will also derive that something that is described as "large" will also be "big." If one learns that a neighbor who is described as "friendly" also does things that are "nice", then that person would also derive that nice people can also be described as friendly. As such, associations between two specific words, concepts, or events can easily become mutually entailed when a relationship between the two is learned and derived.

Combinatorial entailment. While mutual entailment describes how relations between two events are trained and derived, combinatorial (of or relating to a combination of events) entailment describes how multiple relations can develop from simple A and B relations. "Combinatorial entailment refers to a derived stimulus relation in which two or more stimulus relations (trained or derived) mutually combine" (Hayes et al., 2001, p. 30). For instance, having previously trained that A is related to B, the individual hence derives how B is related to A. If we then train that A is related to C (a newly introduced concept), the individual has now derived that not only is B related to A, but also that A is related to C and that C is related to B (without directly training the B and C relations). To revisit our previous examples, if “big” and “large” had been previously trained and the individual now learns that a new concept, “tall”, is similar to “big”, that individual has now derived that not only is “big” related to “tall” and “tall” to “big”, but “large” is also related to “tall” as is “tall” to “large”. As well, learning that the word “sweet” describes someone who is “friendly”, we can derive that our “friendly” neighbor is also both “nice” and “sweet”.

Transformation of stimulus function. Last, the term transformation of stimulus function describes how “the change in functions of one event that stands in relation to another is not mechanical: it is in terms of the underlying relation” (Hayes et al., 2001, p. 32). To illustrate, if we trained that A is “better than” B, and that A is “worse than” C, and that participants earned more money when they selected the option that was “better than” the other, a person could derive (without any formal training) that C is the best and makes the most money, and B is the worst and makes the least money. Stimulus C would exhibit reinforcing properties, whereas B would exhibit aversive properties, or exhibit less reinforcing properties than C (or A). To revisit our previous example, a person might have a negative encounter in the store with a tall, heavy-set person. In the future, tall and/or large people might elicit feelings of anger and of anxiety, leading to avoidance of places where that person has encountered tall, big people. As another example of transformation of stimulus function in everyday experiences, a person who has eaten ice cream in the past is now told to picture an ice cream cone. The person might report enjoying ice cream, but they are also reminded of times when eating ice cream has led to stomach aches (perhaps the person is slightly lactose intolerant). The word “ice cream” can then take on aversive properties as transformation of stimulus function occurs.

Feelings of being embarrassed around others or physical stomach pains can occur again in the context of talk of or pictures/images of “ice cream”. However, that person then starts a new job at a place where their co-workers all get ice cream together after work on Thursdays. While they may have avoided ice cream previously, that person now encounters a context in which “ice cream” may take on different properties, such as positive qualities of getting to spend time with their co-workers, having fun, making jokes, etc. Ice cream, the presence of tall/big people, or things that remind someone of ice cream or big/tall people, then, can take on many other properties, creating incredibly complex networks of things, events, and our interaction with those things and events.

Relational Frame Theory Applied to Stereotype Threat

To depict how RFT might operate in a stereotyped context, let us return to the aforementioned gendered stereotype, “men are better at math”. Mutual entailment would indicate that “males are better at math” has been directly trained, while “Girls are worse than boys at math” would be derived. With combinatorial entailment, in addition to the already trained and derived “males are better at math” and “girls are worse than boys at math”, other combinations such as “Asians are good (or better) at math” could also be trained, leaving “Girls are worse than boys at math” and “Asians are better at math than Americans” to be some of the many derived relations that could form as a result of information added into these relational frames. Finally, transformation of stimulus function might occur when, as a result of the trained and derived information regarding gender/race and math performance, the mention of “math” or “math test” might elicit stress, anxiety, or other punishing properties. Additionally, the presence of things associated with “math” (e.g., calculators, times tables, excel spreadsheet) could also take on similar punishing properties. A more detailed look at how gendered stereotypes can occur from the lens of RFT enables us to better understand the reinforcing or punishing/aversive properties that can occur as a result of stereotyped beliefs/thoughts.

With respect to gender, researchers have used RFT to help discuss and explain why certain behaviors are more likely in different contexts. For instance, Errasti et al. (2019) examined how being in the context of a mixed gender or same gender environmental arrangement affected gendered responses for participants in an IRAP (Implicit Relational Assessment Procedure) application. Briefly, the IRAP is a procedure that has been developed specifically for the "assessment of beliefs, attitudes, and other implicit cognitive elements, intended to overcome the limitations of questionnaires, interviews, or other explicit measures when targeting socially sensitive attitudes" (p. 39). Participants are asked to identify whether the stimuli presented falls into one of two categories (e.g., good versus bad, truth versus lie, male versus female), and are only given a short period of time to respond. In the current study, participants were given initial trials in which they were asked to identify things that are consistent with gendered views and/or views of self versus the opposite gender (i.e., "me" versus "men" for female participants and "me" versus "women" for males). Then, participants are asked to provide answers that are inconsistent with those relations that were just trained. For example, if initially participants were asked to respond in line with gendered stereotypes such as women = weak, men = strong, then in the inconsistent procedure, the correct answers would be women = strong, men = weak. Data on "correct" (or consistent) answers and how fast the participant answered each question are calculated to indicate if a bias is present. The fewer correct or consistent answers provided and the longer the person takes to provide a response indicate the presence of a bias in line with (or not) the thoughts/beliefs that person has indicated at the start of the IRAP.

Errasti and colleagues (2019) found that the environment or social context in which the participants took the IRAP impacted their overall performance: across the board for both men and women, greater gender bias occurred in both the gendered stereotypes conditions and the inconsistent conditions when in the presence of single gender groups versus mixed gender groups. The authors argue that “coexistence in mixed groups can be an attenuating factor for gender stereotypes and gender bias” (p. 46). From an RFT perspective, the stereotypic statements in the presence of same group or mixed groups affected performance for participants, such that each participants previous history with both the stereotypic statements and their environmental conditions brought forth feelings/thoughts and behavior changes that impacted their performance. Stereotype threat research has also found similar results. Inzlicht and Ben-Zeev (2000) showed that females performed better on a difficult math test when taking the test in a group with other females, and that performance decreased when females took the test in a group males. The results of these studies provide evidence that, when gender and gender differences are made salient, this salience may affect the behaviors of individuals that are in line with gender stereotypes and contexts.

Relational frames are useful to evaluate the arbitrary and often novel nature of stereotyped contexts. Verbal stimuli that signal valence in association with other events, such as particular behaviors in idiosyncratic contexts, can immediately alter the functions of related stimuli and contexts through combinatorial entailment. For example, if a verbal statement describes a performance as “good,” then all of the demographic (and other) features of that context may also become varying degrees of “good,” and features that are different from that context might become “bad,” and so on. Stereotyped contexts can simply include stimuli that participate in a frame of coordination or distinction with the instructional stimuli to facilitate or inhibit stereotyped behaviors. That is, when presented with a stereotype such as “females are bad at math” when attempting a math assessment, these verbal stimuli evoke behaviors of skipping, guessing, etc. based on the arbitrary relation of those stimuli to the instructional material and the class of stereotypes as verbal stimuli.

It is important to conceptualize the effect of stereotyped statements in alignment with Relational Frame theory to both maintain integrity of the independent variable and predict the pattern of the independent variable.

Current Research

Currently, meta-analyses that have been published concerning stereotype threat and its effects have indeed found evidence for stereotype threat (Spencer et al., 2016). This aside, observation and interpretation of the phenomenon from a behavior analytic and RFT perspective is not only wanting, but can offer a more precise and moment-to-moment look at what is occurring during stereotyped contexts, or situations in which stereotypes are likely to influence the behaviors during testing. Behaviors that occur in the presence of stereotyped contexts (verbal or otherwise) might include guessing, skipping, spending little time on problems and moving on without checking work. The stimulus features of these contexts are likely varied, often novel, and potentially undetectably subtle to an observer. The use of relational frames is useful to encompass the arbitrary and often novel nature of stereotyped contexts. Stereotyped contexts include stimuli that participate in a frame of coordination or distinction with the instructional stimuli to facilitate or inhibit these behaviors. That is, when presented with a stereotype such as “females are worse at math than males” when attempting a math assessment, these verbal stimuli can evoke behaviors of skipping, guessing, second guessing and hence changing previous answers, spending longer/shorter amounts of time on questions, etc. based on the arbitrary relation of those stimuli to the instructional material and the class of stereotypes as verbal stimuli.

Although research on stereotype threat/lift has been plentiful in the field of social psychology, there are a number of benefits that the research can gain by approaching this topic from a behavior analytic standpoint. First, given the nature of social psychological research, research has only examined this phenomenon in group research rather than single-case designed research. As such, to our knowledge, researchers have also never examined contexts in which the same individuals are confronted with both threat and lift contexts (as most are assigned to a threat, neutral, or lift group/condition). Hence, the effect that these statements have on individuals across tests in terms of confidence in answers and performance have not previously been discussed. Behavior analysis can thus help us to better examine “moment-to-moment” changes within sessions or tests (e.g., question accuracy, time spent on questions, reported confidence within and across tests) that occur during stereotyped contexts. Additionally, research has included questions around how long individuals have spent studying for tests in the current literature, but these questions have been primarily self-report based and usually constitute ranges for total hours spent reviewing material for exams/tests (e.g., 1-3 hours, 6-9 hours), although some studies have given participants specified amounts of “study” or review word sets for later testing (Hess, Emery, et al., 2009; Hess, Hinson, et al., 2009). The current research will be able to include measures of exactly how long one has spent studying the current material for the tests and will enable comparisons of actual study behavior in contexts where stereotype threat or lift are present.

Research Questions

The current research aims to arrange conditions under which stereotype threat and stereotype lift are likely to emerge through a computer-based assessment (an arbitrarily constructed stimulus test, or “memory test”) to observe whether stereotyped statements have a systematic and replicable effect on performance. From previous research, we expect to see performances in line with threat (i.e., suppressed performance) and lift (i.e., improved performance). Thus, this research aims to answer the following questions: 1) What are the effects of threat and lift statements on performance? 2) Does group affiliation (i.e., gender) affect performance on Memory Tests across groups (male vs. female) and individuals? And 3) if performance differences are observed, are other behaviors similarly affected (for instance, are latency to respond, accuracy of answers, latency to complete each task, and confidence in performance affected or not by stereotyped contexts)?

We conducted a series of three pilot studies to establish the general method and procedures prior to the dissertation research. Each will briefly be examined, followed by a more in depth look at the current study.

Method

Pilot Studies

Apparatus and Testing Preparations

Visual Mental Rotation Test (VMRT) (Vandenberg & Kuse, 1978). The VMRT is an assessment, consisting of 24 questions, wherein the participants are asked to compare a sample stimulus of a third-dimensional (3D) object drawn on a two-dimensional apparatus (i.e., an 8 X 11 inch piece of white paper) to 4 other 3D stimuli, and told that only 2 of the stimuli are the same figure as the sample stimulus, but are rotated to be in different positions. The pilot studies utilized a computerized MRT rather than the paper/pencil method by digitally drawing figures similar the original MRT and utilizing those drawn figures as images for the computerized version of the MRT test (see Appendix B for images of the computerized MRT). To score the MRT, per Vandenburg and Kuse (1978), participants receive scores between 0 - 24, only getting points if the indicate which 2 of 4 answers were correct (in other words, no points are given when participants correctly identify 1 of 2 correct answers). Historically, studies have concluded that men tend to outperform women in tests of mental rotation (Guillot et al., 2007; Jones & Healy, 2006; Vandenberg & Kuse, 1978), even when the total number of questions correct is taken into account or partial credit is awarded for partially correct answers (McGlone & Aronson, 2006). In the current research, we report on both scores. For sake of clarity, we have termed the original scoring method the “dichotomous” score, wherein points are given only when participants correctly identified both answers. Dichotomous scores are calculated by taking the total number of questions where both answers were identified correctly and dividing by 24 (as there are 24 question total). The total “weighted” score (as we will call them hereafter) is calculated by taking the total number of correctly identified answers and dividing by 48 (as there are 48 total answers a participant can provide). The weighted scores are more sensitive to overall accuracy, while the dichotomous scores represent a more sensitive depiction of accuracy. Previous research has utilized both scoring methods of MRTs (McGlone & Aronson, 2006), showing similar results across both scoring methods.

Memory Test (MT). For the purposes of the current research, a modified stimulus equivalence task (Sidman & Tailby, 1982) via computer program was created in order to examine the relation between stereotypic suggestions on performance. Four tests were originally created for pilot 1 (with an extra MT created for pilots 2 and 3). Each MT had 24 questions to mimic the format and scoring of the MRT. Participants were presented with 4 training sets per test (with 12 total training sets overall). Each training set consists of 5 different items paired together (a symbol, color, number, 3-letter word, and letter). Each training set has a different symbol, color, number, 3-letter word, and letter, so as to eliminate the possibility of previously trained sets interfering with the training/testing for any other sets. Similar to the MRT, each question consists of a sample stimulus compared to 4 possible answer choices, of which 2 are correct. The computer program calculated how long each participant spent in the training phase (i.e., studying the training sets), as well as how long they spent on each question and how long they spent on the test overall. All information regarding participants answers to the questions (i.e., answers, answer positions, how often they revisited questions, questions correct and incorrect) and their rated confidence level for each question was also automatically recorded.

During the program, participants are first given instructions on how to navigate the training and testing phases (for a detailed script of the instructions, see Appendix C). After instructions were given, the training phase began. Participants were given as much time as they needed to review the 4 training sets relevant to that test. They could browse freely between all 4 sets until they selected that they were finished and ready to move on to the testing phase.

Previous research on stereotype threat and recall, or remembering previously taught information, has primarily been conducted with the aging population and their ability to recall information previously told to them under non-threat and threat conditions (Hess, Emery, et al., 2009; Hess, Hinson, et al., 2009a; Horton et al., 2010; Levy, 1996; Logel et al., 2009; Rahhal et al., 2001; Sekaquaptewa & Thompson, 2003). Such memory tests have generally consisted of showing lists of words to participants and asking them to recall those words at a later point in time. Recall studies on stereotype threat have also looked at how well people remember information told to them by other group members under threat conditions (Sekaquaptewa & Thompson, 2002), recall of information when the task is framed as a memory game versus a geometry test (Huguet & Régner, 2009), and participants' ability to recall new information in high pressure situations of threat (Taylor & Walton, 2011). While the current research will utilize a similar set up for recall (i.e., review new information now, answer questions on that information after), the use of a stimulus equivalence task hasn't been utilized, to our knowledge, in any previous studies.

Examus. All testing for the pilot studies were proctored through *Examus*, a test proctoring service, in order to ensure participants were not cheating by writing or typing training sets down. Due to restrictions put in place due to COVID-19 during pilot data collection, this software enabled participants to participate fully online without the need for a proctor to be physically present while they were taking their tests. Participants first consented to be recorded while training/testing occurred. All participant videos were stored on *Examus*'s secure online system, and each video was analyzed by *Examus*' software, which creates a report regarding behaviors that could be considered "cheating". Only the researchers and *Examus* administrative personnel had access to these videos through *Examu*s. All videos were moved to a secure online box folder for the duration of and post-research.

Some issues arose with the Examus links for a few participants. Some participants did not follow instructions and were not using the recommended updated browser (i.e., Chrome) and attempted to participate on devices other than computers (i.e., tablets and phones). Participants who contacted the researcher and were given new links with instructions on how to remedy the issue. For other participants, Examus did not set up properly on their computers even after fixing other potential issues and attempting 2-3 different links. When this happened, participants were given two options. The first was to use Zoom to record themselves completing testing and were given the direct link to the testing site (enabling them to complete testing and get the same amount of credit) or they were also given the option to cancel with no negative ramifications in SONA. For those participant that chose the zoom option, their Zoom videos were uploaded by the participants to the secure box folder, shared only with the researchers.

Dependent Variables and Measurement

As the independent variable was the presence (or absence) of a gender stereotyped statement (i.e., “men perform better than women”, “women perform better than men”, or “all test-takers perform similarly”), the dependent variables measured were: 1) overall performance/accuracy per test, 2) performance/accuracy on each question per test, reported in both weighted scores (total number of correct responses out of 48 possible answers) and dichotomous scores (total number of questions with both answers correct out of 24), 3) total time spent studying or reviewing the memory sets for each MT, 4) total time spent taking each test, 5) time spent on each question per test, 6) reported confidence level for each question, 7) reported overall level of confidence per test (see description of how this was measured below), 8) total number of times participants visited each question, 9) total number of visits per test, 10) observed within test behaviors, and 11) post-test responses after all tests were finished.

For the purposes of this study, confidence was individually rated by participants on a 5-point Likert scale (i.e., “unconfident” “fairly unconfident”, “neither”, “fairly confident”, and “confident”, respectively). A percentage was assigned to each rating (0% to unconfident, 100% to confident). The program automatically assigned a numerical value to each level (1= unconfident, 5 = confident, etc.). As a result, confidence (or percent confident) was calculated by taking the sum of confidence ratings from all 24 questions per test, subtracting 1 (to account for 1 = unconfident or 0%), and dividing by 4. This gave an overall percentage of confidence for each MRT and MT.

Pilot 1. Participants signed up via UNR's SONA System and were directed to complete the online consent form (via Qualtrics). Upon signing up and completing the consent form, each participant was emailed instructions on how to participate, including their own link to the Examus service and testing site, as well as an individual login/password to participate in the tests. Participants had until the end of the testing window they signed up for in SONA to finish the tests, and could do it at a date, time, and location of their choosing between when they signed up and the testing window closed. All login and password information are specific to the participant, and are non-identifiable. Reminders to complete testing or cancel participation were also sent throughout the testing window to participants who had not yet completed their tests or consent forms.

For pilot 1, 40 students in total completed the study. Appendix D contains demographic information provided in their consent forms and post-session survey on all pilot 1, 2 and 3 participants who completed the study. Of the 40 participants who completed the study, 57.5% identified both their gender and sex as female and 42.5% identified (both gender and sex) as male.

For pilot 1, participants were randomly assigned to 1 of 3 conditions (depending on their identified gender): the control condition (Group X), Group Y, or Group Z. As this study aimed to look at the impact of stereotyped statements on performance, it was necessary to first divide participants by gender and then semi-randomly assign conditions. This was done using the random list order generator from random.org to pre-determine which condition each participant would be assigned to (it should be noted that, as data continued to be collected and disparities between the number of participants in certain conditions emerged due to study cancellations and no-shows, participants were occasionally assigned specific groups by the first author in an attempt to even out experimental and control group participants). All control and experimental conditions consisted of having the participants complete a Mental Rotation Test (MRT), then 4 separate memory tests in the same order (i.e., MRT 17/18, MT 91/92, MT 105/106, MT 221/222, and MT 357/358). All MRT and MTs across conditions included the same instructions, training, and questions. The only difference between each condition was the presentation of a made-up stereotypic statement regarding how well males or females as a group had performed on each test in the past or no statement regarding gendered performance. Those statements were: “All test takers perform similarly (statistically significant at p < .05)”, “Women tend to perform better than men (statistically significant at p < .05)”, and “Men tend to perform better than women (statistically significant at p < .05)”. In the control condition, no stereotyped statements were presented throughout the entirety of the procedures/tests (MRT + MTs). Table 1 depicts the tests given to each group per condition.

Table 1

Pilot 1 tests, order of tests, and group conditions

Tests	Group X		Group Y		Group Z	
	Female N = 8	Male N = 6	Female N = 7	Male N = 6	Female N = 7	Male N = 5
MRT 17 & 18	Neutral	Neutral	Neutral	Neutral	Threat	Lift
MT 91 & 92	Neutral	Neutral	Threat	Lift	Neutral	Neutral
MT 105 & 106	Neutral	Neutral	Neutral	Neutral	Lift	Threat
MT 221 & 222	Neutral	Neutral	Lift	Threat	Neutral	Neutral

MT 357 & 358	Neutral	Neutral	Neutral	Neutral	Threat	Lift

Note. Neutral = statement that all test takers perform similarly; Threat = statement does not favor the group the individual identifies with; Lift = statement favors the group the individual identifies with.

In group Y, no stereotyped statement appeared before the MRT test, but did appear in memory tests 1 (MT 92) and 3 (MT 222). In group Z, a stereotyped statement appeared for the MRT (18), as well as memory tests 2 (106) and 4 (358). As the stereotyped statements specified for each test were the same for all participants in groups Y and Z in pilot 1, it was necessary to break both conditions down by male and female participants given how each statement could impact these two groups differently (i.e., what serves as a stereotype threat against females is at the same time a stereotype lift for males and vice-versa).

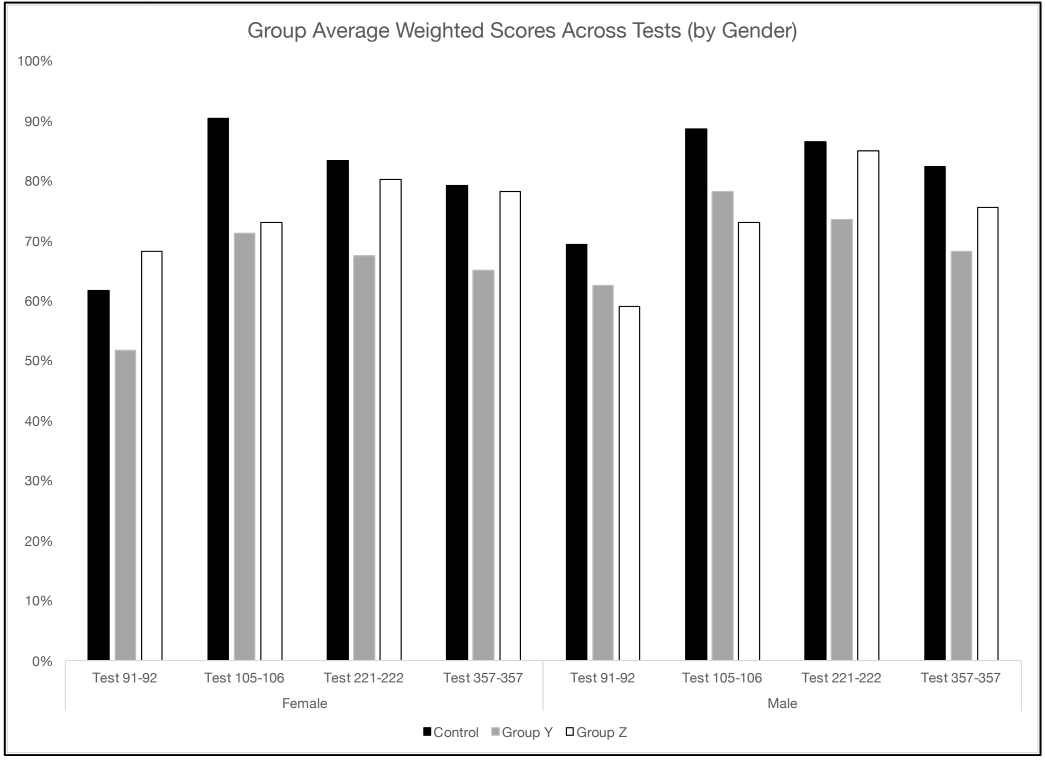
Participants were given detailed instructions on how to complete the MRT and MTs before the MRT and first MT (thereafter, participants had the option to “show instructions” in the program if they wanted to at any point during testing) and had an unlimited amount of time to “study” or review the sample stimuli prior to taking each MT. No studying was required for the MRT test. For each test and question, participants needed to identify 2 of 4 answers that they thought were correct for the question being asked, and also how confident they were in each answer by ranking their confidence on a 5-point Likert scale, ranging from “unsure” to “confident.” Participants were not able to move on to the next question if they did not provide 2 answers and a confidence level for each question. Participants could go back to previous questions and skip ahead, as long as they have previously provided answers to those questions. Once participants completed each test and hit “finish”, they were no longer be able to access the previous test. The program directed them directly to the next test.

Once participants completed all 5 tests, they were automatically directed to a Qualtrics survey to fill out a brief exit survey (see Appendix A for all exit survey questions) to report how well they thought they did on each test, what their study techniques were, how they thought the stereotyped statements impacted their thoughts/performance, and any other feedback they had regarding the tests. Participants were not debriefed at this time due to the likelihood that that information could impact potential future participants; however, participants were able to opt in to be informed of the main aim and overall findings of the research via email once the research concluded. A repeated measures mixed group design was utilized to examine the effect of threat and lift statements across groups. Additionally, a within-subject design (i.e., single-subject) was also utilized to more closely examine individual differences within groups. Condition arrangements for pilot 1 designs were 1 of the following for each participant: A-B-A-C-A or B-A-C-A-B for females and A-C-A-B-A or C-A-B-A-C for males, wherein A = neutral or no threat/lift present, B = threat statements present, and C = lift statement present.

Figure 1.1 depicts the weighted group averages for males and females across tests.

Figure 1.1

Pilot 1 weighted score averages across tests (by gender)



Note. Total N = 38 (Control N = 14, Group Y N = 11, Group Z N = 11).

There were two findings from pilot 1 that are worth noting, as they informed the two other pilot studies and dissertation research. The first finding was that across all groups, many participants had their lowest scores in the first MT (MT 91/92). Six participants in each group had their lowest score in this condition (43% of participants in Group X, 54% in group Y, and 54% in Group Z). This pattern was likely indicative of a potential testing effect, wherein this was the first MT they were taking and they had perhaps not prepared as well as they might need to in the future.

The second finding of note was in comparing the lowest scores for individual participants. For both the control group (X) and group Y, Where MT #2 (105/106) was neutral for both groups, the number of people and proportion of participants who had their lowest score in MT #2 was 1 (7%) for group X and 0 (0%) for Group Y. However, for Group Z, where MT #2 was a lift for females and a threat for males, 5 of 11 participants had their lowest score in MT #2 (2 males (threat), 3 females (lift)). Thus, 46% of participants overall had their lowest score in this condition. This suggests that the presence of a stereotyped statement may have had an effect on performance for Group Z, even if participants noted that effect or not in their post-tests.

Pilot 2. For pilot 2, 39 students in total completed the study. Of the 39 participants who completed the study, 61.5% identified both their gender and sex as female, 28.2% identified (both gender and sex) as male, and 4 participants (10.3%) identified their gender as genderqueer and their sex as female.

Participants were also assigned to 1 of 3 groups as in pilot 1. For pilot 2, all participants completed a neutral MRT (17), followed by 2 randomized neutral MTs, then 3 subsequent randomized MTs that introduced threat and lift statements. All dependent variables remained the same from pilot 1, and all data was calculated in the same way as in pilot 1. Additionally, the same group design was again used in pilot 2. However, an A-A-A-B-C-B or A-A-A-C-B-C design was used for pilot 2. Table 2 depicts the Pilot 2 test order and group conditions.

Table 2

Pilot 2 order of tests and group conditions

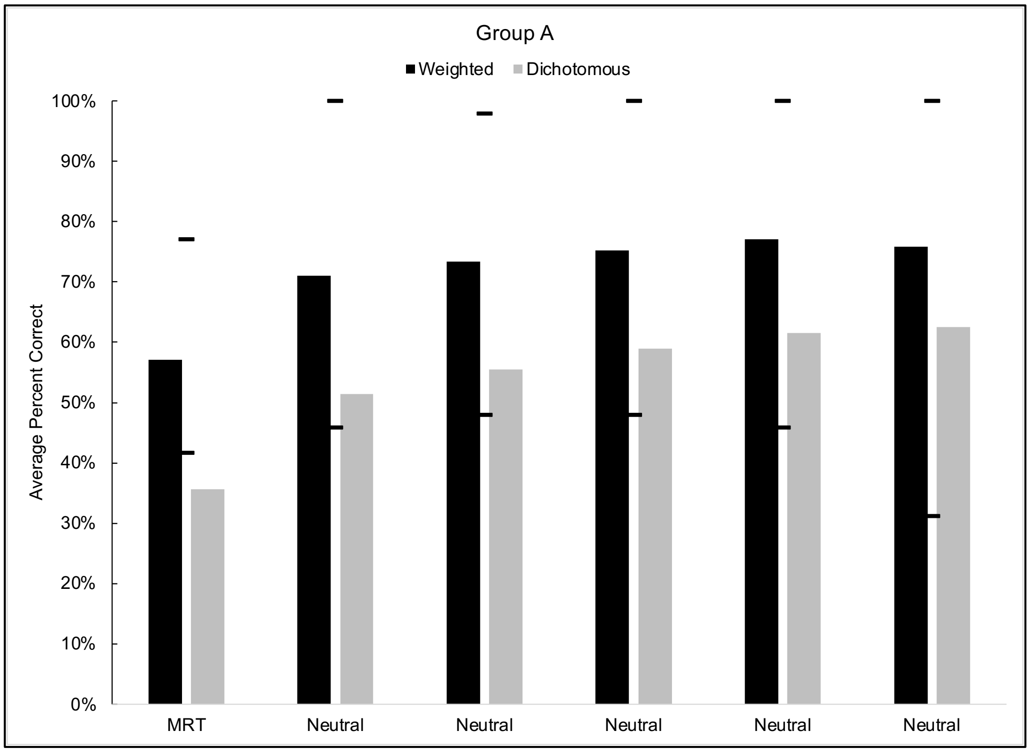
Tests	Group A		Group B		Group C	
	Female	Male	Female	Male	Female	Male
MRT 17	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
MT 1	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
MT 2	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
MT 3	Neutral	Neutral	Threat	Threat	Lift	Lift
MT 4	Neutral	Neutral	Lift	Lift	Threat	Threat
MT 5	Neutral	Neutral	Threat	Threat	Lift	Lift

Note. Neutral = statement that all test takers perform similarly, Threat = statement does not favor the group the individual identifies with, Lift = statement favors the group the individual identifies with.

Overall group data for groups A-C are presented below (Figures 2.1-2.3). Group A data (Figure 2.1) show an overall increasing trend in weighted and dichotomous scores across tests, with a slight decrease from the fourth MT to the fifth and final MT.

Figure 2.1

Control group average weighted and dichotomous scores with weighted score ranges

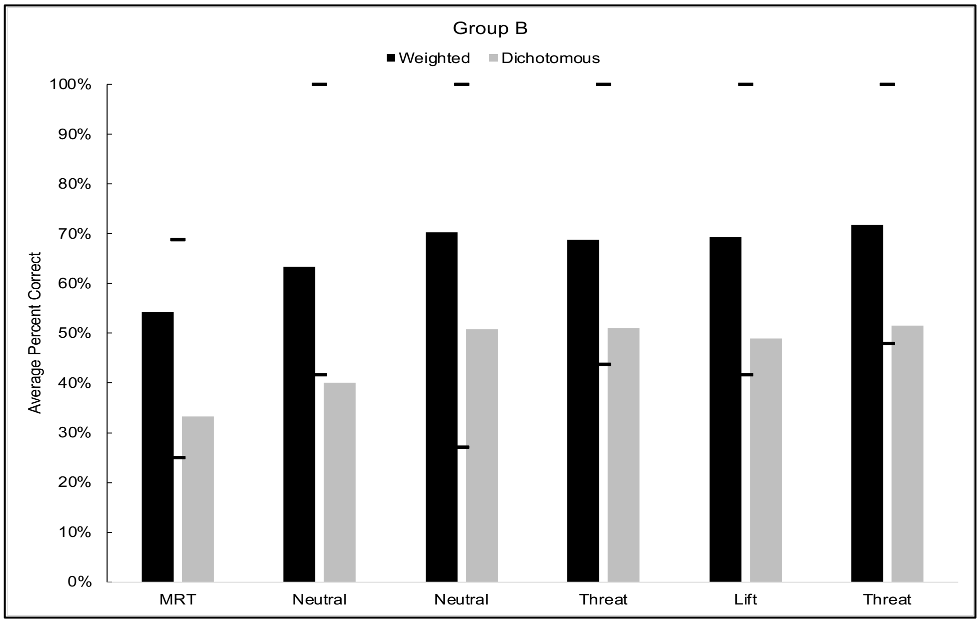


Note. Total N = 12 (Male N = 3; Female N = 9). Weighted average percentage is calculated as any correct responses out of 48, dichotomous refers to whether each question received the two correct answers and is scored out of 24. Lines next to the weighted bars indicate the range of weighted scores.

Group B data, depicted in Figure 2.2, show an increasing trend from the MRT, MT # 1 and MT #2, then show a slight decrease in MT #3, which introduces the first threat statement. MT's # 4 (lift statement) and # 5 (second threat statement) again show an increasing trend after the decrease in MT # 3. With the exception of a higher overall average for Group B participants than the control group in the MRT, all memory test score averages were lower than the control group, with the largest score different between groups in MT #4 (10.4% lower average score for Group B than the control group).

Figure 2.2

B group average weighted and dichotomous scores with weighted score ranges

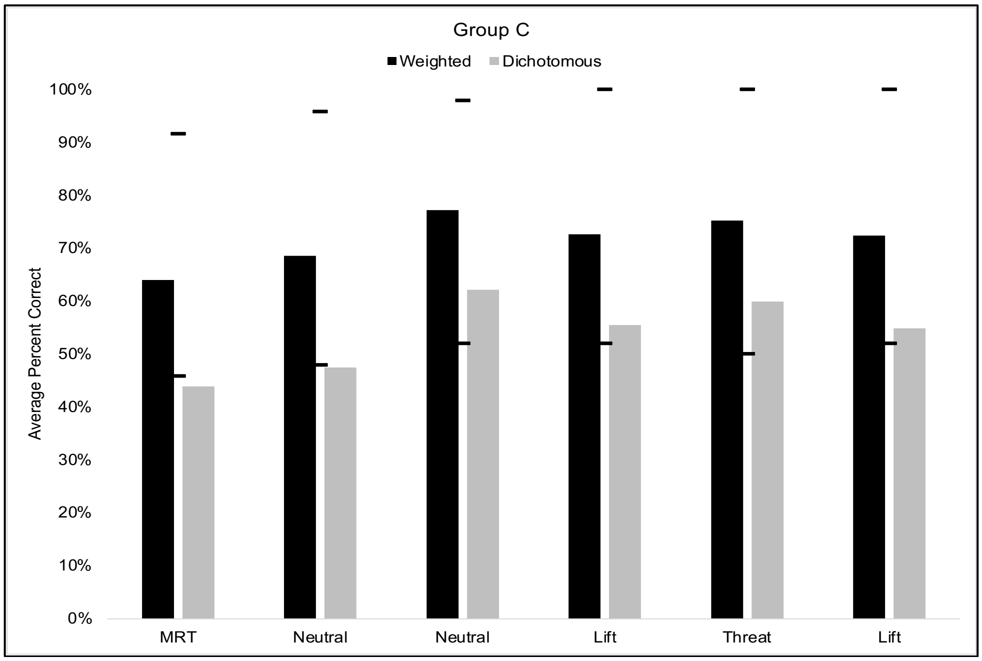


Note. Total N = 15 (Male N = 5; Female N = 10). Lines next to the weighted bars indicate the range of scores.

Group C average weighted and dichotomous scores differed more from the control group than did Group B (Figure 2.3). Although we again saw an increasing trend from the MRT through MT's #1 and #2, the average for MT #3 (lift statement) decreased slightly (similar to Group B). While the average for MT #4 (threat statement) again increased for Group C, the overall average decreased in MT #5 (the second lift statement). Thus, overall group data show lower overall scores when lift statements were presented, and higher overall scores when the threat statement was present.

Figure 2.3

C group average weighted and dichotomous scores with weighted score ranges



Note. Total N = 13 (Male N = 3; Female N = 10). Lines next to the weighted bars indicate the range of scores.

Some important patterns emerged from pilot 2. While there were no significant differences (after running a repeated measures ANCOVA, with gender as the covariate) between performances amongst groups A, B, and C (as was similar in pilot 1), participants individual data showed different patterns across tests depending upon statements presented. As well, while participants generally did better from MT #1 and MT #2 to MT #3 in groups A and B, more participants in group C (where MT #3 was a threat for all participants in group C) saw a decrease in test scores from MT #2 to #3 in group C (53.3%), while only 41,7% in group A and 38.5% in group B (where the statement was a lift) saw a decrease. Although some participants went on to perform worse in the subsequent lift condition for group C, the initial reaction to the threat statement was more pronounced than for those that saw a neutral statement, and was even more pronounced than those who saw a lift statement. Very similar to patterns observed in pilot 1, more participants in Group B saw their highest scores or best performances in MT #2 (neutral) before MTs 3-5 (T-L-T). Thus, although participants may have done better in the subsequent threat or lift conditions compared to the other threat and lift conditions, performance scores never surpassed those in MT 2.

Pilot 3. 8 students in total completed pilot 3. Of the 8 participants who completed the study, 75% identified both their gender and sex as female and 25% identified (both gender and sex) as male.

The design remained the same as in pilot 2, but with one difference occurring in MT #4 for all experimental group participants (i.e., Groups F and G). All participants were given a return to baseline condition in MT #4 (i.e., A-A-A-B-A-B or A-A-A-C-A-C). Group D (control) remained the same as in pilot 2. Table 3 depicts the pilot 3 test groups and types of statements that were presented.

Table 3

Pilot 3 order of tests and group conditions

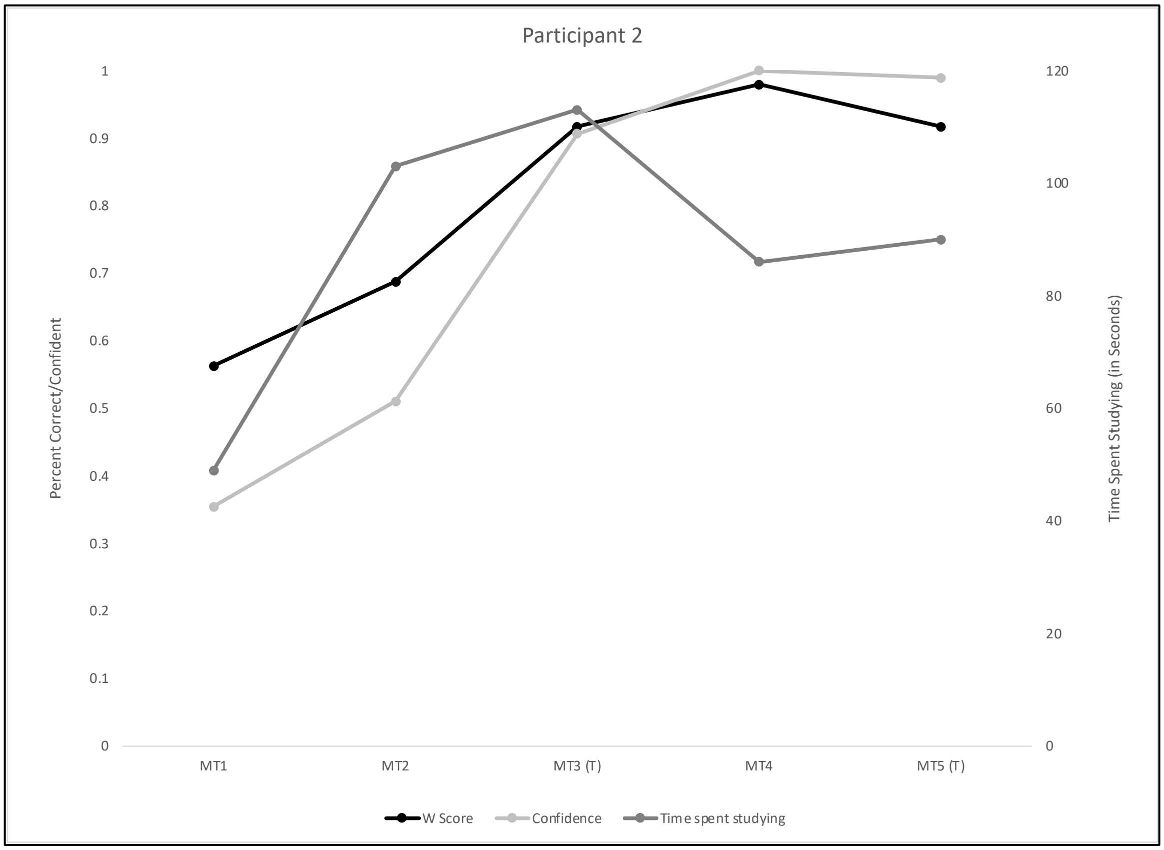
Tests	Group D		Group F		Group G	
	Female	Male	Female	Male	Female	Male
MRT 17	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
MT 1	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
MT 2	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
MT 3	Neutral	Neutral	Threat	Threat	Lift	Lift
MT 4	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
MT 5	Neutral	Neutral	Threat	Threat	Lift	Lift

Note. Neutral = statement that all test takers perform similarly, Threat = statement does not favor the group the individual identifies with, Lift = statement favors the group the individual identifies with.

Data collected from 8 participants show similar patterns as seen in Pilot 2. Only 7 participants' data were analyzed, as one participant rushed through the MTs (spending 2-8 seconds studying for each). Of the remaining 7 participants, 5 of 7 (71.4%) corresponded with similar patterns as seen in pilot 2 (i.e., threats higher, neutrals higher, or lifts higher). The remaining 2 displayed an increasing trend pattern, but interestingly, both had higher scores on MT 2 than on MT 3. Below is one example of individual participant data from pilot 3.

Figure 3.1

Participant 2: Weighted scores, confidence, and time spent studying



Note. Participant was part of group F – female. Data falls into the "neutrals higher" category, as the participant scored better on the neutral condition in between the two threat MTs. Participant 2 stated that the stereotyped statements impacted their performance accordingly: "Made me want to do better because I am a woman and didn't want to 'lose' to a man".

Although pilot 3 remained similar to pilot 2, the return to baseline condition (neutral statement in between two threats or lifts) helped to isolate the effect of threat or lift statements on performance scores, confidence, time spent testing and studying, etc. Thus, clear patterns observed in two neutral conditions (2 and 4) could more easily be compared with the two threat or lift conditions.

To briefly summarize, data and patterns observed from pilots 1-3 indicated performance differences in threat conditions compared to their neutral and lift counterparts. Although performance differences were not always clear to observe across groups, group Z in pilot 1, group C in pilot 2, and group F in pilot 3, all saw worse performances when the first threat statement was introduced after one or two neutral MTs. Although pilot 2 isolated patterns better than in pilot 1, and pilot 3 eliminated a third statement for participants with a return to baseline in MT 4, the switching of statements between individual tests could have created different responses across time.

As well, the shorter gendered statements could have been easily glossed over in pilots 1-3, with little justification provided as to why one gender might have outperformed the other or not. Thus, scripts linking participants' performances to a "validated" memory test, complete with information as to why performance on memory tests is important, were added and provided to participants in the current study.

Current Study

Participants and Setting

As in pilots 1-3, all participants were students at UNR who signed up through the SONA system. 60 students started testing, with 57 completing the study. To ensure equal numbers of participants across groups and genders, we aimed to collect data on 10 participants per group (50 total) with as evenly distributed as possible males and females in each group. Appendix E contains demographic information provided in participant consent forms and post-session survey on all current study participants who completed the study. Of the 57 participants who completed the study, 49.1% identified both their gender and sex as female and 50.9% identified (both gender and sex) as male. The average age of participants was 20.3, ranging from 18-29, with both a median of 20 and mode of 19. Most participants again identified their race as White/Caucasian (64.9%), and 15.8% identifying as “other”, 8.8% identifying as Black or African American, and 7% identifying as Filipino. Other racial categories participants identified were: Chinese, Japanese, Asian Indian, Korean, and Native Hawaiian. For ethnicity, 68.4% of participant identified as Non-Latinx/Non-Hispanic, while 17.5% identified as Mexican or Chicano/a and 12.3% identified as “other”, which included Salvadorian and Guatemalan. Participants identified their year in school as the following: 35.1% Freshmen, 17.5% as Sophomores, 19.3% as Juniors, 24.6% as Seniors, and 1.8% as Graduate Student.

Participants were also asked to indicate their current majors and minors. The most common majors among participants were Psychology (24.6%), Neuroscience (19.3%), and Kinesiology (10.5%). Participants also reported the following majors: Agricultural Science, Business, Management, Marketing, Civil Engineering, Microbiology, Criminal Justice, Public Health, Social Work and “Undeclared”. The most common minors provided were: Undeclared (66.7%), HDFS (5.3%), Substance Abuse (3.5%), Immunology (3.5%) and CAS (5%).

We analyzed data for only 49 participants in total for the current research because three of the 57 met the criteria for “rushing” per the previously established criteria, three participants experienced program errors (two hit back buttons that restarted the tests they were taking without restarting the timer, and one was accidentally assigned the same test twice). One participant was removed because they fell asleep during testing, and the remaining participant was removed because they were an extra participant in their group and the last one to participate.

Our exclusion criteria remained the same as in pilots 1-3 (must be 18 years of age). In the pilots, some participants spent so little time reviewing the study set that their scores were generally below the average and unrepresentative of the performance on the AMAT that may or may not have been influenced by a threat or lift statement. To address the issue of participants who rushed through training and testing, thereby providing data that isn't representative of studying and retention in the AMAT's, the analysis excludes data from participants who spent less than 40 seconds studying across all 6 MTs, and another participant was recruited in their place.

Participants accessed the study in one of two ways. The first was through Examus (as participants in pilots 1-3 had done). We added an extra check for systems compatibility in the consent form for participants to determine if their computer would meet compatibility standards before attempting the testing. However, some participants still had issues with compatibility even after passing initial checks. The second option, added to increase access to testing, enabled participants to participate via Zoom. We created a separate SONA study for "in lab" participation, wherein participants signed up for live timeslots to meet with a proctor over Zoom. The proctor guided these participants to complete their consent forms, gave brief instructions (as were provided in the email Examus participants received), shared the link to the site and their login/password info, asked the participants share their screen (desktop), leave their webcam and microphones on, and the proctor then recorded the participant via zoom taking the tests. The proctor kept their camera and microphone off the entire time the participant was taking their exams. A total of 22 participants completed testing via Examus) and 28 participated via Zoom. With the exception of group I, all groups had both Zoom and Examus participants.

While we arranged for participation via Examus and Zoom to be as similar as possible, there were some minor differences in the two options for participation. Specifically, participants via Zoom actually interacted with a proctor (all of whom were female) for the testing description and at the end when the proctor ended the recording. Via Examus, participants only interacted with another person via the main participation email sent from the researcher, or if they messaged the researcher with questions. Although we anticipated the Zoom option might be more likely to result in participants paying attention and not rushing through testing, “rushing” (as per our specified criteria) still occurred across both groups. Also, the participant who fell asleep did so during a Zoom session. While participation was slightly different across the two participation options, similar testing behaviors were observed across participants in both SONA options.

Apparatus and Testing Preparations

The same Memory Tests (MTs) used in the three pilot studies were again used in the current research as were. However, the current arrangement included a couple of changes that differed from pilots 1-3. First, the MRT was not administered for any participants. This allowed us to isolate our analyses to memory tests alone and observe if similar patterns to the pilots occurred in testing. Second, an extra MT was added, making the total number of MTs that all participants took six regardless of assigned group. Thus, all participants received 6 memory tests (again, in random order).

Third, to increase the legitimacy and similarity to standard operating procedures of the MTs, we changed the language and scripts for the memory tests and renamed the MTs the Atlanta Memory Adaptation Tests (AMATs). As a note, there is no such test as the AMAT, this statement was intentionally deceptive in an attempt to establish credibility of the test and the stereotyped statements that were given prior to each. In addition, we included a longer script prior to MT 1 (the first set of three tests all in the same condition) and MT 4 (the first set of three sets in the next condition), and the same statements that occurred at the top of the screen in the previous MTs now scrolled across the screen during testing (whereas these statements were static for the pilot studies). Rather than including a short note on any changes in gendered statements both prior to each test and at the top of the screen (as was previously the case for pilots 1-3), we separated gendered statements into what we labeled “scripts” and “statements” in the computer program. Scripts described, in detail, the purpose of the AMATs, how the AMATs predicted success, and whether or not the versions of the AMAT they were taking were gendered (modeled after Naphan-Kingery & Kemmelmeier, 2018), and appeared prior to AMAT 1 and 4. The shorter statement simply re-stated what script the participant had seen in each AMAT and scrolled across the top of the screen for all AMATs. To clarify, participants only received scripts twice (before the first AMAT and AMAT 4), whereas the statements appeared on the top of every AMAT, such that the statements reinforced what was said in the previous script. Appendix F shows the scripts and statements associated with each version of each AMAT.

Dependent Variables and Measurement

We aimed to address the same research questions and dependent variables as discussed previously in pilots 1-3.

Design

From pilot studies 1-2, mixed statements presented to each participant (i.e., threat, lift, neutral) produced issues with being able to analyze impact of statements across participants. In pilot 3, we attempted to isolate this issue by limiting each participants exposure to only two conditions (neutral + lift or neutral + threat). For the current research, the program only switched the gendered statements once to better analyze the impact of the statements, instead of switching back and forth as in the previous 3 pilots. A mixed model factorial design (within and between groups) was used for the current research. Participants were assigned to one of five groups. Table 4 depicts the order and group conditions.

Procedures

All procedures remained the same as pilots 1-3, with participants all signing up via SONA. As aforementioned, participants had the option to sign up to participate in the in-lab version of the study or the online version (Examus). For in-lab (Zoom) participants, a brief email reminder and consent form link was sent to all participants. Those participants that had not completed the consent form prior to the start of their sessions were given time to complete the consent form before testing started. Once consents were complete, participants were read a brief script with information about how the testing process would work and when participation/recording in Zoom would end.

Participants who completed the study online (via Examus) received an emailed consent form links if they hadn't completed the consent form upon signing up. Those who completed consent forms and who indicated their computer passed the systems check received an Examus link, login and password info, as well as a complete description of how to complete testing. The lead researcher also sent reminders throughout the testing window to complete testing or cancel participation to those who had not completed their tests. To keep SONA slots open for participants who completed consent forms and were eligible to participate, participants who signed up via SONA but did not complete the consent forms were given a deadline by which to complete consents (usually 1-2 days). They received reminders to complete consents each day if they hadn't completed the consent form. If they did not complete consents in the specified time frame, their participation was cancelled by the researchers. An email notified removed participants that they could sign up again and complete the consent forms if they still wished to participate.

As aforementioned, we assigned participants to 1 of 5 groups. All participants completed six AMATs, with the first script occurring prior to AMAT 1, followed by AMATs 1-3, then the second script, followed by AMATs 4-6. Table 4 depicts this information for all groups.

Table 4

Order of tests and group conditions

Tests	Group H	Group I	Group J	Group K	Group L
	Script 1	*Script 1*	*Script 1*	*Script 1*	*Script 1*
AMAT 1	Neutral	Neutral	Neutral	Threat	Lift
AMAT 2	Neutral	Neutral	Neutral	Threat	Lift
AMAT 3	Neutral	Neutral	Neutral	Threat	Lift
	Script 2	*Script 2*	*Script 2*	*Script 2*	*Script 2*
AMAT 4	Neutral	Threat	Lift	Neutral	Neutral
AMAT 5	Neutral	Threat	Lift	Neutral	Neutral
AMAT 6	Neutral	Threat	Lift	Neutral	Neutral

Note. Neutral = statement that all test takers perform similarly. Threat = statement does not favor the group with which the individual identifies. Lift = statement favors the group with which the individual identifies.

Participants received the same post-test assessment (with added questions), as described in detail in the section “participant post-test responses and performance” and recording process through Examus in pilots 1-3.

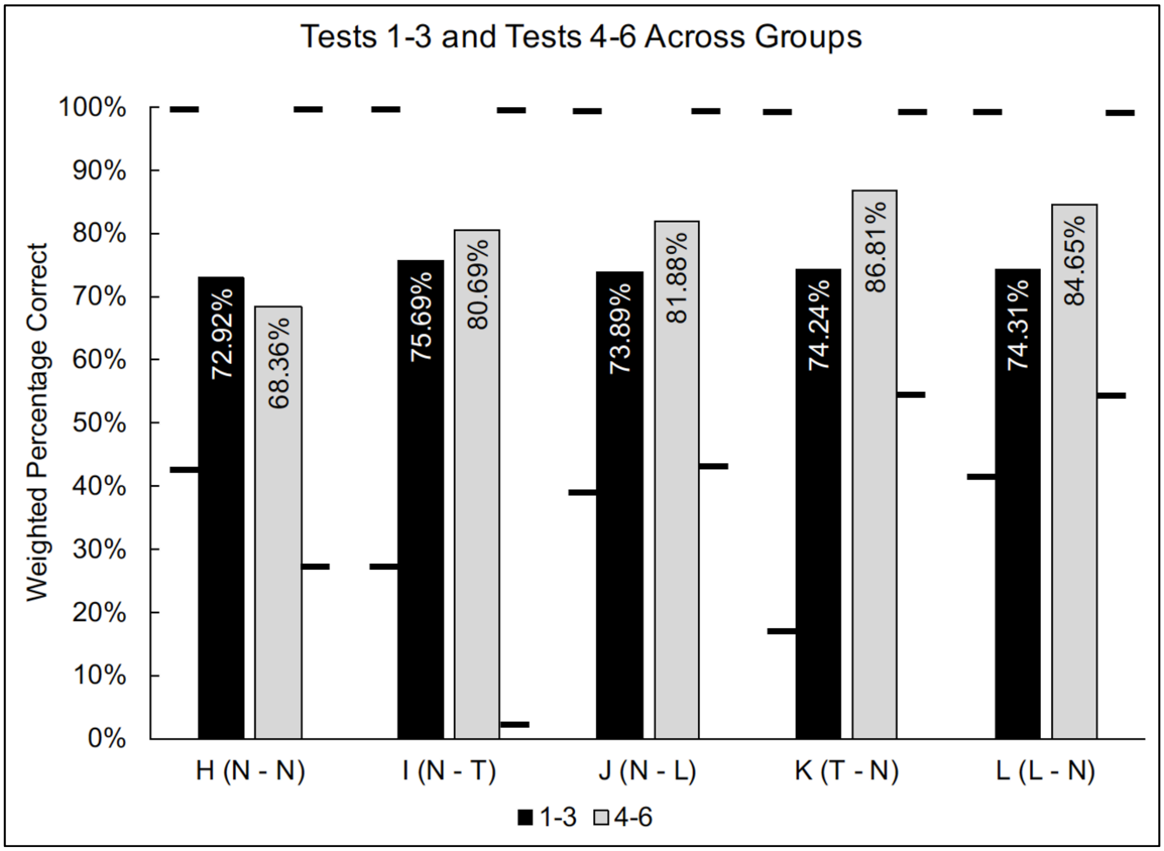
Results

Group Results

Figure 4.1 depicts the overall weighted averages for AMATs 1-3 and 4-6 for each of the 5 groups, along with the ranges in scores across participants in AMATs 1-3 and 4-6. With the exception of the control group, all group average scores increased from the first 3 AMATs to the subsequent 3 AMATS (4-6).

Figure 4.1

AMATs 1-3 and 4-6 averages across groups (weighted)



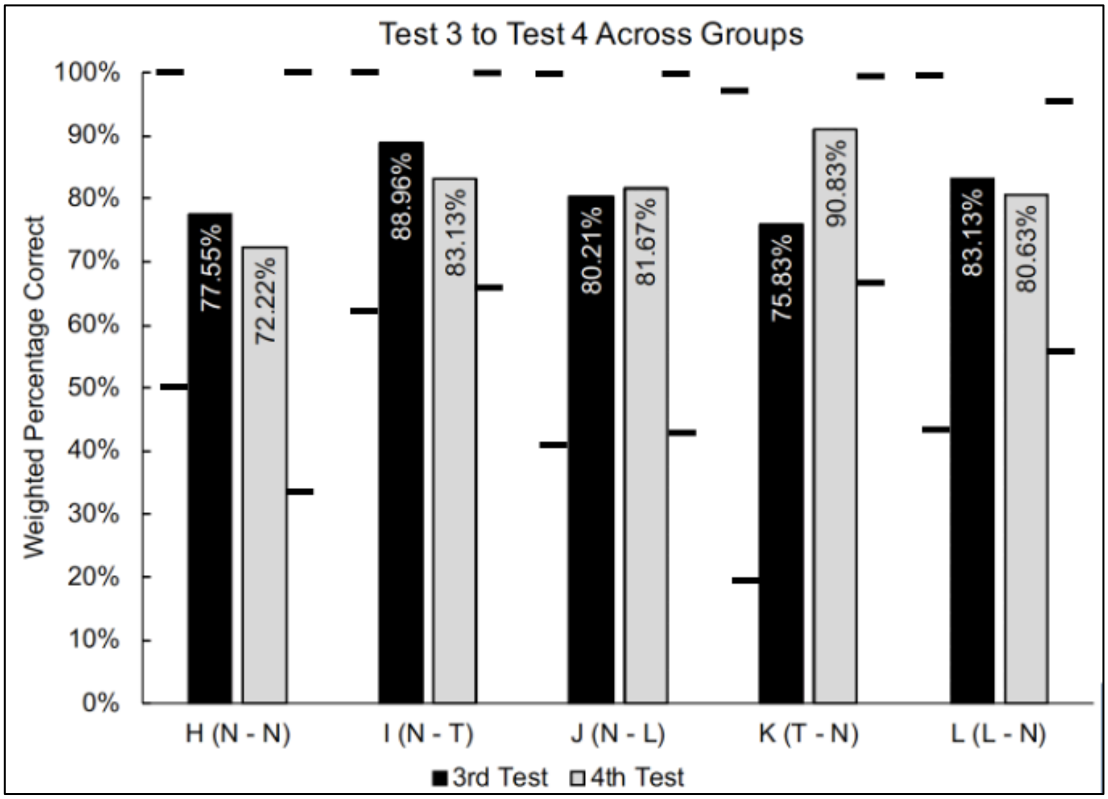
Note. Black range brackets beside the AMAT 1-3 and AMAT 4-6 average score bars show the range in scores across all participants and test ranges in each group. Ranges are as follows: Group H: AMAT 1-3 (43.8% - 100%), AMAT 4-6 (27.1%- 100%); Group I: AMAT 1-3 (27.1% - 100%), AMAT 4-6 (2.1% - 100%); Group J: AMAT 1-3 (39.6% - 100%), AMAT 4-6 (43.8% - 100%); Group K: AMAT 1-3 (18.8% - 100%), AMAT 4-6 (56.3% - 100%); Group L: AMAT 1-3 (41.7% - 100%), AMAT 4-6 (56.3% - 100%).

Although the range bars indicate participants in all groups were able to achieve 100% on at least one test in AMATs 1-3 and 4-6, it is interesting to note the lowest score obtained in each group. The data showed that the two lowest scores and hence largest ranges in performance occurred in the only two threat conditions (i.e., AMATs 4-6 for group I and AMATs 1-3 for group K).

These data should be interpreted with caution. Data from pilots 1-3 revealed one potential issue from grouping AMATs 1-3 and 4-6 together for analysis. Specifically, the majority of participants performed worse in the first MT. We hypothesize that this was because it was their first experience with taking these tests and many struggled adjusting to how questions were asked and with providing two answers to each question and a confidence level. That is, in Pilots 1-3 there was evidence of a testing effect where performance improved from Test 1 to Test 2 regardless of the condition for that test. This consistency of this results warranted further analyses to examine patterns in average scores across groups. Specifically, to rule out the practice effect present in the first test, we analyzed the data by comparing the group averages for AMAT 3 to 4 (representing the change in stereotyped statements). Figure 4.2 depicts the overall averages across each group for AMAT 3 and AMAT 4, along with the ranges in scores across participants in each group (similar to Figure 4.1). By isolating changes between AMATs 3 and 4, we could examine the direct effect observed from one condition to the next, after the new script was read between AMATs 3 and 4.

Figure 4.2

AMAT 3 and AMAT 4 averages across groups (weighted)



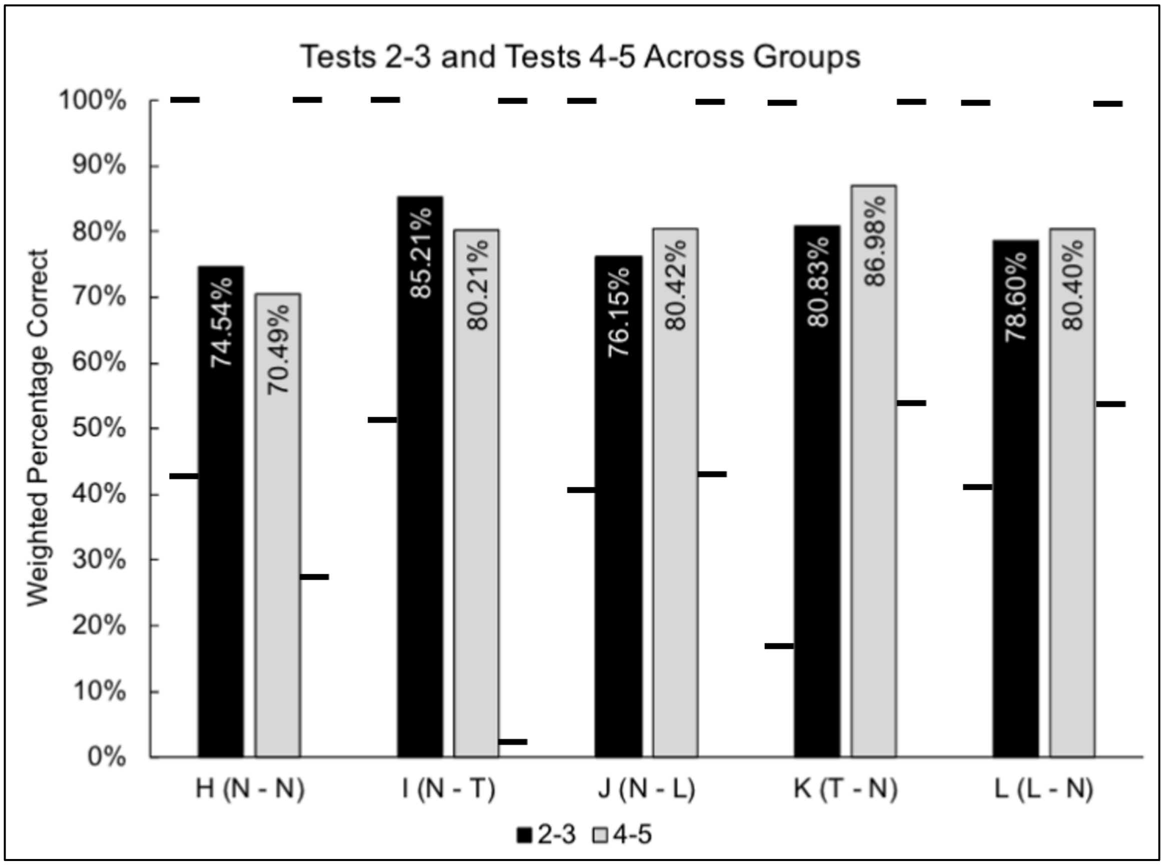
Note. Black range brackets beside the AMAT 3 and AMAT 4 average score bars show the range in scores across all participants in each group. Ranges are as follows: Group H: AMAT 3 (50% - 100%), AMAT 4 (33.3%- 100%); Group I: AMAT 3 (66.7% - 100%), AMAT 4 (33.3% - 100%); Group J: AMAT 3 (41.7% - 100%), AMAT 4 (43.8% - 100%); Group K: AMAT 3 (18.8% - 97.9%), AMAT 4 (68.8% - 100%); Group L: AMAT 3 (45.8% - 100%), AMAT 4 (58.3% - 95.8%).

A different pattern emerged in isolating averages for AMAT 3 and 4 across groups. Although the control group data remained similar to the analysis of averages for AMATs 1-3 and 4-6 in which performance averages were better for the first 3 tests than the last 3, the patterns for Groups I and L now reversed. Specifically, participants in both groups performed better in AMAT 3 than in AMAT 4. For group I, we observed a 5.82% decrease in average score from the neutral to threat conditions, and for group J participants, the data show a slight increase (1.46%) from the neutral to the lift condition. Group K data depict the largest change in average scores, with a 15% lower score average in the threat condition compared to the subsequent neutral condition. For the final group, group L, the data show a slight decrease in performance from the lift condition to the subsequent neutral condition (2.5%). Overall, averages across all participants for each group display the typical, expected stereotype threat and stereotype lift patterns. That is, participants (on average) showed increased performance scores in lift conditions and lower performance scores in threat conditions in comparison to the preceding or subsequent neutral conditions (although, the average differences were small for some groups, primarily those groups who received lift statements: groups J and L).

To see if this pattern remained consistent across AMATs 2-3 and 4-5, we analyzed these averages together. Figure 4.3 displays group averages for both AMATS 2 and 3 together, as well as 4 and 5 together, along with the ranges in scores across all 4 tests.

Figure 4.3

AMAT 2-3 and AMAT 4-5 averages across groups (weighted)



Note. Black range brackets beside the AMAT 3 and AMAT 4 average score bars show the range in scores across all participants in each group. Ranges are as follows: Group H: AMAT 2-3 (43.75% - 100%), AMAT 4-5 (27.1%- 100%); Group I: AMAT 2-3 (52.1% - 100%), AMAT 4-5 (2.1% - 100%); Group J: AMAT 2-3 (41.7% - 100%), AMAT 4-5 (43.8% - 100%); Group K: AMAT 2-3 (18.8% - 100%), AMAT 4-5 (56.3% - 100%); Group L: AMAT 2-3 (41.7% - 100%), AMAT 4-5 (56.3% - 100%).

The control group data show a 4.1% decrease in scores from AMATs 2-3 to 4-5, consistent with averages across AMATs 1-3 and 4-6 and AMATs 3 and 4. For group I, we observed a slightly smaller gap in averages scores from neutral to threat conditions (a 5% decrease down from 5.82% when isolating data from AMATs 3 & 4). For group J participants, the data show a larger increase from the neutral to the lift condition when averaging AMATs 2-3 and 4-5 than with AMATs 3-4 alone (4.27% up from 1.46%). Group K data show a smaller overall change in average scores (6.15% more accurate performance in the neutral AMATs 4-5 in comparison to the threat AMATs 2-3). This was down from a 15% change in average score observed when AMATs 3-4 averages were isolated for comparison. The largest change in overall averages across conditions again remained the largest in Group K. Interestingly, for group L, performance averages reversed from AMATs 3-4 alone, showing a 1.8% increase in the neutral AMATs 4-5 than in the lift AMATs 2-3 (averages from AMATs 3-4 showed a 2.5% higher average performance in the lift condition as opposed to the threat condition). Patterns tended to remain similar across groups in comparing average performance between AMATs 3 and 4 and AMATs 2-3 and 4-5 (with the exception of group L). Appendix G displays the same results, replaced with dichotomous scores, for further information.

Tables 5 and 6 show the results of a repeated measures ANOVA conducted across each test for each group. The results were examined based on an alpha of .05. The main effect for the within-subjects factor was not significant presents the ANOVA results across all groups. The means of the within-subjects factor across each group is presented in Table 6.

Table 5

Repeated Measures ANOVA Results

Source	*df*	*SS*	*MS*	*F*	*p*	η_p^2
Group H						
Within-Subjects						
Within Factor	3	0.04	0.01	0.59	.628	0.07
Residuals	24	0.49	0.02			
Group I						
Within-Subjects						
Within Factor	3	0.10	0.03	0.81	.499	0.08
Residuals	27	1.16	0.04			
Group J						
Within-Subjects						
Within Factor	3	0.05	0.02	0.81	.498	0.08
Residuals	27	0.60	0.02			
Group K						
Within-Subjects						
Within Factor	3	0.12	0.04	1.67	.226	0.16
Residuals	27	0.64	0.02			
Group L						
Within-Subjects						
Within Factor	3	0.04	0.01	0.76	.528	0.08
Residuals	27	0.48	0.02			

Table 6

Means Table for Within-Subject Variables

	Group H		**Group I**		**Group J**		**Group K**		**Group L**	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
AMAT 2	.72	.18	.68	.33	.72	.19	.86	.12	.76	.15
AMAT 3	.78	.17	.81	.20	.80	.20	.76	.27	.83	.17
AMAT 4	.72	.28	.70	.25	.82	.19	.91	.12	.81	.11
AMAT 5	.69	.22	.69	.29	.79	.22	.83	.17	.85	.19

Note. n = 9 (Group H); n = 10 (Groups I, J, K, & L).

Group comparisons of between and within subjects were also analyzed with a mixed model analysis of variance (ANOVA). The results were examined based on an alpha of .05. The main effect for Groups was not significant, *F*(4, 44) = 0.96, *p* = .437, indicating the levels of Group were all similar for AMAT 3, AMAT 4, AMAT 2, and AMAT 5. The main effect for the within-subjects factor was not significant, *F*(3, 132) = 0.94, *p* = .422, indicating the values of AMATs 2-5 were all similar for AMATs 2-5. The interaction effect between the within-subjects factor and Group was not significant, *F*(12, 132) = 1.00, *p* = .455, indicating that the relationships between AMATS 2-5 were similar between the levels of Group. Table 7 presents the mixed model ANOVA results.

Table 7

Mixed Model ANOVA Results

Source	*df*	*SS*	*MS*	*F*	*p*	η_p^2
Between-Subjects						
Group	4	0.31	0.08	0.96	.437	0.08
Residuals	44	3.53	0.08			
Within-Subjects						
Within Factor	3	0.06	0.02	0.94	.422	0.02
Group: Within Factor	12	0.26	0.02	1.00	.455	0.08
Residuals	132	2.84	0.02			

We also conducted a repeated measures ANOVA on all 5 groups for dichotomous scores to see if any results were statistically significant. In the same way that no results were statistically significant for weighted scores across groups in tests 2-5, results across dichotomous scores were also not statistically significant.

Additionally, repeated measures ANOVAs were also conducted on all AMATS (1-6). Given the issue with AMAT 1 tending to produce the lowest overall performance scores across most participant in each group (e.g., all experimental group averages displayed the lowest scores in AMAT 1), statistically significant results were shown for groups I, K, and L. Within effects, however, were only significant for Groups I and K. For Group I, AMAT 1 was significantly less than AMAT 6, *t*(9) = -3.78, *p* = .027, AMAT 1 was significantly less than AMAT 4, *t*(9) = -4.39, *p* = .011, and AMAT 1 was significantly less than AMAT 3, *t*(9) = -6.48, *p* < .001. For Group K, AMAT 1 was significantly less than AMAT 4, *t*(9) = -3.85, *p* = .017 and AMAT 1 was significantly less than AMAT 6, *t*(9) = -3.53, *p* = .027. No other significant differences were found between AMATs in Groups I and K. Again, these results were only significant when in comparison with AMAT 1. Appendix H displays these results.

In sum, the only results which were found to be statistically significant across all groups and all tests occurred when AMAT 1 data was included for analysis. As discussed previously, AMAT 1 scores tended to be lower than any other tests due to a testing effect, which presented an issue for analysis across all AMATs. However, when AMAT 1 data was not included for comparison, results from group analyses AMATS 3-4 and 2-3, 4-5 were not statistically significant.

Individual Results by Group and Gender

To better observe individual participants' performances across conditions from AMAT 3 to 4, the remaining figures (4.4 – 4.18) depict the individual participant results for all participants in the control group, separated by gender. These graphs represent participants across all groups (i.e., H-L). The bars in each graph depict participant performance scores on the primary y-axis, while the lines represent the total time spent studying for each participant in AMATs 3 and 4 (on the secondary y-axis). An example of an individual participant's results across each AMAT for all groups are also included.

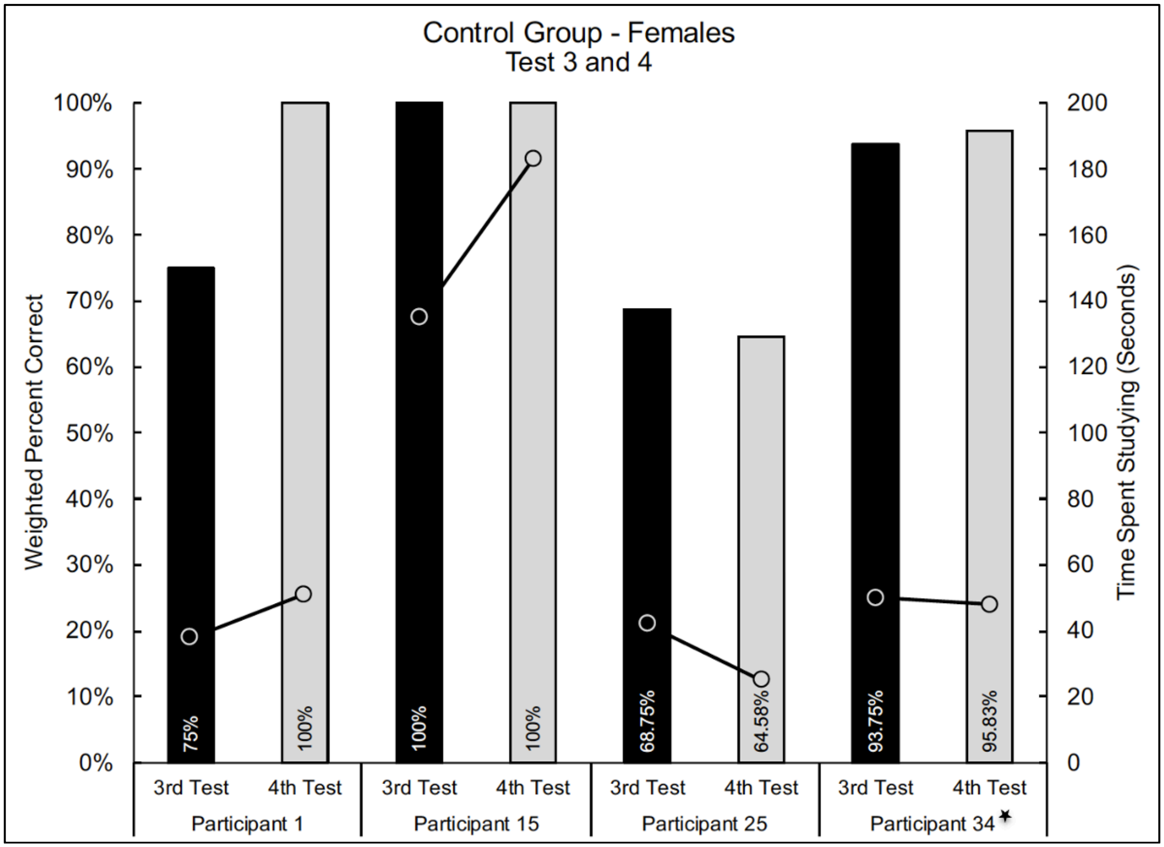
For further information, the same graphs, depicting performance for each individual across AMATS 2-3 and 4-5 are provided for each group in Appendix I.

Group H

Figures 4.4 and 4.5 display the data for the control group, or Group H.

Figure 4.4

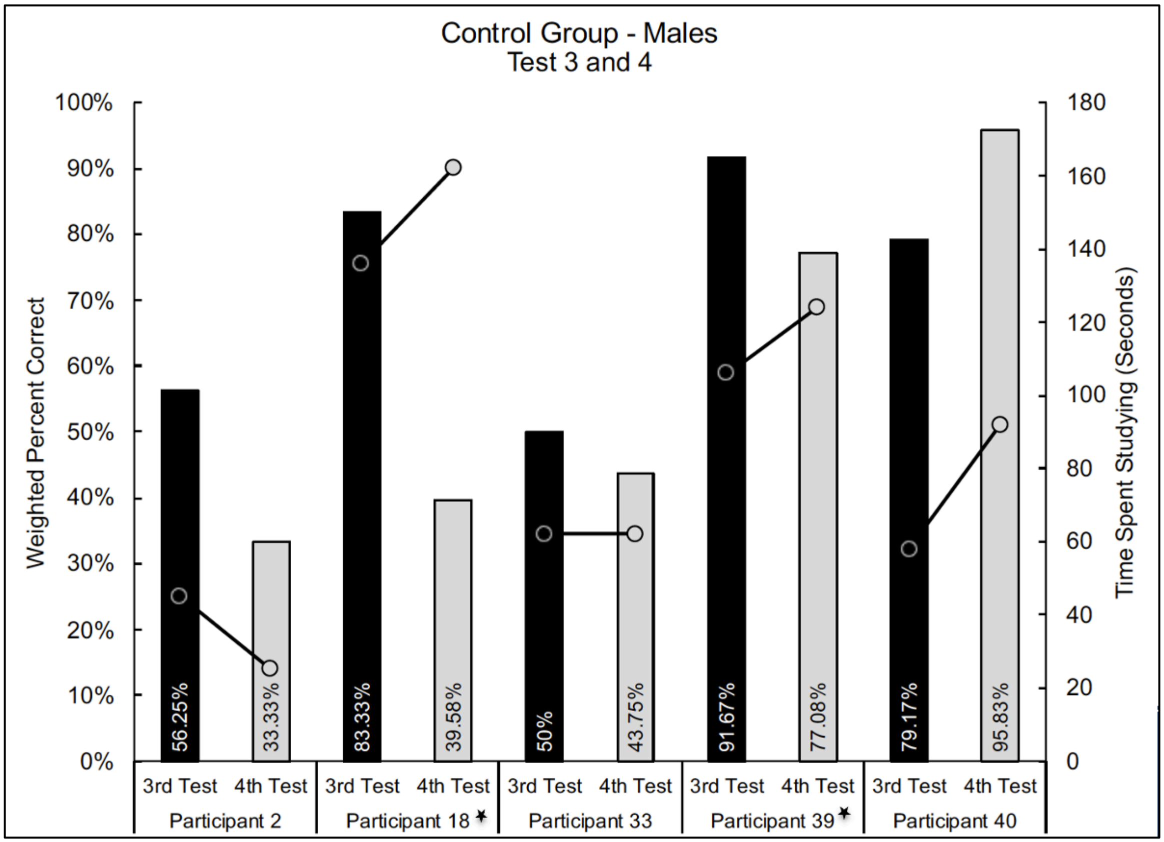
Group H female individual participant data across AMATs 3-4



Note. n = 4. The x-axis depicts AMAT 3 (3rd test) and AMAT 4 (4th test) for each individual participant in group H (control) for females. The bars represent the weighted percent correct on AMATs 3 and 4 for each participant (y-axis), while the lines represent total time spent studying for each test (secondary y-axis). Participants with an (*) participated in the study via Zoom.

Figure 4.5

Group H male individual participant data across AMATs 3-4



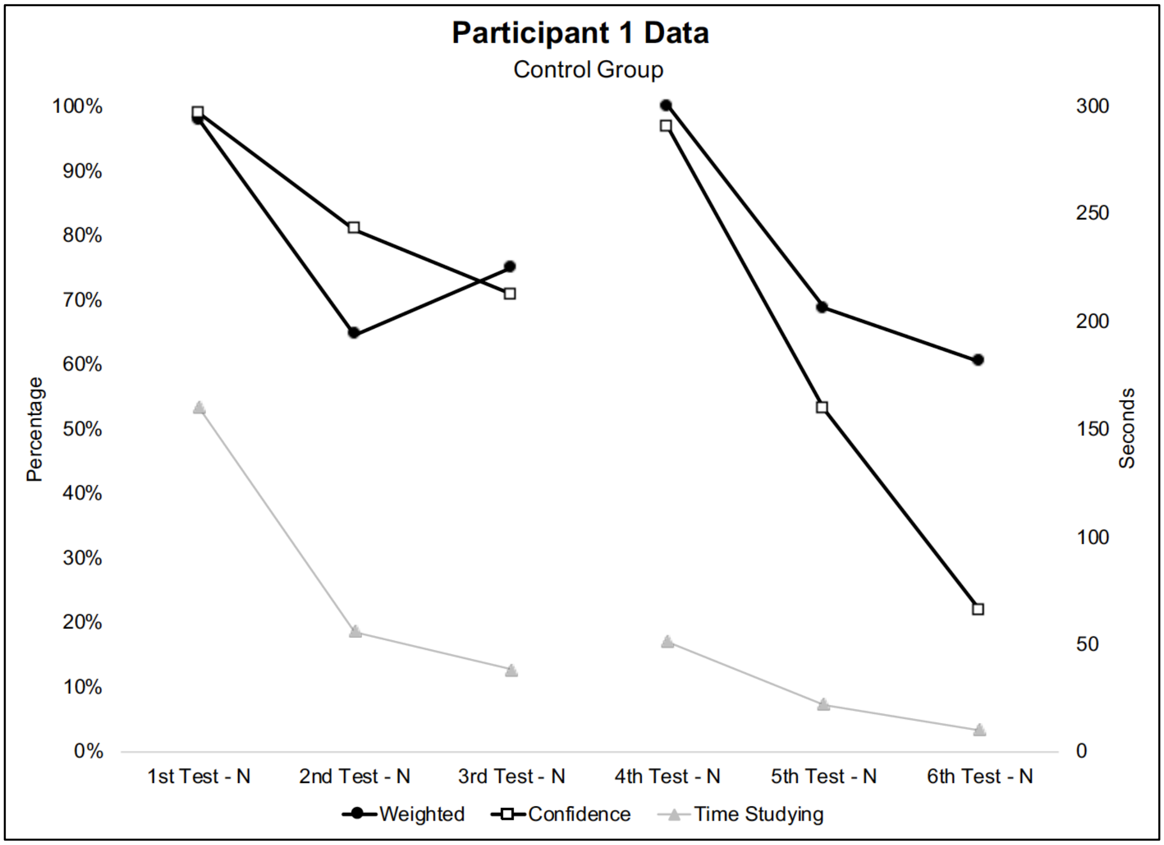
Note. n = 5. The x-axis depicts AMAT 3 (3rd test) and AMAT 4 (4th test) for each individual participant in group H (control) for females. The bars represent the weighted percent correct on AMATs 3 and 4 for each participant (y-axis), while the lines represent total time spent studying for each test (secondary y-axis). Participants with an (*) participated in the study via Zoom.

Overall across the control group, two participants performed better on AMAT 4 than AMAT 3 (2 females, 1 male), one participant performed the same (female), and five participants performed worse in AMAT 4 than AMAT 3 (1 female, 4 males). Five participants studied longer from AMAT 3 to 4 (2 females, 3 males), one studied the same amount as the AMAT prior (male), and three participants studied less from AMAT 3 to 4. These results are consistent with all three control group data from pilots 1-3, wherein control group participants overall tended to perform better up through memory test 3 or 4, then performance scores tended to decrease thereafter.

To illustrate individual data patterns in performance, confidence, time spent studying, and responses, figure 4.6 depicts the data observed from participant 1, who was a female in the control condition (group H).

Figure 4.6

Participant 1 data (female – H group)



Note. Participant stated in response to how the statements in the test impacted them: "I got stressed out when I started to run out of time and also when I got confused from seeing multiple things that I thought were just for one group"

For participant 1, data initially show that they spent the most amount of time (160 seconds) studying for the first AMAT for which they received a 97.9% weighted score.

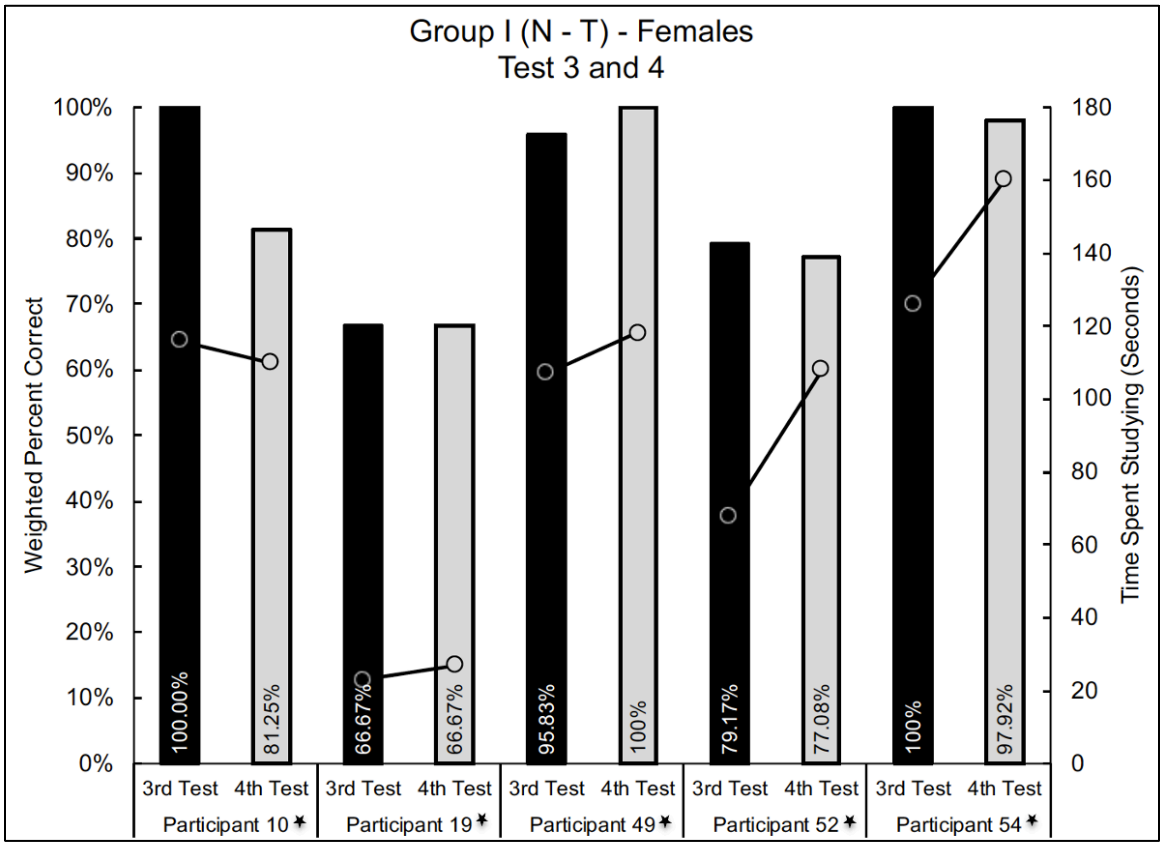
Then, their data show a decrease in score and time spent studying (64.5% and 56 seconds respectively), followed by an increase in scores from AMAT 3-4. Time spent studying and performance decrease in AMAT 5 (68.8%) and decrease further in AMAT 6 (60.4%). Although the highly accurate first test performance had been uncommon across all pilots and overall participant data for the current research, the increase in test scores through AMAT 4, followed by decreased scores thereafter, is consistent with previous control group data. Many control group participants reported frustration with seeing the same statement over and over again, and many studied less as time went on. One participant (whose data was removed from analysis) fell asleep during testing while he was taking one of the later AMATs. Although boredom and frustration were common trends observed across participants in the control group in their reported statements and behaviors during testing, these patterns remained consistent with the control groups in pilots 1-3.

Group I

Figures 4.7 and 4.8 depict the individual participant results for all group I participants.

Figure 4.7

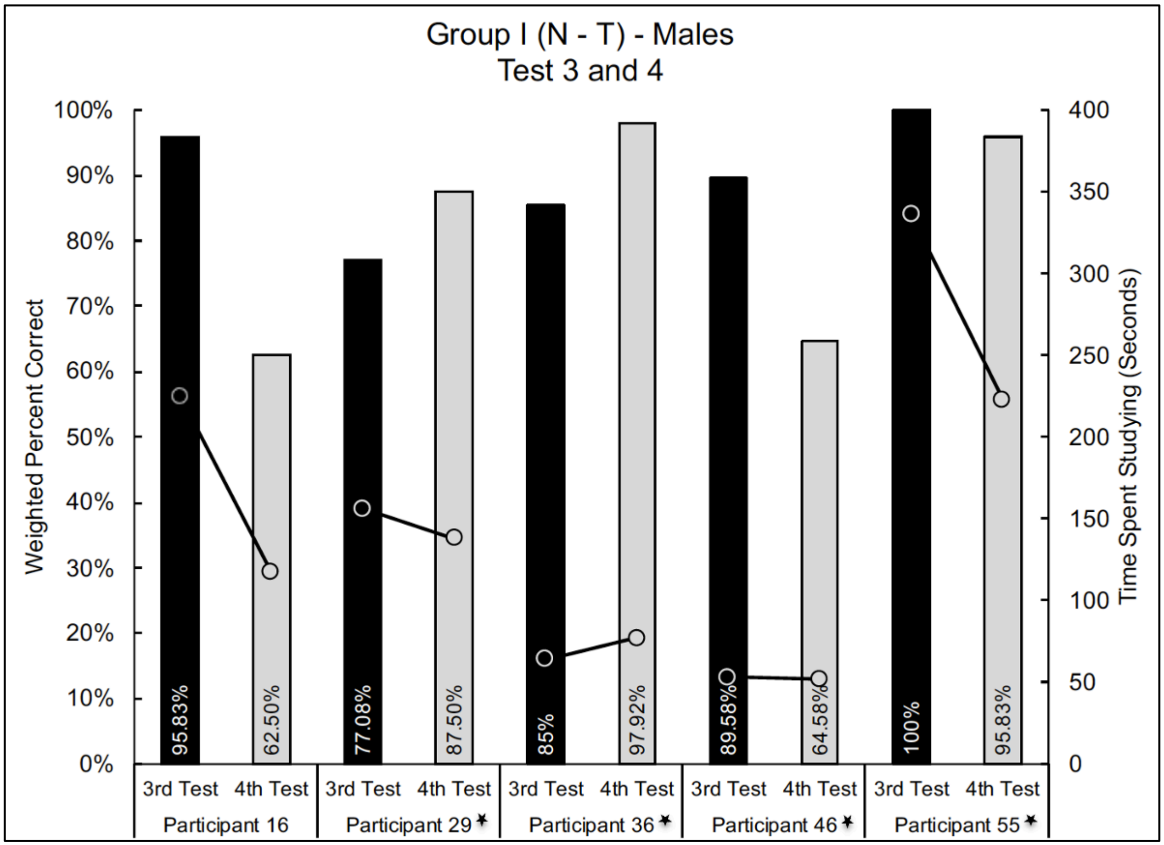
Group I female individual participant data across AMATs 3-4



Note. n = 5. The x-axis depicts AMAT 3 (3rd test) and AMAT 4 (4th test) for each individual participant in group H (control) for females. The bars represent the weighted percent correct on AMATs 3 and 4 for each participant (y-axis), while the lines represent total time spent studying for each test (secondary y-axis). Participants with an (*) participated in the study via Zoom.

Figure 4.8

Group I male individual participant data across AMATs 3-4



Note. n = 5. The x-axis depicts AMAT 3 (3rd test) and AMAT 4 (4th test) for each individual participant in group H (control) for females. The bars represent the weighted percent correct on AMATs 3 and 4 for each participant (y-axis), while the lines represent total time spent studying for each test (secondary y-axis). Participants with an (*) participated in the study via Zoom.

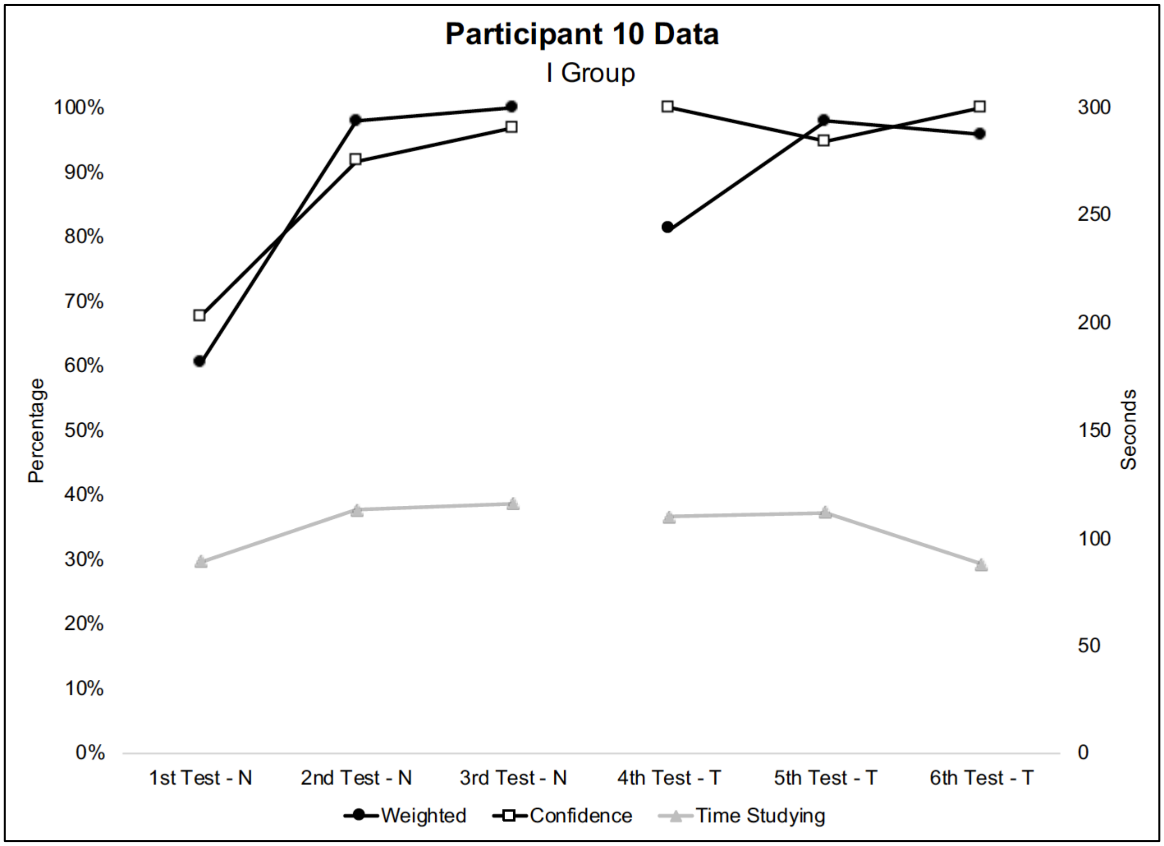
Group I included three participants who performed with higher accuracy on AMAT 4 than the previous AMAT 3 (1 female, 2 males), one participant performed the same from AMAT 3 to 4 (female), and six participants performed with lower accuracy in AMAT 4 than AMAT 3 (3 females, 3 males). Thus, most participants (70%) performed with lower accuracy or the same from the neutral condition to the subsequent threat condition. Five participants studied longer from AMAT 3 to 4 (2 females, 3 males), one studied the same amount as the prior AMAT (male), and three participants studied less from AMAT 3 to 4.

Four of five males studied less in the neutral AMAT 3 to the threat AMAT 4, with three of those four performing worse from AMAT 3 to 4. However, four of five females studied longer in the threat condition than in the neutral condition prior, with only one of those females performing better.

Figure 4.9 depicts participant 10's data on performance, confidence, time spent studying, and responses to the gendered statement.

Figure 4.9

Participant data (female – I group)



Note. Participant stated that they thought they would perform better between threat and neutral conditions, and in response to the question of how it impacted them, said: "I feel like i [sic] became more focused because I wanted to prove the statement wrong"

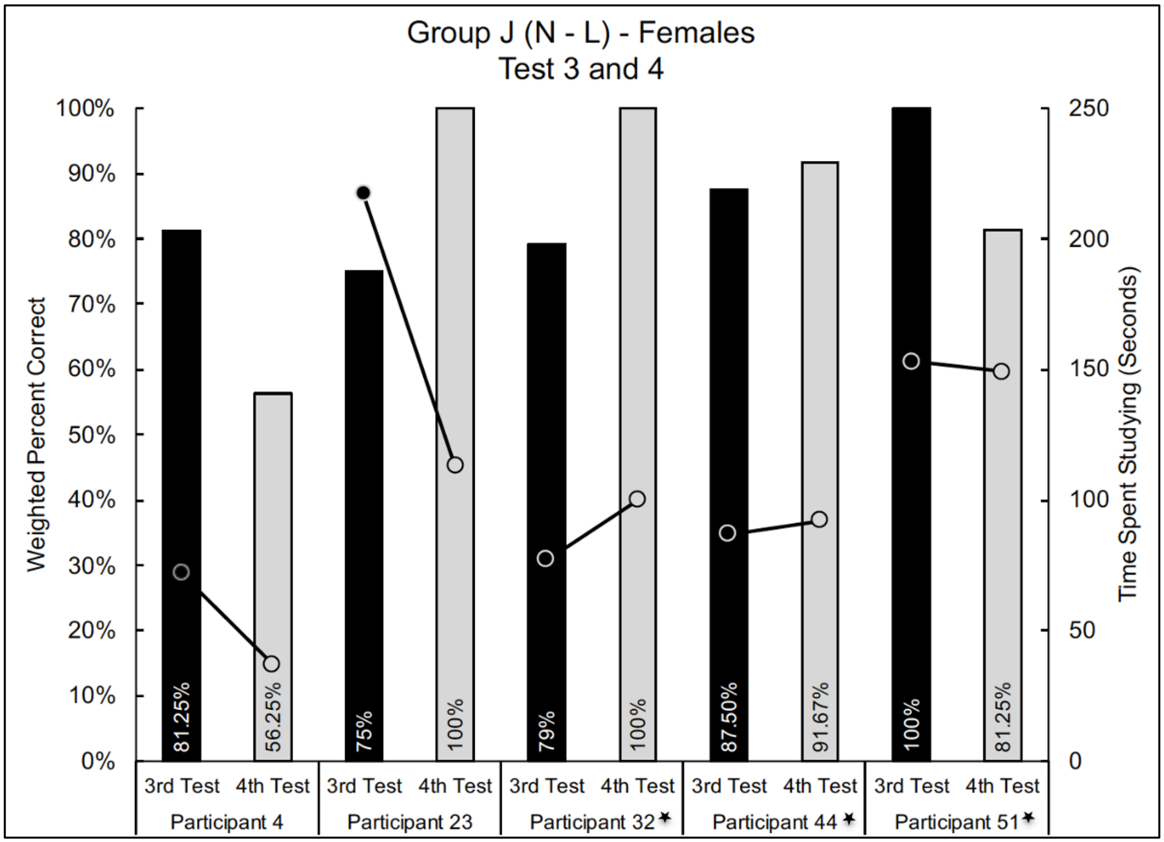
Participant 10's data show an increasing trend from AMAT 1 on with the exception of the first introduced threat statement (AMAT 4). Although the participant noted that they thought they would do better in the threat conditions, the impact of the initial threat observed from AMAT 3-4 indicates the opposite. Although participant 10 did improve in scores in AMATs 5 and 6, and the overall averages in scores for AMATs 1-3 and 4-6 for this participant indicate improvement into the threat condition, it seems that the threat statement still impacted performance as indicated by the decrease in score for AMAT 4 when the threat statement was first introduced.

Group J

Figures 4.10 and 4.11 depict the individual participant results for all group J participants.

Figure 4.10

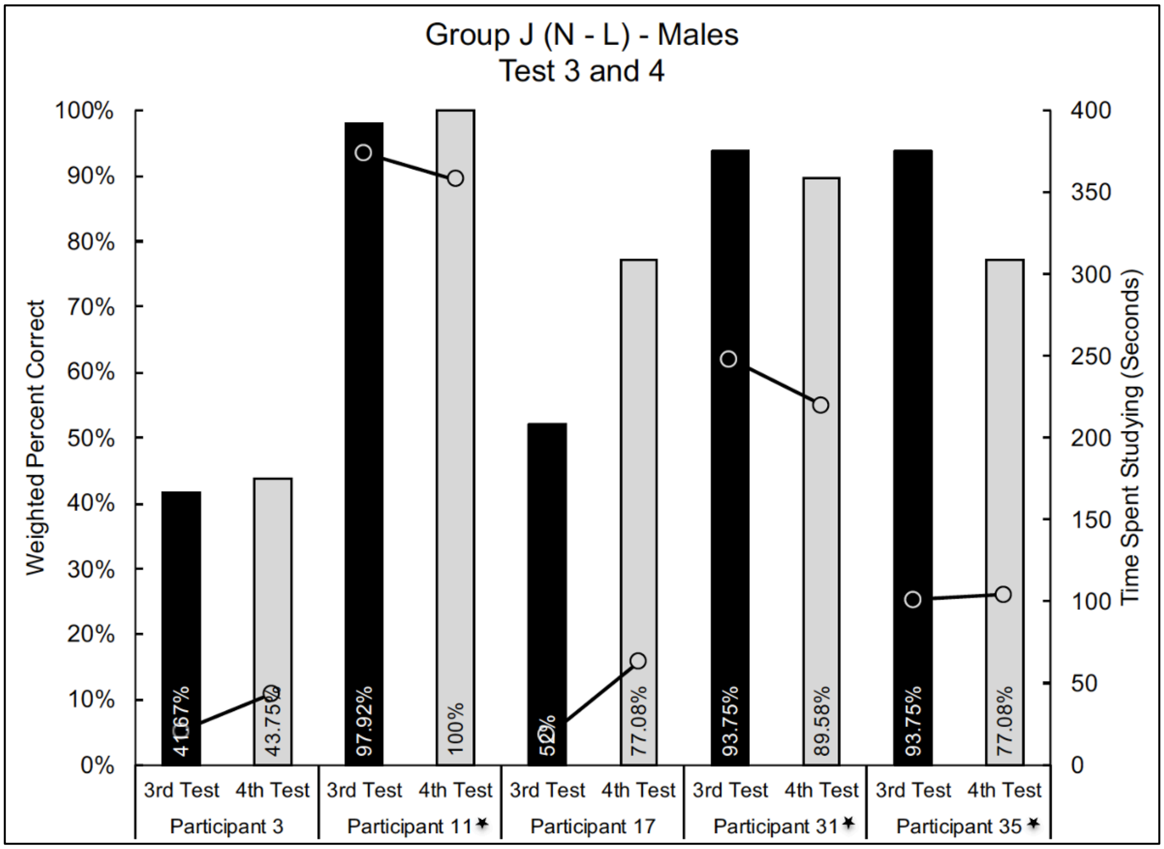
Group J female individual participant data across AMATs 3-4



Note. n = 5. The x-axis depicts AMAT 3 (3rd test) and AMAT 4 (4th test) for each individual participant in group H (control) for females. The bars represent the weighted percent correct on AMATs 3 and 4 for each participant (y-axis), while the lines represent total time spent studying for each test (secondary y-axis). Participants with an (*) participated in the study via Zoom.

Figure 4.11

Group J male individual participant data across AMATs 3-4



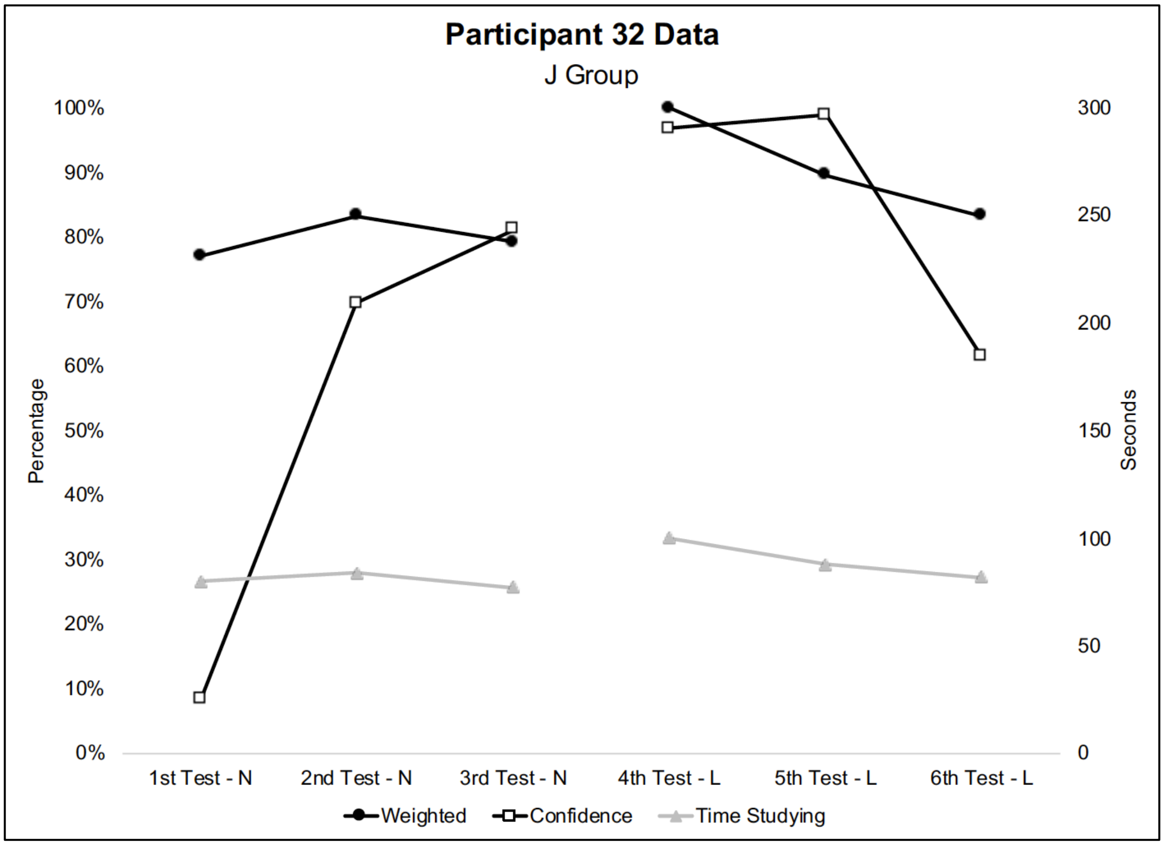
Note. n = 5. The x-axis depicts AMAT 3 (3rd test) and AMAT 4 (4th test) for each individual participant in group H (control) for females. The bars represent the weighted percent correct on AMATs 3 and 4 for each participant (y-axis), while the lines represent total time spent studying for each test (secondary y-axis). Participants with an (*) participated in the study via Zoom.

Group J included six participants who performed with higher accuracy in the lift condition AMAT 4 than the previous neutral condition AMAT 3 (3 females, 3 males) and four participants who performed with lower accuracy in AMAT 4 than AMAT 3 (2 females, 2 males). An even number of participants studied less or more between AMAT 3 and 4 (i.e., five and five).

Figure 4.12 shows data from one participant in group J (participant 32).

Figure 4.12

Participant 32 data (female – J group)



Note. Participant noted "I was a little distracted because I was thinking about the subliminal messages" in response to the lift statement.

Participant 32's data show overall higher accuracy across all lift conditions AMATs (4-6) than neutral performance. Although they rated their confidence in the first and final AMATs lower than the rest of each respective group (1-2 and 4-6), confidence was also higher for the lift tests as opposed to the neutral tests.

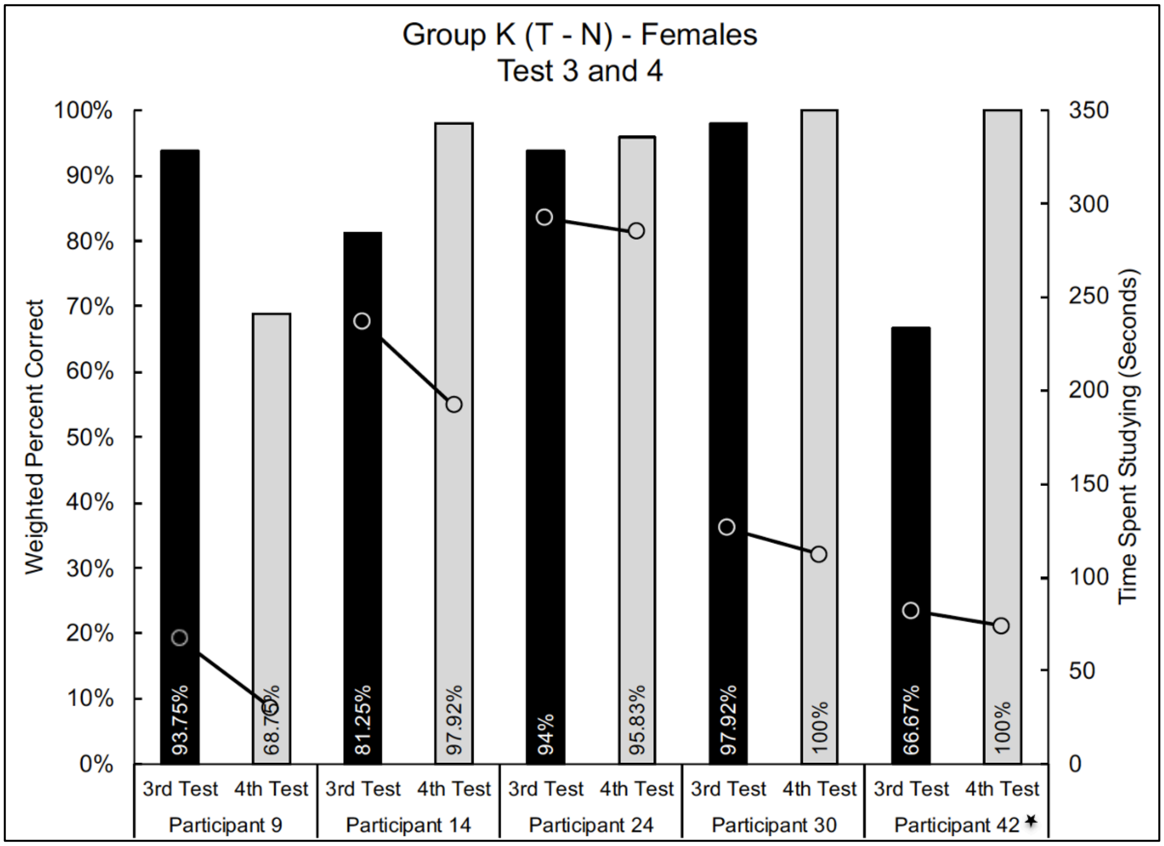
Of note, however, is the participant's response to how the gendered statement impacted their performance. Their feedback indicated that they were "a little distracted because [they were] thinking about the subliminal messages." Although the participant demonstrated higher accuracy on all lift condition AMATs (i.e., 4-6) than they did in all neutral conditions, this participant seemed to note that the statement was more distracting, rather than resulting in them performing better, worse, or the same. Participant 32 did spend longer studying in AMAT 4 than in any previous or subsequent AMAT (100 seconds, as opposed to 77 in AMAT 3 and 88 in AMAT 5).

Group K

Figures 4.13 and 4.14 depict the individual participant results for all group K participants.

Figure 4.13

Group K female individual participant data across AMATs 3-4

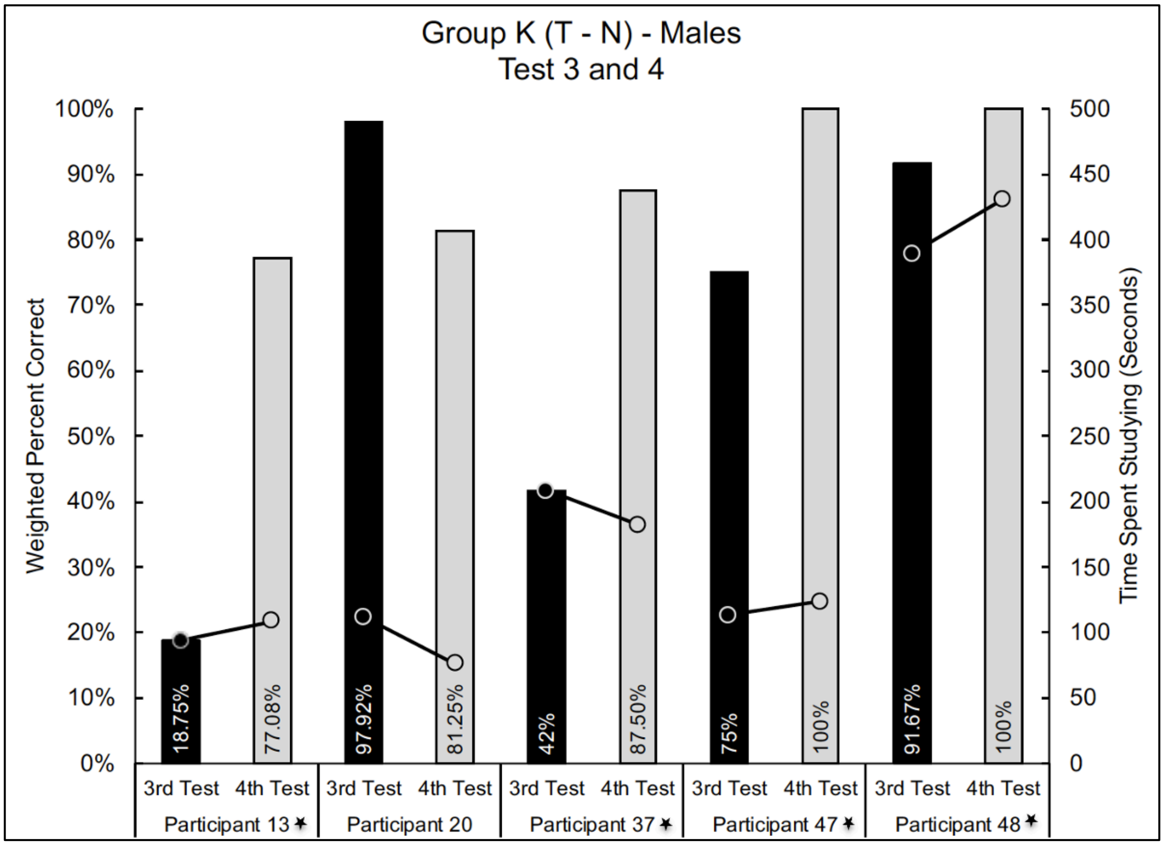


Note. n = 5. The x-axis depicts AMAT 3 (3rd test) and AMAT 4 (4th test) for each individual participant in group H (control) for females. The bars represent the weighted percent correct on AMATs 3 and 4 for each participant (y-axis), while the lines represent total time spent studying for each test (secondary y-axis). Participants with an (*)

participated in the study via Zoom.

Figure 4.14

Group K male individual participant data across AMATs 3-4



Note. n = 5. The x-axis depicts AMAT 3 (3rd test) and AMAT 4 (4th test) for each individual participant in group H (control) for females. The bars represent the weighted percent correct on AMATs 3 and 4 for each participant (y-axis), while the lines represent total time spent studying for each test (secondary y-axis). Participants with an (*) participated in the study via Zoom.

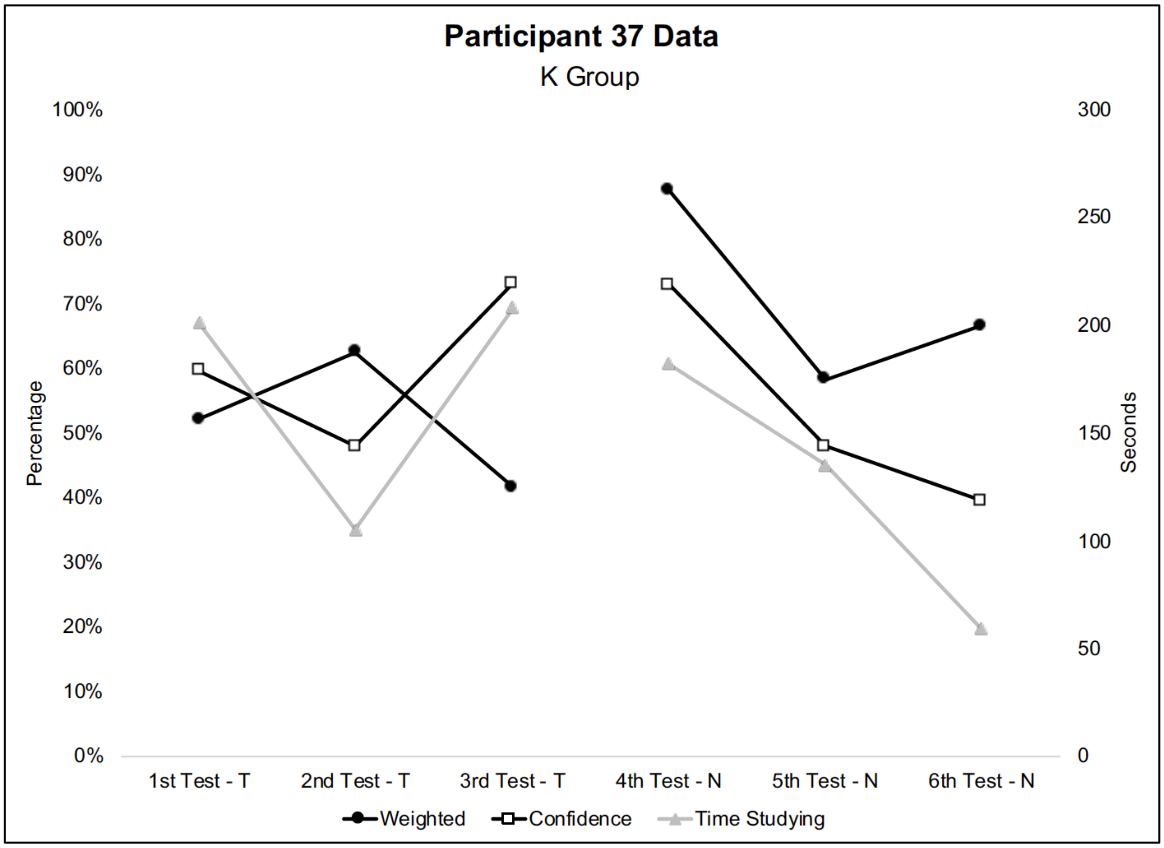
Group K data show the most uniform performance changes across participants. Eight of the ten participants (80%) performed with higher accuracy in the neutral condition (AMAT 4) than the previous threat condition (AMAT 3) (4 females, 4 males) with only two participants performing with lower accuracy in AMAT 4 than AMAT 3 (1 female, 1 male). Interestingly, only three participants (all males) studied longer in AMAT

4 than 3, as most participants studied less in AMAT 4 but overall showed higher accuracy. The two participants who did perform with lower accuracy from the threat to the neutral condition studied considerably less. Participant 9 decreased from studying for 67 seconds in the threat condition to 30 seconds in the subsequent neutral condition (37 seconds less), and participant 20 studied 36 seconds less (112 to 76).

Figure 4.15 shows the data for one of the group K female participants, participant 37.

Figure 4.15

Participant 37 data (female – K group)



Note. Participant stated that they thought they would perform the same between threat and neutral conditions, and in response to the question of how it impacted them, stated “It might've caused a slight change but I'm not sure. It was funny at first but after looking at it a couple times I ignored it.”

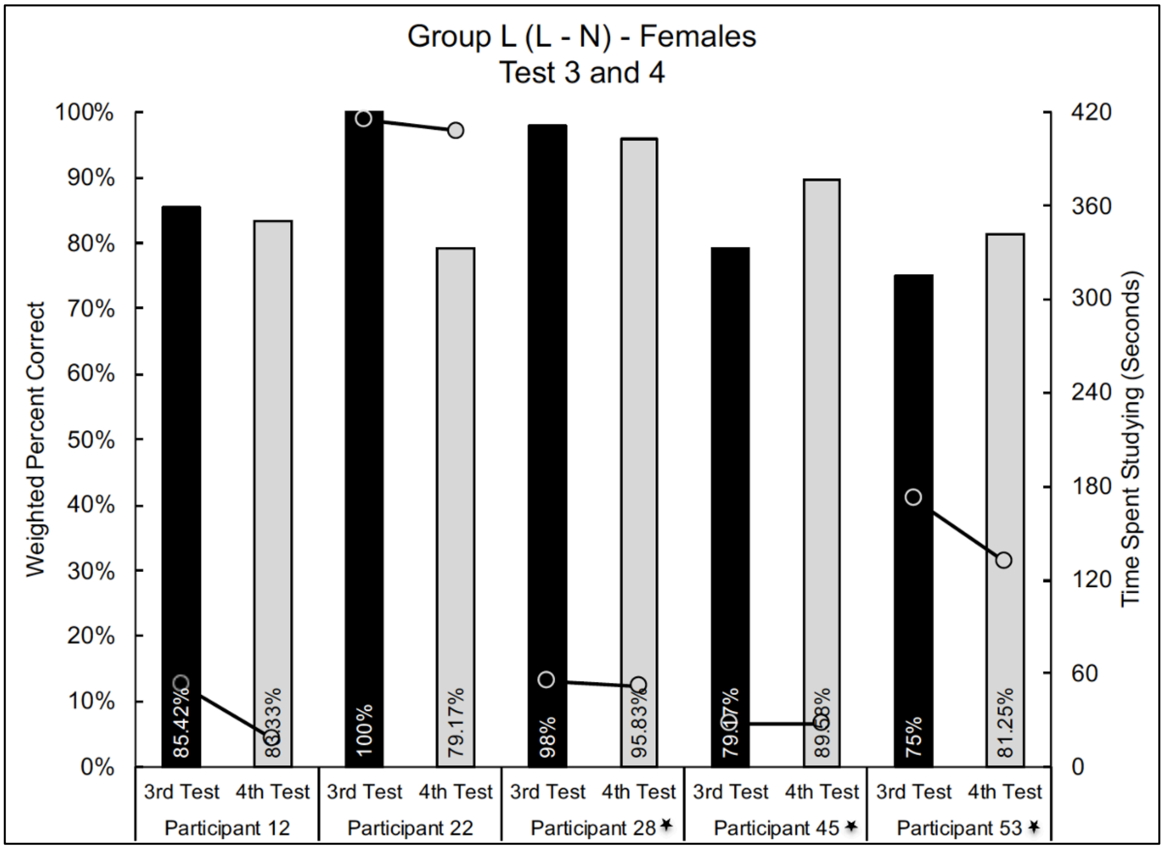
Although participant 37 marked they thought they thought they would perform the same in the threat conditions (AMATs 1-3) in relation to the neutral conditions (AMATs 4-6), they did acknowledge that the statement may have caused a “slight change.” Their data, however, show a sizeable change in score from AMAT 3 to AMAT 4, with AMAT 4 showing the most accurate performance for participant 37 overall (87.5%) and AMAT 3 showing the lowest score (41.7%). Interestingly, they studied the most for AMATs 1 and 3 (201 and 208 seconds, respectively), and yet had their lowest accuracy performances in those AMATs (which were both threats). In addition, their confidence was highest in AMAT 3, while their performance score displayed the lowest performance score that participant demonstrated throughout all tests.

Group L

Figures 4.16 and 4.17 depict the individual participant results for all group L participants.

Figure 4.16

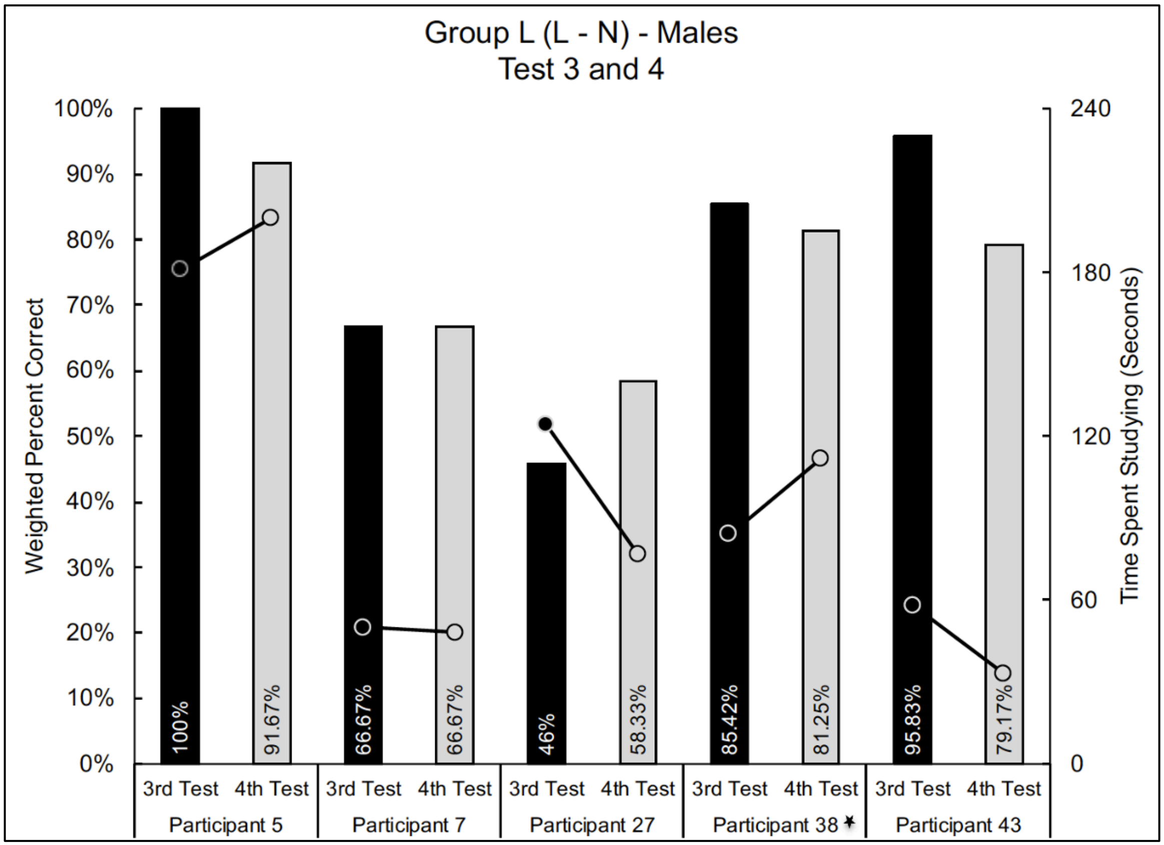
Group L female individual participant data across AMATs 3-4



Note. n = 5. The x-axis depicts AMAT 3 (3rd test) and AMAT 4 (4th test) for each individual participant in group H (control) for females. The bars represent the weighted percent correct on AMATs 3 and 4 for each participant (y-axis), while the lines represent total time spent studying for each test (secondary y-axis). Participants with an (*) participated in the study via Zoom.

Figure 4.17

Group L male individual participant data across AMATs 3-4



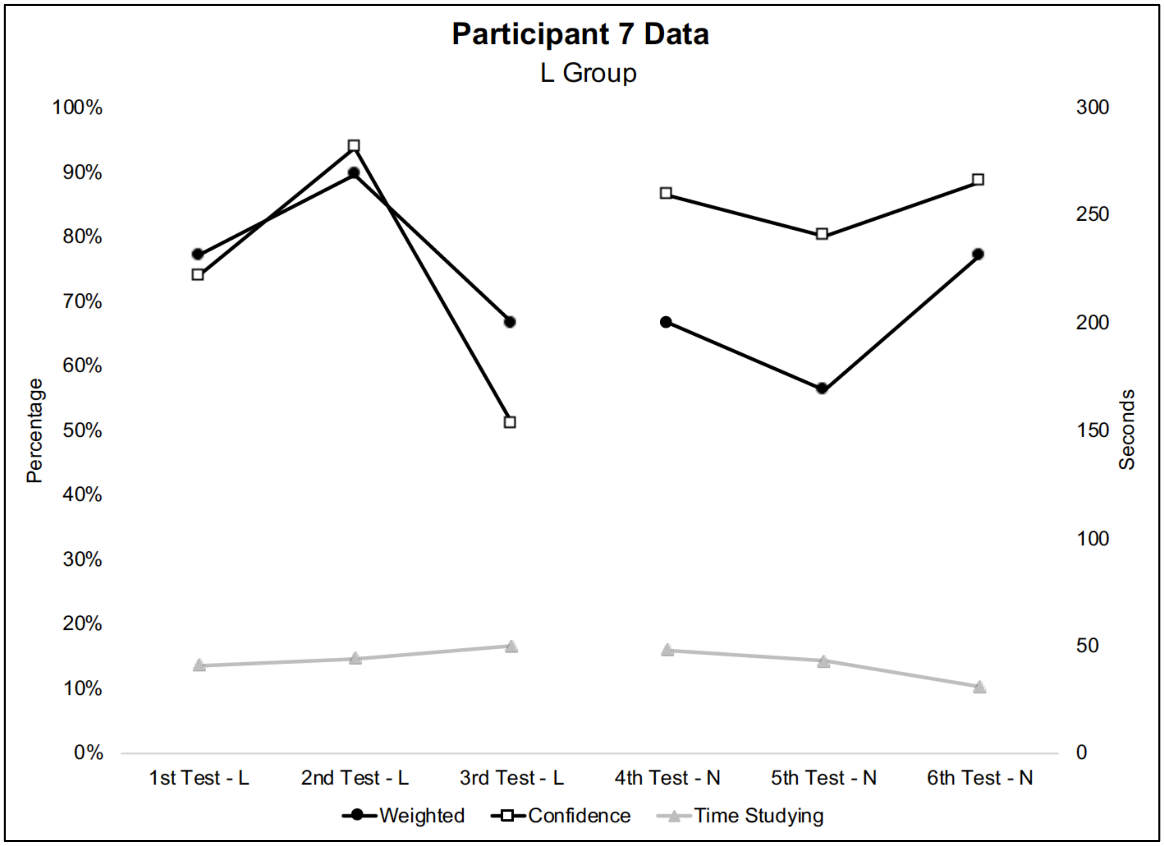
Note. n = 5. The x-axis depicts AMAT 3 (3rd test) and AMAT 4 (4th test) for each individual participant in group H (control) for females. The bars represent the weighted percent correct on AMATs 3 and 4 for each participant (y-axis), while the lines represent total time spent studying for each test (secondary y-axis). Participants with an (*) participated in the study via Zoom.

Group L included three participants who performed with higher accuracy in the neutral condition (AMAT 4) than the previous lift condition (AMAT 3) (2 females, 1 male), one participant performed the same from AMAT 3 to 4 (male), and six participants performed with lower accuracy in AMAT 4 than AMAT 3 (3 females, 3 males). Thus, most participants (70%) had lower or the same performance scores from the lift condition to the subsequent neutral condition. Only two participants studied longer (both males) from AMAT 3 – 4, and one participant studying the same amount from 3-4.

Figure 4.18 shows the data for one of the group L male participant, participant 7.

Figure 4.18

Participant 7 data (male – L group)



Note. Participant stated that they thought they would perform better between lift and neutral conditions, and in response to the question of how it impacted them, said: “I thought the pressure of expecting to do better was a lot and may have impacted my memory due to pressure of expectations.

For participant 7, performance remained the same between AMAT 3 and 4. However, data isolating participant 7’s AMAT 2-3 and 4-5 performance, as well as 1-3 and 4-6 performance, reveal an overall increase in performance averages for the lift condition. Although the participant discussed the pressure of expectations to do well, and technically performed the same across AMATs 3-4, their performance was in line with their answer to the direct impact question that they would perform better in the lift conditions.

For further detail regarding overall group confidence comparisons in relation to time spent studying and performance scores, Appendix J displays group averages, confidence, and time spent studying across AMATs 1-6 for each group.

In sum, data trends in the individual participant performance revealed interesting patterns. The current research more clearly showed the impact of the threat/lift statements once these statements were isolated. Still, the effect observed in testing between conditions directly prior to and after the longer gendered scripts often showed the most robust differences in performance. An important feature of the experiment was participant post-test responses in comparison to their performance. The following section, will discuss this data.

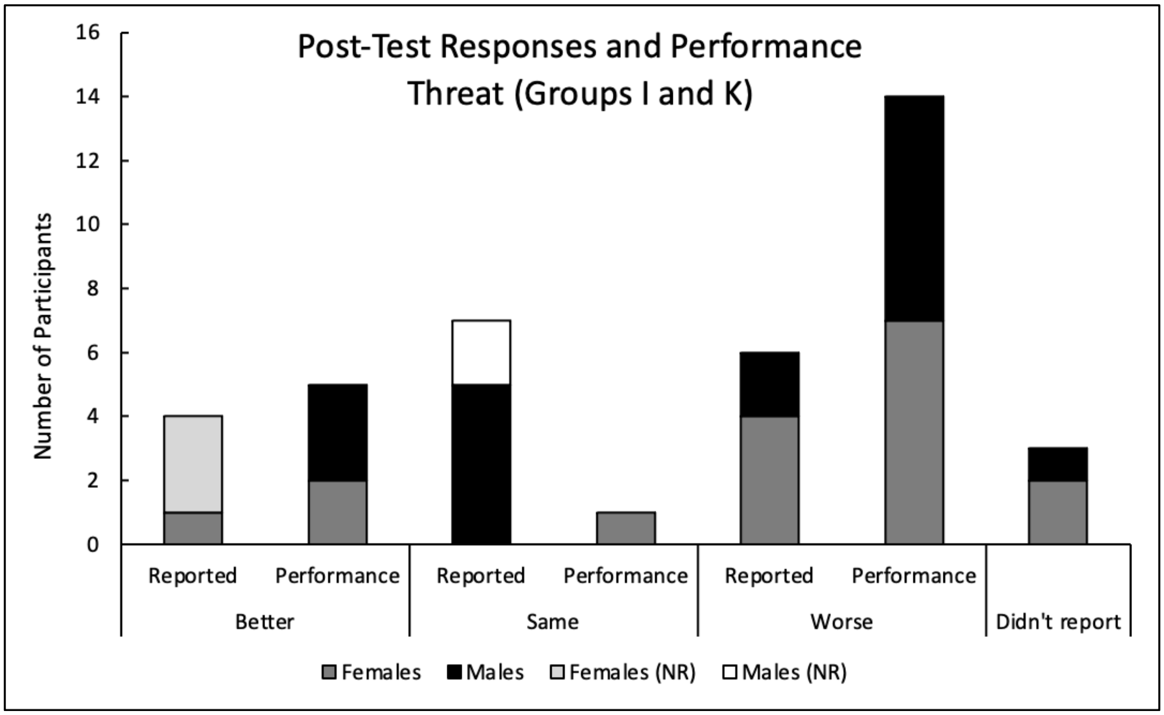
Participant Post Test Responses and Performance

Data from the post-tests in pilots 1-3 revealed interesting responses from participants with respect to their encountered lift and threat statements. However, as these responses were open-ended, a useful qualitative analysis of these statements was difficult to accurately portray. To enable a more consistent analysis, we altered the post-test to include another question, which was displayed whenever the participant selected that they saw one of the two gendered statements (see Appendix A for the Post-test exit survey and added questions). Participants who selected that they saw “men perform better than women” or “women perform better than men” were asked how they thought they would perform on those tests in relation to the other (neutral) AMATs they took. Participants could select that they thought they would perform better, the same, or worse on those tests (there was also an “other” option where participants could write in their responses). Figures 4.19 and 4.20 display participants responses to these questions for those that experienced threat scripts/statements (Groups I and K) and lift scripts/statements (Groups J and L). These are analyzed by difference in AMAT 3 and 4 scores.

Data for all participants who experienced threat statements (i.e., Groups I and K) are depicted in Figure 4.19, separated by gender.

Figure 4.19

Post-test responses and performance changes for participants in threat conditions



Note. n = 20. Performance is reported on score differences between AMATs 3 to 4. Females (NR) and Males (NR) were coded for participants who did not formally report they thought they did better, the same, or worse but who mentioned one option in their comments.

Four participants reported that they thought they did better on the threat tests than the neutral tests (one formally and three informally in the comments section), while seven reported that they thought they performed the same as in the neutral tests, six reported they thought they did worse in the threat tests than the neutral tests, and three did not formally report on the impact or mention anything related to the gendered statements in their comments. Five participants had higher performance scores in the threat condition, only one performed the same as their neutral test, and 14 participants had lower performance scores in the threat test than in the neutral test. Thus, participants in these groups seemed to underestimate the impact of these statements on their performance between AMATs 3 and 4. Specifically, most reported that they were not impacted by the received statements (e.g., participant 16 – "did not effects [sic] my performance/thoughts at all").

Ultimately, the threat condition did appear to impact participants as 70% of participants had lower performance scores in the threat conditions than the neutral conditions and overall group averages in score changes between AMAT 3 and 4 for both groups I and K indicated an overall lower performance for participants in the threat condition. These data seem to support the current stereotype threat literature.

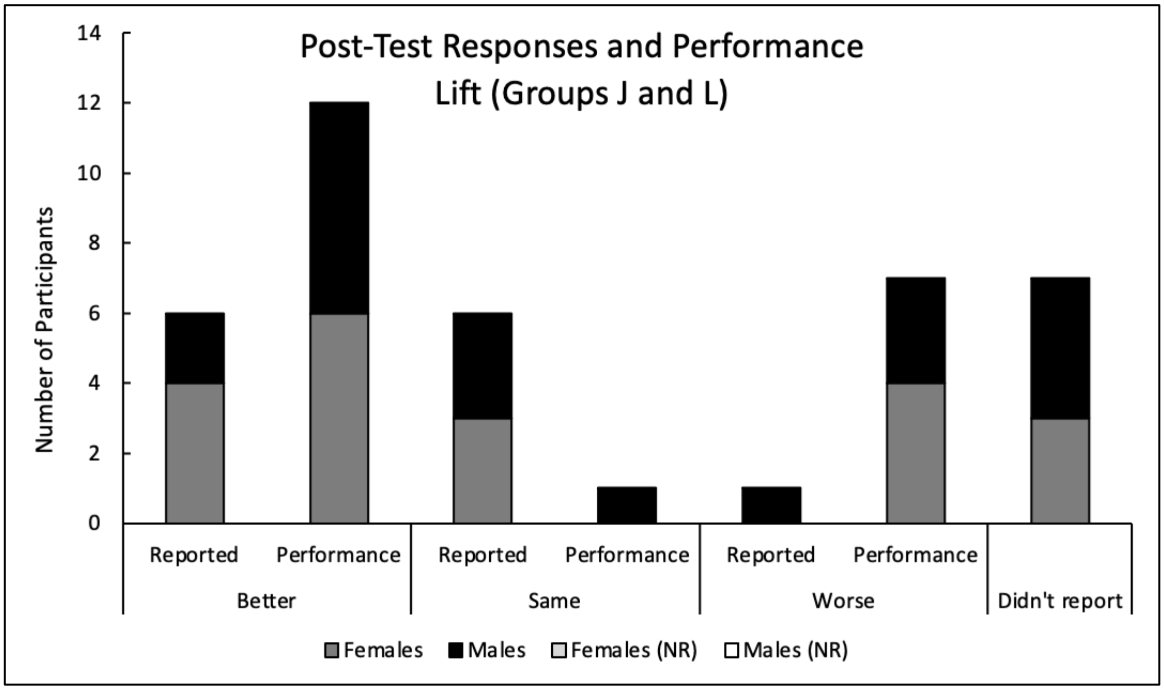
Although most participants performed had lower performance scores in the threat condition, some interesting trends were also noted from those who had higher performance scores in the threat condition. Participants who indicated that they thought they would do better in the threat condition all discussed ways in which they changed their testing behavior to make it more likely that they performed better on the AMATs for which the threat scripts and statements were present. For instance, one participant stated that: "I applied myself more than I have in previous psychology surveys to prove to myself that I as an individual am equal to men (if not better)". Another participant stated

"I feel like i [sic] became more focused because I wanted to prove the statement wrong". Similar reactions and responses have also been noted in previous research (i.e., easier tasks can produce the opposite effect when a threat is present).

Data for all participants who experienced lift statements (i.e., Groups J and L) are depicted in Figure 4.20, separated by gender.

Figure 4.20

Post-test responses and performance changes for participants in lift conditions



Note. n = 20. Females (NR) and Males (NR) were coded for participants who did not formally report they thought they did better, the same, or worse but who mentioned one option in their comments.

Six participants reported that they thought they did better on the threat tests than the neutral tests. Six reported that they thought they performed the same as in the neutral tests. One reported they thought they did worse on the threat tests than the neutral tests. Seven did not formally report on the impact or mention anything related to the gendered statements in their comments. Twelve participants performed with higher accuracy in the lift test. One performed the same as they had in the neutral test. Seven participants performed with lower accuracy in the lift test than in the neutral test.

Although the impact of the lift condition compared to the neutral condition was smaller than the data from the threat groups (i.e., I and K), participants ultimately had higher performance scores in the lift condition (60% of participants demonstrate higher accuracy in the lift condition than the neutral conditions). This was consistent with group averages across AMATs 3 and 4 in which participants had higher performance scores in the lift conditions.

Discussion

The current research aimed to arrange conditions under which stereotype threat and stereotype lift would be likely to emerge through a memory test to observe whether stereotyped statements have a systematic and replicable effect on performance. From previous research, we expected to see performances in line with threat (i.e., suppressed performance) and lift (i.e., improved performance). When all AMATs were analyzed together, our results showed improved scores across all groups but the control group (H) from AMATs 1-3 to 4-6. The two lowest scores obtained across participants in all groups, however, occurred in the two threat conditions (AMATs 4-6 for group I and AMATs 1-3 for group K). As was noted in the pilot studies, the first test tended to produce the lowest performance scores for most participants. Given this, we also analyzed data comparing AMATs 3 to 4 and 2-3 to 4-5. While the control group still displayed decreases in average scores over time, group I now showed higher performance scores in AMAT 3 and 2-3 compared to AMAT 4 and 4-5, or the neutral conditions compared to the threat conditions. Additionally, group L averages showed a decrease in score from AMAT 3-4 from the lift condition to the neutral condition (i.e., 83.1% - 80.6%), although this pattern reversed slightly when averages from AMATs 2-3 and 4-5 were analyzed (i.e., 78.6% - 80.4%).

Although the average data across all AMATs showed that participants in all groups except the control group generally displayed higher performance scores in AMATs 4-6 than 1-3, analyses isolating AMATs 2-3 and 4-5 as well as 3 and 4 showed patterns more consistent with stereotype threat and stereotype lift research. These patterns were more consistent and produced larger changes in performance across participants in groups where threat scripts and statements were present (i.e., groups I and K), as evidenced by overall group averages in performance scores for AMATs 3 and 4, AMATs 2-3 and 4-5, and the total number of participants who performed with lower accuracy on the tests where threat statements were present (70% of participants versus 60% of participants who received lift statements performing better in lift conditions). Previous pilot 1-3 research displayed inconsistent group results, wherein some threat and lift conditions saw higher or lower performance scores depending on when the threat and lift statements were introduced. As well, the switching back and forth between statements created situations wherein it was much more difficult to analyze how the gendered statement impacted the participant. Thus, the methods utilized in the current research did lend better to a more clear analysis of impact.

Results from repeated measures ANOVAs as well as a mixed model analysis of variance ANOVA did not yield statistically significant results when AMAT 2-5 data were analyzed. However, when all AMAT data was analyzed, results for Groups I, K, and L were statistically significant when examined based on an alpha of .05. These results only indicated within effects when the first AMAT was included for analysis with groups I and K. Although the statistical significance of our data was very limited to a particular context (i.e., AMAT 1 data needed to be included), inducing a gendered stereotype in a novel context in which gendered stereotypes were not likely to be relevant still produced overall group averages and trends that did support the typical patterns that have been well discussed in stereotype threat and lift research. That is, participants were more accurate or received higher performance scores in lift conditions compared to the neutral conditions (i.e., groups J and L), and were less accurate or received lower performance scores in the threat conditions compared to the neutral conditions (i.e., groups I and K).

With respect to whether gender affiliation impacted performance across groups, our data indicate that males and females tended to perform very similarly across each experimental group. For our threat groups: 60% of females and 60% of males (group I) and 80% of females and 80% of males (group K) showed a decrease in performance scores in the threat condition compared to the neutral condition. For the lift groups: 60% of females and 60% of males (groups J and L) showed an increase in performance scores in the lift condition compared to the neutral condition. Thus, no differences in performance across males and females was observed in the current research.

Some variables we examined across the pilot studies and in the current research that might impact performance scores were confidence and time spent studying. In general, time spent studying decreased over time for groups K and L moving into the neutral conditions (70% of participants), while groups I and J saw more equal distributions of participant studying for longer/shorter between AMATs 3 and 4 (50%/50%). Thus, participants studied longer when they moved from the neutral condition into a threat or lift condition, as the tendency in the control group and groups K and L was for participants to study less over time. As for confidence, overall average confidence was higher in the neutral conditions than the threat and lift conditions for groups I and L, higher in the lift condition than the neutral condition for group J, and higher in the threat condition for group K. Differences across confidence averages, however tended to be very small.

To our knowledge, a repeated measures mixed group design across 6 separate tests has not been utilized in previous stereotype threat research. This mixed methods design enabled a more nuanced evaluation of the effects of stereotyped statements on performance. That is, we were able to analyze changes at both the group and individual level. Despite non statistically significant results across groups and AMATs 2-6, individual data across participants showed a majority of participants showing higher accuracy in lift conditions and lower accuracy in threat conditions. These mixed method enabled us to better assess patterns, trends, and impact of the stereotypic statements.

Further, the repeated measures design allowed for analysis of patterns in performance across time, under the different stereotype statement conditions. One pattern we noticed as a result of the repeated measures design was that some participants would show an initial reaction to the change in statement, and then their scores would improve thereafter. It is possible, then, that the scripts and statements were most impactful upon the first exposure, after which participant were occasionally able to recover from the effect of the threat. Although more data is needed to analyze this pattern, the repeated measures design can offer a look into how threat and lift statements can impact participants over time and potentially lead to new research questions in stereotype threat/lift research.

Lastly, by using a novel task that was not associated with specific gender stereotypes, we were able to distill out the potential impacts of stereotyped statements within an RFT framework. Although the current research did not implement a task with which a gendered stereotype necessarily exists, our aim was to see if a gendered statement could evoke changes in behavior on a more novel task by utilizing similar language often used to evoke stereotypic patterns. Our results show changes in our overall group averages and a majority of individual participants' responses that suggest that the statement itself could have an impact on participants responding to a novel task. This is consistent with the RFT framework: suggestions of gender differences, originally paired with common stereotypes, are likely to have evoked behavior changes for the individual previously. When those statements were applied to the novel task, it seems that the functions of those stereotyped statements also were brought to bear on the participants' behaviors. Although these statements occurred in a more novel context than stereotypes have typically been evoked, the suggestion itself can evoke similar responding in different contexts (i.e., combinatorial entailment). As a result, the statements alter the function of the test questions and memory sets, such that the participant might have a different frame of engagement with the task, and might study more or less, report more or less confidence, etc. (i.e., transformation of stimulus function). Such research can provide a starting point for a more comprehensive analysis of stereotype threat and lift from an RFT framework.

Although the standard stereotypes and tasks for which stereotype threat and lift tend to occur in the literature (e.g., during difficult math assessments) were not utilized in the current research, by manipulating general "stereotypic" statements to see if these would induce changes in behavior during a novel task, we were able to examine further conditions or situations in which stereotype threat and lift might occur. The pervasiveness of stereotypes that have been shown to impact behavior in particular situations (e.g., females taking math tests) can, as is discussed in RFT, relate to and impact behavior in other areas/situations in which stereotypes could be relevant. By utilizing both individual and group data analyses, we were able to detect overall changes across participants while also checking for consistencies and inconsistencies in the data through individual participants' data. Importantly, individual data allowed us to analyze contexts in which typical threat/lift patterns were observed or not. In general, two patterns emerged from this data. Either participants reported more "stress" or "pressure" to do well in threat conditions, or that they "tried harder" to "prove the statement wrong." Both reactions are consistent with previous research: however, this research has generally shown participants consistently engaging in one pattern or the other, rather than both within the same research arrangement.

Limitations and Future Research

Given the novelty of the current research, a number of limitations exist. First and foremost, the sample size obtained in the current research and previous pilots created a barrier to statistical analyses. Low numbers of participants (10 in each group) limited the generalizability of the data to similar experiments. Future research should recruit larger sample sizes to fully analyze stereotypic statement impact in less formal or novel contexts in which stereotypes could be relevant.

As was evident in the current research and the pilot studies, there were issues with being able to analyze data from the first MT/AMAT, given that the was the first introduction participants had to how the tests would operate. In order to be able to use data from the first tests, especially in a repeated measures design, future research should examine the use of a short practice test administered prior to testing, as has been utilized in other research (e.g., Vandenberg & Kuse, 1978), more explicit instructions participants would need to engage with (e.g., time limit on how long they have to read instructions), or other program modifications that enable participants to enter the testing situation more prepared to take the memory tests.

Although the current research does not formally meet the standards of current stereotype threat and lift research, future research should examine conditions under which these criteria can be met. Our data did show trends in the directions consistent with stereotype threat and lift research in our group averages and individual participant responses, which is a promising first step toward threat and lift research. However, trends in the data were not statistically significant, which creates a barrier to the applicability of our results. Larger sample sizes are critical to any replications/extensions of the current research and future research. Given our current data trends across groups and individuals, it is likely that, with larger sample sizes, we will be able to see patterns more consistent with stereotype threat and lift displayed in a more novel context.

Another limitation was the testing or practice effect consistently encountered in pilots 1-3 and the current experimental arrangement. The first memory test (MT 1/AMAT 1) tended to produce the lowest score across participants in all groups. This made analyses of statement impact in MT1 or AMAT 1 difficult to interpret. Future research should address how this problem can be mitigated, either through more brief practice tests (as observed in VMRT administration) or by reviewing the expectations and how to take the tests with the participant ahead of testing.

In addition, the low rates of participants who were able to participate in the study via Examus, warranting the addition of the Zoom session version of the study, also presented a limitation, as not all participants participated in the same manner. Although we attempted to arrange testing in Zoom to be as similar to the Examus sessions as possible, differences occurred that could have differentially impacted participants. To ensure testing was similar across groups, we compared data across Zoom participants versus Examus participants for each group. Participants in each group tended to do better on some tests and worse on others across most groups, indicating there was variability in responding across groups and that one group didn't tend to outperform the other. Group J tended to show higher performance scores across tests with the zoom group, but there were also fewer participants in the Examus condition, meaning that one person's lower score would have been weighted more heavily. Future research should endeavor to ensure that all participants access the experiment in the same manner in order to reduce the possible confounds associated with different testing platforms.

It is also important to note that the current research was conducted during the pandemic, which meant changes to our original planned methods were warranted to allow participants to participate online. Although this could have presented a limitation to the current research, it is also worth noting that, even amongst an entirely online participant pool, we still saw results consistent with stereotype threat and lift. Thus, stereotype threat may not just be an in-person, classroom-based phenomenon wherein an experimenter must physically be present with participants (as argued by some researchers, e.g., Pennington et al., 2016) in order to see effects.

Future research should examine if similar patterns, as have been shown in stereotype threat and lift research, can be observed in other novel tasks. While known stereotypes about performance differences tend to explain what we can formally call stereotype threat or stereotype lift, as per their respective definitions, it is possible that such effects could be replicated in other circumstances not yet discussed in the extant research. As stereotypes are prolific in everyday language and practices, it is possible that just statements that one group will perform differently than another could result in changes in performance and behavior. If stereotypic statements can impact behavior under conditions which have yet to be explored, as can be analyzed from an RFT framework, this would show the extent to which such statements can impact behavior.

In order to provide a more comprehensive analysis of how the RFT framework operates in stereotype threat and lift, future research should examine ways in which this can better occur. Although the current research utilized multiple quantitative metrics (confidence ratings, time spent studying/taking each test, number of times questions were revisited), and one qualitative metric (i.e., open-ended answer to statement impact) to measure statement impact, a more comprehensive analysis of each participants verbal behavior with respect to testing would enable a better analysis at the individual level. Thirteen participants reported not being affected by the gendered statements during testing, but clearly showed an impact in how long they studied and their overall confidence in their answers. Others reported an impact different from what their data showed (e.g., performing “better” but reporting they thought they performed “worse”). As the opportunity to discuss their interaction with these statements was removed from the situation itself, and this was one of the last questions the participants were asked, participants were much more likely to proceed quickly through these questions and less likely to elaborate on their process during testing. Implementing a “think aloud” procedure, wherein participants are instructed to read and say everything out loud that they are seeing/thinking during testing (Guss, 2018) might be instructive in future research aiming to analyze this impact more cohesively.

Despite barriers and limitations to the current study, behavior analysts would benefit from expanding our knowledge of stereotyped language and its impact on individuals and society. We often focus on the impact we can make for each individual client or student and can often fail to account for influences that culture presents and that we encounter or even create, whether we are aware of it or not. Fred Keller (1968) famously argued that “*the student is always right.* He is not asleep, not unmotivated, not sick, and he can learn a great deal if we provide the right contingencies of reinforcement” (p. 88, italics original). We would do well in acknowledging that, to truly account for what is meant by “the student is always right,” we must also consider that people frequently contact gendered, racialized, or other situations that impact their interactions with people and situations.

References

Alter, A. L., Aronson, J., Darley, J. M., Rodriguez, C., & Ruble, D. N. (2010). Rising to the threat: Reducing stereotype threat by reframing the threat as a challenge. *Journal of Experimental Social Psychology*, *46*(1), 166–171. https://doi.org/10.1016/j.jesp.2009.09.014

Angel, D. (2014, April 21). *Neil Degrasse Tyson on being Black and Women in Science*. [Video]. You tube. https://youtu.be/z7ihNLEDiuM

Armenta, B. E. (2010). Stereotype boost and stereotype threat effects: The moderating role of ethnic identification. *Cultural Diversity and Ethnic Minority Psychology*, *16*(1), 94–98. https://doi.org/10.1037/a0017564

Aronson, J., Lustina, M. J., Good, C., Keough, K., Steele, C. M., & Brown, J. (1999). When white men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology*, *35*(1), 29–46. https://doi.org/10.1006/jesp.1998.1371

Benbow, C. P., & Stanley, J. C. (1980). Sex differences in mathematical ability: Fact or artifact? *Science*, *210*(4475), 1262–1264. https://doi.org/10.1126/science.7434028

Ben-Zeev, T., Fein, S., & Inzlicht, M. (2005). Arousal and stereotype threat. *Journal of Experimental Social Psychology*, *41*(2), 174–181. https://doi.org/10.1016/j.jesp.2003.11.007

Burnette, J. L., Pollack, J. M., & Hoyt, C. L. (2010). Individual differences in implicit theories of leadership ability and self-efficacy: Predicting responses to stereotype threat. *Journal of Leadership Studies*, *3*(4), 46–56. https://doi.org/10.1002/jls.20138

Cadinu, M., Maass, A., Frigerio, S., Impagliazzo, L., & Latinotti, S. (2003). Stereotype threat: The effect of expectancy on performance. *European Journal of Social Psychology*, *33*(2), 267–285. https://doi.org/10.1002/ejsp.145

Chateignier, C., Dutrévis, M., Nugier, A., & Chekroun, P. (2009). French-Arab students and verbal intellectual performance: Do they really suffer from a negative intellectual stereotype? *European Journal of Psychology of Education*, *24*(2), 219–234. https://doi.org/10.1007/BF03173013

Cole, N. (1997). *The ETS gender study: How females and males perform in educational settings*. Princeton, NJ. Educational Testing Service.

Crisp, R. J., Bache, L. M., & Maitner, A. T. (2009). Dynamics of social comparison in counter-stereotypic domains: Stereotype boost, not stereotype threat, for women engineering majors. *Social Influence*, *4*(3), 171–184. https://doi.org/10.1080/15534510802607953

Crowley, K., Callanan, M. A., Tenenbaum, H. R., & Allen, E. (2001). Parents explain more often to boys than to girls during shared scientific thinking. *Psychological Science*, *12*(3), 258–261. https://doi.org/10.1111/1467-9280.00347

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. https://doi.org/10.1177/014616702237644

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition* (1st ed.). Sage Publications, Inc.

Eccles, J. S., & Jacobs, J. E. (1986). Social forces shape math attitudes and performance. *Signs: Journal of Women in Culture and Society*, *11*(2), 367–380. https://doi.org/10.1086/494229

Eriksson, K., & Lindholm, T. (2007). Making gender matter: The role of gender-based expectancies and gender identification on women's and men's math performance in Sweden. *Scandinavian Journal of Psychology*, *48*(4), 329–338. https://doi.org/10.1111/j.1467-9450.2007.00588.x

Errasti, J., Martinez, H., Rodriguez, C., Marquez, J., Maldonado, A., & Menendez, A. (2019). Social context in a collective IRAP application about gender stereotypes: Mixed versus single gender groups. *Psychological Record*, *69*(1), 39–48. https://doi.org/10.1007/s40732-018-0320-1

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist*, *43*(2), 95–103. https://doi.org/10.1037/0003-066X.43.2.95

Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of School Psychology*, *53*(1), 25–44. https://doi.org/10.1016/j.jsp.2014.10.002

Forbes, C. E., & Schmader, T. (2010). Retraining Attitudes and Stereotypes to Affect Motivation and Cognitive Capacity Under Stereotype Threat. *Journal of Personality and Social Psychology*, *99*(5), 740–754. https://doi.org/10.1037/a0020971

Frome, P. M., & Eccles, J. S. (1998). Parents' influence on children's achievement-related perceptions. *Journal of Personality and Social Psychology*, *74*(2), 435–452. https://doi.org/10.1037/0022-3514.74.2.435

Gonzales, P. M., Blanton, H., & Williams, K. J. (2002). The effects of stereotype threat and double-minority status on the test performance of Latino women. *Personality and Social Psychology Bulletin*, *28*(5), 659–670. https://doi.org/10.1177/0146167202288010

Guillot, A., Champely, S., Batier, C., Thiriet, P., & Collet, C. (2007). Relationship between spatial abilities, mental rotation and functional anatomy learning. *Advances in Health Sciences Education*, *12*(4), 491–507. https://doi.org/10.1007/s10459-006-9021-7

Guss, C. (2018). What is going through your mind? Thinking aloud as a method in cross-cultural psychology. *Frontiers in Psychology, 9* (1292), 1-11. https://doi.org/10.3389/fpsyg.2018.01292

Hayes, S. C. (2016). Acceptance and commitment therapy, relational frame theory, and the third wave of behavioral and cognitive therapies – Republished article. *Behavior Therapy*, *47*(6), 869–885. https://doi.org/10.1016/j.beth.2016.11.006

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational frame theory: A post-Skinnerian account of human language*. Plenum Publishers.

Hess, T. M., Emery, L., & Queen, T. L. (2009a). Task demands moderate stereotype threat effects on memory performance. *Journals of Gerontology - Series B Psychological Sciences and Social Sciences*, *64*(4), 482–486. https://doi.org/10.1093/geronb/gbp044

Hess, T. M., Emery, L., & Queen, T. L. (2009b). Task demands moderate stereotype threat effects on memory performance. *Journals of Gerontology - Series B*

Psychological Sciences and Social Sciences, *64*(4), 482–486. https://doi.org/10.1093/geronb/gbp044

Hess, T. M., Hinson, J. T., & Hodges, E. A. (2009a). Moderators of and mechanisms underlying stereotype threat effects on older adults' memory performance. *Experimental Aging Research*, *35*(2), 153–177. https://doi.org/10.1080/03610730802716413

Hess, T. M., Hinson, J. T., & Hodges, E. A. (2009b). Moderators of and mechanisms underlying stereotype threat effects on older adults' memory performance. *Experimental Aging Research*, *35*(2), 153–177. https://doi.org/10.1080/03610730802716413

Horton, S., Baker, J., Pearce, W., & Deakin, J. M. (2010). Immunity to popular stereotypes of aging? Seniors and stereotype threat. *Educational Gerontology*, *36*(5), 353–371. https://doi.org/10.1080/03601270903323976

Hoyt, C. L., Johnson, S. K., Murphy, S. E., & Skinnell, K. H. (2010). The impact of blatant stereotype activation and group sex-composition on female leaders. *Leadership Quarterly*, *21*(5), 716–732. https://doi.org/10.1016/j.leaqua.2010.07.003

Huguet, P., & Régner, I. (2009). Counter-stereotypic beliefs in math do not protect school girls from stereotype threat. *Journal of Experimental Social Psychology*, *45*(4), 1024–1027. https://doi.org/10.1016/j.jesp.2009.04.029

Humphreys, L. G. (1975). "Educational uses of tests with disadvantaged students": Addendum. *American Psychologist*, *30*(1), 95–96. https://doi.org/10.1037/0003-066X.30.1.95

Hyde, J. S., Fennema, E., Ryan, M., Frost, L. A., & Hopp, C. (1990). Gender comparisons of mathematics attitudes and affect: A meta-analysis. *Psychology of Women Quarterly*, *14*(3), 299–324. https://doi.org/10.1111/j.1471-6402.1990.tb00022.x

Inzlicht, M., & Ben-Zeev, T. (2000). A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science*, *11*(5), 365–371. https://doi.org/10.1111/1467-9280.00272

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. https://doi.org/10.1037/0022-3514.63.6.932

Jensen, A. R. (1980). *Bias in mental testing*. Free Press.

Jones, C. M., & Healy, S. D. (2006). Differences in cue use and spatial memory in men and women. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1598), 2241–2247. https://doi.org/10.1098/rspb.2006.3572

Keller, J. (2002). Blatant stereotype threat and women's math performance: Self-handicapping as a strategic means to cope with obtrusive negative performance expectations. *Sex Roles*, *47*(3–4), 193–198. https://doi.org/https://doi.org/10.1023/A:1021003307511

Keller, J., & Dauenheimer, D. (2003). Stereotype threat in the classroom: Dejection mediates the disrupting threat effect on women's math performance. *Personality and Social Psychology Bulletin*, *29*(3), 371–381. https://doi.org/10.1177/0146167202250218

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*, *119*(1), 80–113. https://doi.org/10.1037/a0025648

Lawrence, J. S., Marks, B. T., & Jackson, J. S. (2010). Domain identification predicts black students' underperformance on moderately-difficult tests. *Motivation and Emotion*, *34*(2), 105–109. https://doi.org/10.1007/s11031-010-9159-8

Levy, B. (1996). Improving memory in old age through implicit self-stereotyping. *Journal of Personality and Social Psychology*, *71*(6), 1092–1107. https://doi.org/10.1037/0022-3514.71.6.1092

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research*, *43*(2), 139–161. https://doi.org/10.2307/1169933

Logel, C., Iserman, E. C., Davies, P. G., Quinn, D. M., & Spencer, S. J. (2009). The perils of double consciousness: The role of thought suppression in stereotype threat. *Journal of Experimental Social Psychology*, *45*(2), 299–312. https://doi.org/10.1016/j.jesp.2008.07.016

Martinez, A., & Christnacht, C. (2021, January 26). *Women are nearly half of U.S. workforce but only 27% of STEM workers*. United States Census Bureau.

Marx, D. M., Stapel, D. A., & Muller, D. (2005). We can do it: The interplay of construal orientation and social comparisons under threat. *Journal of Personality and Social Psychology*, *88*(3), 432–446. https://doi.org/10.1037/0022-3514.88.3.432

McGlone, M. S., & Aronson, J. (2006). Stereotype threat, identity salience, and spatial reasoning. *Journal of Applied Developmental Psychology*, *27*(5), 486–493. https://doi.org/10.1016/j.appdev.2006.06.003

Mendes, W. B., Blascovich, J., Lickel, B., & Hunter, S. (2002). Challenge and threat during social interactions with white and black men. *Personality and Social Psychology Bulletin*, *28*(7), 939–952. https://doi.org/10.1177/014616720202800707

Nadler, J. T., & Clark, M. H. (2011). Stereotype threat: A meta-analysis comparing African Americans to Hispanic Americans. *Journal of Applied Social Psychology*, *41*(4), 872–890. https://doi.org/10.1111/j.1559-1816.2011.00739.x

Naphan-Kingery, D., & Kemmelmeier, M. (2018). Gender identification moderates the effects of stereotype threat: A replication and extension. Unpublished manuscript.

Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, *93*(6), 1314–1334. https://doi.org/10.1037/a0012702

O’Brien, L. T., & Crandall, C. S. (2003a). Stereotype threat and arousal: Effects on women’s math performance. *Personality and Social Psychology Bulletin*, *29*(6), 782–789. https://doi.org/10.1177/0146167203029006010

O’Brien, L. T., & Crandall, C. S. (2003b). Stereotype threat and arousal: Effects on women’s math performance. *Personality and Social Psychology Bulletin*, *29*(6), 782–789. https://doi.org/10.1177/0146167203029006010

Osborne, J. W. (2001). Testing stereotype threat: Does anxiety explain race and sex differences in achievement? *Contemporary Educational Psychology*, *26*(3), 291–310. https://doi.org/10.1006/ceps.2000.1052

Pennington, C. R., Heim, D., Levy, A. R., & Larkin, D. T. (2016). Twenty years of stereotype threat research: A review of psychological mediators. *PLoS ONE*, *11*(1). https://doi.org/10.1371/journal.pone.0146487

Pennington, C. R., Litchfield, D., McLatchie, N., & Heim, D. (2019). Stereotype threat may not impact women's inhibitory control or mathematical performance: Providing support for the null hypothesis. *European Journal of Social Psychology*, *49*(4), 717–734. https://doi.org/10.1002/ejsp.2540

Picho, K., & Schmader, T. (2018). When do gender stereotypes impair math performance? A study of stereotype threat among Ugandan adolescents. *Sex Roles*, *78*(3–4), 295–306. https://doi.org/10.1007/s11199-017-0780-9

Pronin, E., Steele, C. M., & Ross, L. (2004). Identity bifurcation in response to stereotype threat: Women and mathematics. *Journal of Experimental Social Psychology*, *40*(2), 152–168. https://doi.org/10.1016/S0022-1031(03)00088-X

Purdie-Vaughns, V., Steele, C. M., Davies, P. G., Ditlmann, R., & Crosby, J. R. (2008). Social identity contingencies: How diversity cues signal threat or safety for African Americans in mainstream institutions. *Journal of Personality and Social Psychology*, *94*(4), 615–630. https://doi.org/10.1037/0022-3514.94.4.615

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*(1), 55–71. https://doi.org/10.1111/0022-4537.00201

Rahhal, T. A., Hasher, L., & Colcombe, S. J. (2001). Instructional manipulations and age differences in memory: Now you see them, now you don't. *Psychology and Aging*, *16*(4), 697–706. https://doi.org/10.1037/0882-7974.16.4.697

Rydell, R. J., McConnell, A. R., & Beilock, S. L. (2009). Multiple social identities and stereotype threat: Imbalance, accessibility, and working memory. *Journal of*

Personality and Social Psychology, *96*(5), 949–966. https://doi.org/10.1037/a0014846

Schmader, T. (2002). Gender identification moderates stereotype threat effects on women's math performance. *Journal of Experimental Social Psychology*, *38*(2), 194–201. https://doi.org/10.1006/jesp.2001.1500

Schmader, T., Forbes, C. E., Shen Zhang, & Berry Mendes, W. (2009). A metacognitive perspective on the cognitive deficits experienced in intellectually threatening environments. *Personality and Social Psychology Bulletin*, *35*(5), 584–596. https://doi.org/10.1177/0146167208330450

Seibt, B., & Förster, J. (2004). Stereotype threat and performance: How self-stereotypes influence processing by inducing regulatory foci. In *Journal of Personality and Social Psychology* (Vol. 87, Issue 1, pp. 38–56). https://doi.org/10.1037/0022-3514.87.1.38

Sekaquaptewa, D., & Thompson, M. (2002). The differential effects of solo status on members of high- and low-status groups. *Personality and Social Psychology Bulletin*, *28*(5), 694–707. https://doi.org/10.1177/0146167202288013

Sekaquaptewa, D., & Thompson, M. (2003). Solo status, stereotype threat, and performance expectancies: Their effects on women's performance. *Journal of Experimental Social Psychology*, *39*(1), 68–74. https://doi.org/10.1016/S0022-1031(02)00508-5

Shewach, O. R., Sackett, P. R., & Quint, S. (2019). Stereotype threat effects in settings with features likely versus unlikely in operational test settings: A meta-analysis.

Journal of Applied Psychology, *104*(12), 1514–1534. https://doi.org/10.1037/apl0000420

Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. matching to sample: An expansion of the testing paradigm. *Journal of the Experimental Analysis of Behavior*, *37*(1), 5–22. https://doi.org/10.1901/jeab.1982.37-5

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology*, *67*, 415–437. https://doi.org/10.1146/annurev-psych-073115-103235

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28. https://doi.org/10.1006/jesp.1998.1373

Stanley, J. C. (1971). Predicting college success of the educationally disadvantaged. *Science*, *171*(3972), 640–647. https://doi.org/10.1126/science.171.3972.640

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*(6), 613–629. https://doi.org/10.1037/0003-066X.52.6.613

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797–811. https://doi.org/10.1037/0022-3514.69.5.797

Stone, J. (2002). Battling doubt by avoiding practice: The effects of stereotype threat on self-handicapping in White athletes. *Personality and Social Psychology Bulletin*, *28*(12), 1667–1678. https://doi.org/10.1177/014616702237648

Stone, J., & McWhinnie, C. (2008). Evidence that blatant versus subtle stereotype threat cues impact performance through dual processes. *Journal of Experimental Social Psychology*, *44*(2), 445–452. https://doi.org/10.1016/j.jesp.2007.02.006

Taylor, V. J., & Walton, G. M. (2011). Stereotype threat undermines academic learning. *Personality and Social Psychology Bulletin*, *37*(8), 1055–1067. https://doi.org/10.1177/0146167211406506

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, *47*(2), 599–604. https://doi.org/10.2466/pms.1978.47.2.599

von Hippel, C., Issa, M., Ma, R., & Stokes, A. (2011). Stereotype threat: Antecedents and consequences for working women. *European Journal of Social Psychology*, *41*(2), 151–161. https://doi.org/10.1002/ejsp.749

Walton, G. M., & Cohen, G. L. (2003). Stereotype lift. *Journal of Experimental Social Psychology*, *39*(5), 456–467. https://doi.org/10.1016/S0022-1031(03)00019-2

Walton, G. M., & Spencer, S. J. (2009). Latent ability. *Psychological Science*, *20*(9), 1132–1139. https://doi.org/10.1111/j.1467-9280.2009.02417.x

Walton, G. M., Spencer, S. J., & Erman, S. (2013). Affirmative meritocracy. *Social Issues and Policy Review*, *7*(1), 1–35. https://doi.org/10.1111/j.1751-2409.2012.01041.x

Zajonc, R. B. (1965). Social facilitation. *Science*, *149*(3681), 269–274. https://doi.org/10.1126/science.149.3681.269

Appendix A

Survey Questions from Exit Survey

Thank you for completing the MRT and memory tests! We appreciate your participation in this research and request that you answer 4 final questions about the memory strategies you used during testing and any other experiences you had. A response is required to this survey to be entered into the gift card drawing. This should take between 5-8 minutes of your time.

1. Please enter your name (first and last):

The first 4 questions will refer to the Mental Rotation Test (MRT):

2. For the MRT, what were some strategies you used to solve each question? (check all that apply)
    a. I imagined myself being stationary and the whole object rotating with respect to me.
    b. I first imagined rotating a part of the object and then checked whether the rest of the object could be rotated in the same way to match the target
    c. I imagined the objects being stationary as I moved around them to view them from different perspectives
    d. I imagined that the object was an animal (e.g., a snake) and where its head or arms or tail would be
    e. I counted the number of cubes in the four straight segments of the object
    f. I counted the number of cubes in just the two end segments of the object
    g. I examined the directions of the four segments of the object with respect to each other
    h. I examined the directions of the two end segments with respect to each other
    i. None of the above
3. Were there other memory techniques you used that weren't listed above? If so, please list below:
4. How much effort (i.e., use of strategies) would you say you put into each question on the MRT?
    a. I put effort into my answers the entire time and guessed on very few questions
    b. I mostly put effort into my answers, but started guessing when time was getting close to running out
    c. I guessed and put effort into my answers at about equal proportions
    d. I guessed a good deal of the time, but put effort into a few questions
    e. I guessed most of the time
5. How confident are you in your overall performance on the MRT?
    a. I think I got 95-100% of the questions I attempted correct
    b. I think I got 85-95% of the questions I attempted correct
    c. I think I got 70-85% of the questions I attempted correct
    d. I think I got 50-70% of the questions I attempted correct

   e. I think I got 35-50% of the questions I attempted correct
   f. I think I got fewer than 35% of the questions I attempted correct
6. Prior to your participation in this study, have you taken some form of a mental rotation test before?
   a. Yes
   b. No
   c. I don't remember

The following questions refer to your experiences during the Memory Tests:

7. What were some strategies you used to answer any memory test questions? (select all that apply): Were there other memory techniques you used that weren't listed above?
   a. I attempted to memorize each individual set
   b. I used acronyms to remember sets (e.g., RBT for "Rip, Black, and T")
   c. I created images in my head of information from the sets (e.g., I pictured a "purple bun-nY" to help remember "purple, bun, and Y" together
   d. I used mnemonic devices to remember sets (e.g., "16 ripe oranges were left" to help remember "16, Rex, Orange, and L"
   e. I related sets to my own experiences (e.g., "my favorite coffee place HUB is by the Blue river")
   f. I memorized some and used the process of elimination for others (e.g., "I know X is not correct, so it must be Y")
   g. None of the above
8. How much effort (i.e., use of strategies) would you say you put into each question on the Memory Tests?
   a. I put effort into my answers the entire time and guessed on very few questions
   b. I mostly put effort into my answers, but started guessing when time was getting close to running out
   c. I guessed and put effort into my answers at about equal proportions
   d. I guessed a good deal of the time, but put effort into a few questions
   e. I guessed most of the time
9. Were there other techniques you used that weren't listed above?
10. How many memory tests do you feel confident you performed well on overall?
    a. All 5 memory tests
    b. 4 of the memory tests
    c. 3 of the memory tests
    d. 2 of the memory tests
    e. 1 of the memory tests
    f. None of the memory tests
11. Which statements did you see during testing? (select all that apply)
    a. "There were no performance differences between groups on this test"
    b. "Women tend to perform better than men on this test"
    c. "Men tend to perform better than women on this test"
12. (Display logic: if the participant selected 11c, they received the following): "How did reading "men tend to do better than women" impact your performance during testing?

   a. I thought I would perform better on those tests
   b. I thought I would perform the same as on other tests
   c. I thought I would perform worse on those tests
   d. Other: (fill in)
13. (Display logic: if the participant selected 11b, they received the following): “How did reading "women tend to do better than men" impact your performance during testing?
    a. I thought I would perform better on those tests
    b. I thought I would perform the same as on other tests
    c. I thought I would perform worse on those tests
    d. Other: (fill in)
14. Given the statements above, how did these impact your thoughts/performance during testing?

The remaining questions refer to your current educational history and whether you want to opt into the gift card raffle:

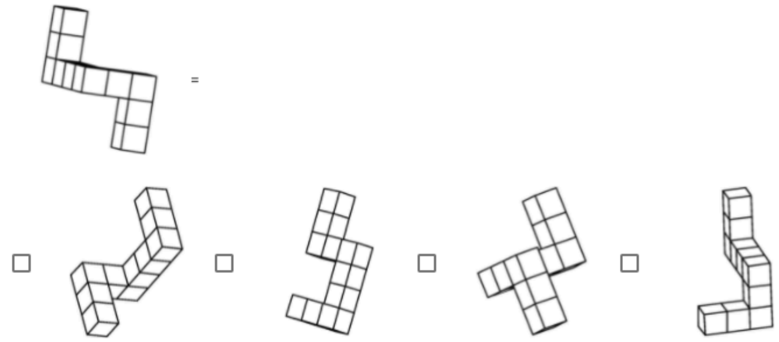
15. Given the number of credits you have accumulated thus far, what year are you in college?
    a. Freshmen
    b. Sophomore
    c. Junio
    d. Senior
    e. Graduate Special
    f. Graduate Student
16. What is your current major and minor?
17. Please check all that apply:
    a. I have taken and/or am taking at least one Women's Studies Course
    b. I have taken and/or am taking Psychology of Gender
    c. I have taken and/or am taking Psychology 101
    d. None of the above
18. Do you have any further questions?
19. Would you like to be entered into the gift card drawing? If so, enter an email where we can contact you to arrange gift card distribution if your name is drawn.
    a. Yes (make sure your email address is entered correctly:
    b. No
20. As we are still collecting data, results from the study are not yet available. However, at the completion of this study, we are happy to share our results with any participants who wish to receive an update. Would you prefer we send you an email update regarding the results of this study? If so, enter the email address you want it to be sent to.
    a. Yes (make sure your email address is entered correctly:)
    b. No

Appendix B

Figure B1

Mental Rotation Test Samples

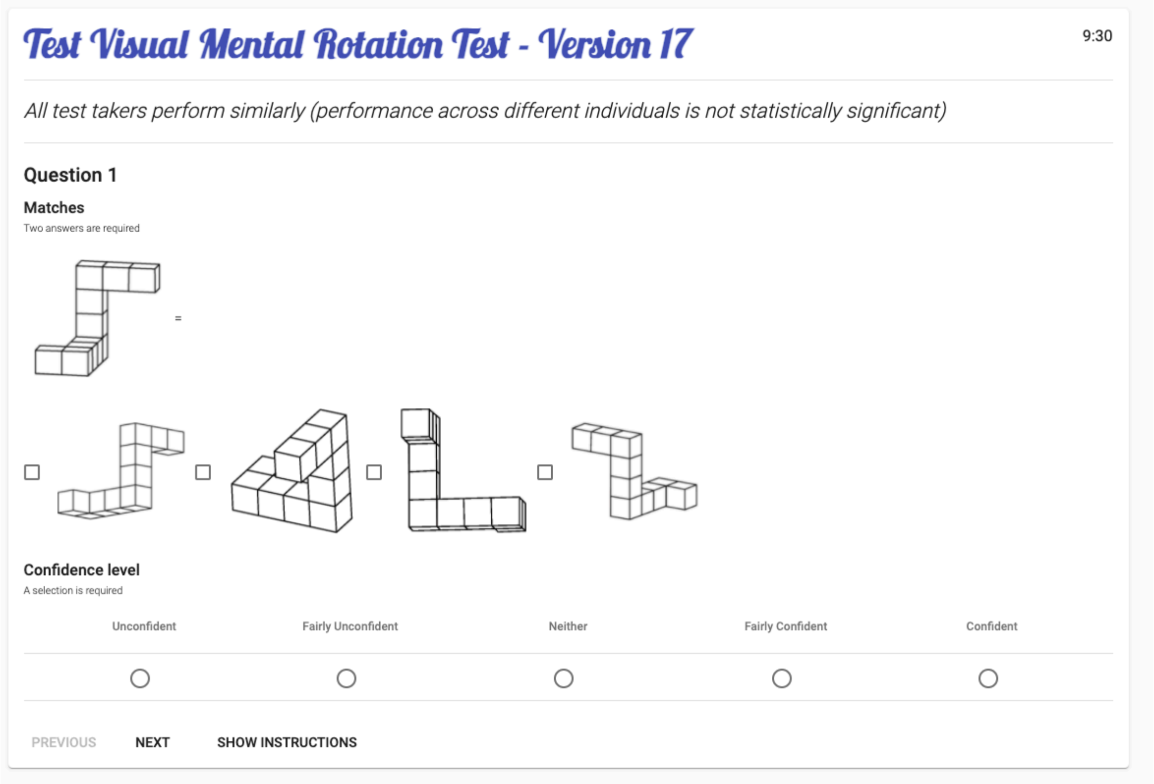
MRT – Computer Drawn figures in stimulus test program



MRT Question Sample

Figure B2

Mental Rotation Test Samples



Test Visual Mental Rotation Test - Version 17

9:30

All test takers perform similarly (performance across different individuals is not statistically significant)

Question 1

Matches

Two answers are required

=

□ □ □ □

Confidence level

A selection is required

Unconfident	Fairly Unconfident	Neither	Fairly Confident	Confident
○	○	○	○	○

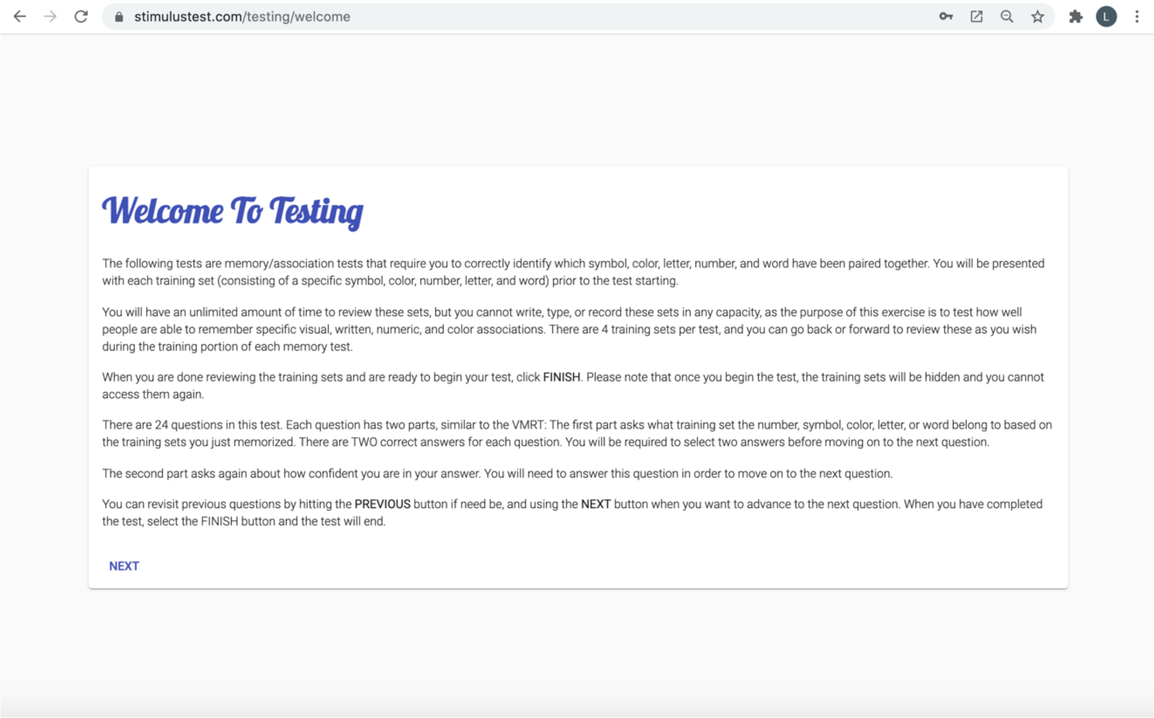
PREVIOUS NEXT SHOW INSTRUCTIONS

Appendix C

Figure C1

Memory Test (MT) Samples

Test Instructions (both MT's and MRT)



Welcome To Testing

The following tests are memory/association tests that require you to correctly identify which symbol, color, letter, number, and word have been paired together. You will be presented with each training set (consisting of a specific symbol, color, number, letter, and word) prior to the test starting.

You will have an unlimited amount of time to review these sets, but you cannot write, type, or record these sets in any capacity, as the purpose of this exercise is to test how well people are able to remember specific visual, written, numeric, and color associations. There are 4 training sets per test, and you can go back or forward to review these as you wish during the training portion of each memory test.

When you are done reviewing the training sets and are ready to begin your test, click **FINISH**. Please note that once you begin the test, the training sets will be hidden and you cannot access them again.

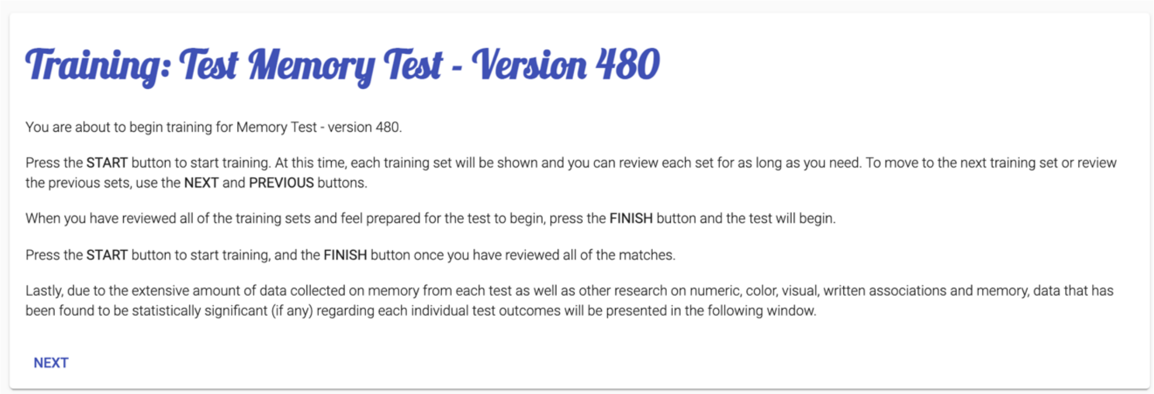
There are 24 questions in this test. Each question has two parts, similar to the VMRT: The first part asks what training set the number, symbol, color, letter, or word belong to based on the training sets you just memorized. There are TWO correct answers for each question. You will be required to select two answers before moving on to the next question.

The second part asks again about how confident you are in your answer. You will need to answer this question in order to move on to the next question.

You can revisit previous questions by hitting the **PREVIOUS** button if need be, and using the **NEXT** button when you want to advance to the next question. When you have completed the test, select the FINISH button and the test will end.

Figure C2

Memory Test (MT) Samples



Training: Test Memory Test - Version 480

You are about to begin training for Memory Test - version 480.

Press the **START** button to start training. At this time, each training set will be shown and you can review each set for as long as you need. To move to the next training set or review the previous sets, use the **NEXT** and **PREVIOUS** buttons.

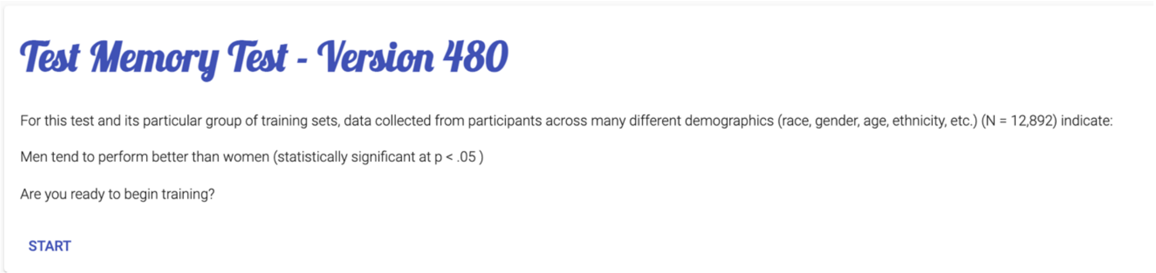
When you have reviewed all of the training sets and feel prepared for the test to begin, press the **FINISH** button and the test will begin.

Press the **START** button to start training, and the **FINISH** button once you have reviewed all of the matches.

Lastly, due to the extensive amount of data collected on memory from each test as well as other research on numeric, color, visual, written associations and memory, data that has been found to be statistically significant (if any) regarding each individual test outcomes will be presented in the following window.

Figure C3

Memory Test (MT) Stereotypic statement introduction



Test Memory Test - Version 480

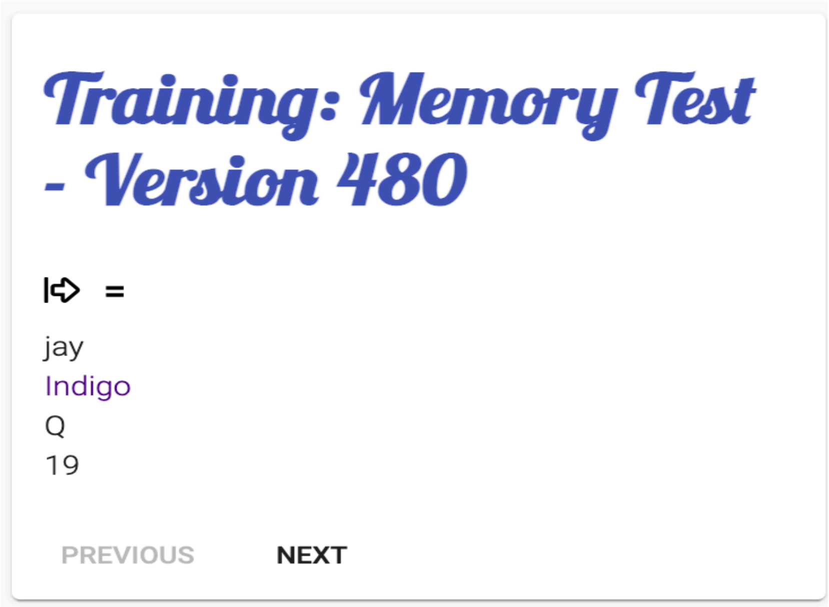
For this test and its particular group of training sets, data collected from participants across many different demographics (race, gender, age, ethnicity, etc.) (N = 12,892) indicate:

Men tend to perform better than women (statistically significant at p < .05)

Are you ready to begin training?

Figure C4

Memory Test (MT) Training Set Sample



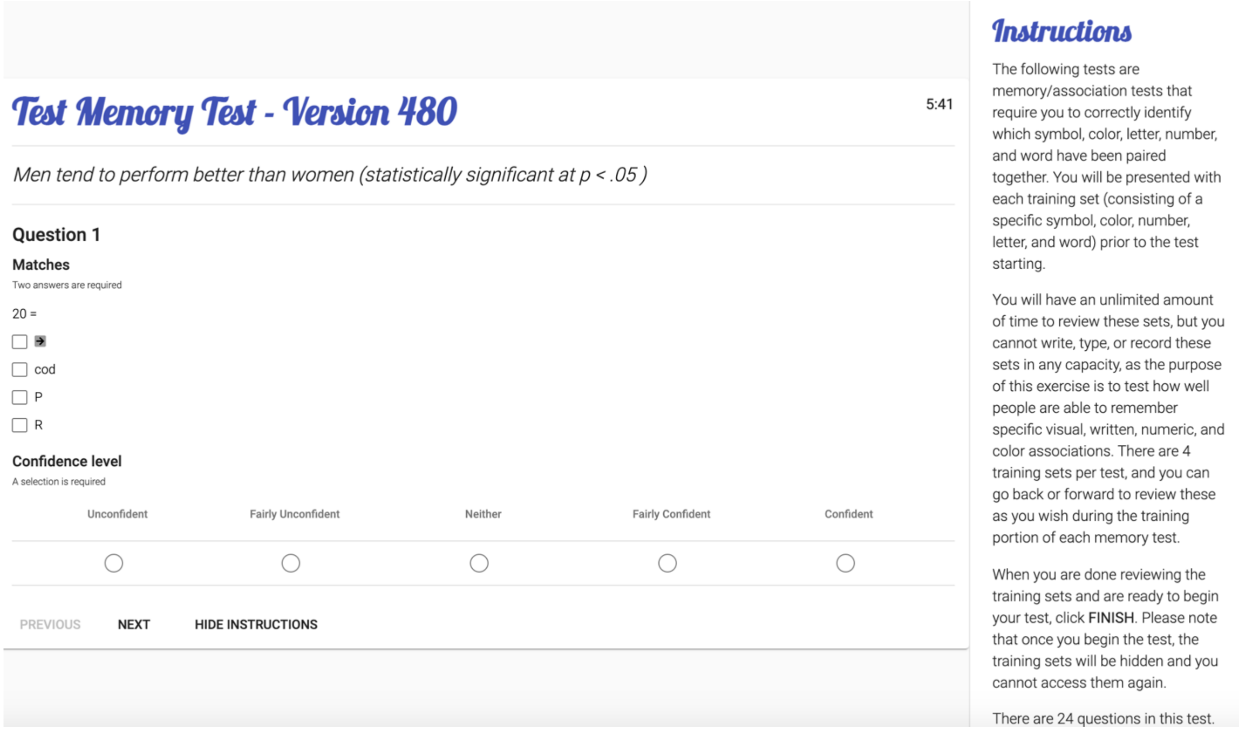
Training: Memory Test - Version 480

⇥ =

jay
Indigo
Q
19

Figure C5

Memory Test (MT) Question Sample (with instructions on side)



Test Memory Test - Version 480

5:41

Men tend to perform better than women (statistically significant at $p < .05$)

Question 1

Matches

Two answers are required

20 =

□ ➔

□ cod

□ P

□ R

Confidence level

A selection is required

Unconfident	Fairly Unconfident	Neither	Fairly Confident	Confident
○	○	○	○	○

Instructions

The following tests are memory/association tests that require you to correctly identify which symbol, color, letter, number, and word have been paired together. You will be presented with each training set (consisting of a specific symbol, color, number, letter, and word) prior to the test starting.

You will have an unlimited amount of time to review these sets, but you cannot write, type, or record these sets in any capacity, as the purpose of this exercise is to test how well people are able to remember specific visual, written, numeric, and color associations. There are 4 training sets per test, and you can go back or forward to review these as you wish during the training portion of each memory test.

When you are done reviewing the training sets and are ready to begin your test, click **FINISH**. Please note that once you begin the test, the training sets will be hidden and you cannot access them again.

There are 24 questions in this test.

Appendix D

Table D1

Participant Demographics Table

Participant Demographics (Pilots 1-3)

		Pilot 1			Pilot 2			Pilot 3		
		Mean	Median	Mode	Mean	Median	Mode	Mean	Median	Mode
	Age:	22.425	21	21	19.69	18	18	19.5	20	20
		Range:	18-49		Range:	18-44		Range:	18-21	
		Count		**% of Sample**	**Count**		**% of Sample**	**Count**		**% of Sample**
Race	White / Caucasian	25		62.5%	26		66.7%	5		62.5%
	Chinese	5		12.5%	3		7.7%	0		0.0%
	Black and/or African American	2		5.0%	2		5.1%	0		0.0%
	American Indian / Alaskan Native	1		2.5%	1		2.6%	0		0.0%
	Japanese	3		7.5%	1		2.6%	1		12.5%
	Filipino	7		17.5%	3		7.7%	3		37.5%
	Vietnamese	2		5.0%	0		0.0%	0		0.0%
	Asian Indian	0		0.0%	3		7.7%	0		0.0%
	Native Hawaiian	1		2.5%	0		0.0%	1		12.5%
	Korean	0		0.0%	1		2.6%	0		0.0%
	Others (please indicate):	3		7.5%	3		7.7%	1		12.5%
Ethnicity	Non-Latinx / Non-Hispanic	28		70.0%	30		76.9%	6		75.0%
	Mexican / Chicano or Chicana	6		15.0%	5		12.8%	1		12.5%
	Other (please indicate):	6		15.0%	4		10.3%	0		0.0%

	Salvadorian	1	2.5%	0	0.0%	0	0.0%
	Asian	2	5.0%	0	0.0%	0	0.0%
	Costa Rican	0	0.0%	1	2.6%	0	0.0%
	Chilean	1	2.5%	1	2.6%	0	0.0%
	Cuban	0	0.0%	0	0.0%	0	0.0%
	Venezuelan	0	0.0%	1	2.6%	0	0.0%
	Basque	1	2.5%	0	0.0%	0	0.0%
Gender	Female / Woman	23	57.5%	24	61.5%	6	75.0%
	Male / Man	17	42.5%	11	28.2%	2	25.0%
	Genderqueer	0	0.0%	4	10.3%	0	0.0%
Sex	Female	23	57.5%	28	71.8%	6	75.0%
	Male	17	42.5%	11	28.2%	2	25.0%
Year in School	Freshmen	7	17.5%	26	66.7%	3	37.5%
	Sophomore	6	15.0%	4	10.3%	2	25.0%
	Junior	12	30.0%	2	5.1%	1	12.5%
	Senior	11	27.5%	7	17.9%	2	25.0%
	Graduate Special	1	2.5%	0	0.0%	0	0.0%
Current / Previous Classes	Psychology 101	33	82.5%	39	100.0%	8	100.0%
	Psychology of Gender	6	15.0%	1	2.6%	1	12.5%
	Women's Studies	3	7.5%	2	5.1%	1	12.5%
MRT Study Techniques	I imagined myself being stationary and the whole object rotating with respect to me.	11	27.5%	18	46.2%	2	25.0%

	I first imagined rotating a part of the object and then checked whether the rest of the object could be rotated in the same way to match the target	20	50.0%	18	46.2%	6	75.0%
	I imagined the objects being stationary as I moved around them to view them from different perspectives	6	15.0%	10	25.6%	3	37.5%
	I imagined that the object was an animal (e.g., a snake) and where its head or arms or tail would be	1	2.5%	1	2.6%	1	12.5%
	I counted the number of cubes in the four straight segments of the object	10	25.0%	19	48.7%	5	62.5%
	I counted the number of cubes in just the two end segments of the object	19	47.5%	26	66.7%	5	62.5%
	I examined the directions of the four segments of the object with respect to each other	15	37.5%	22	56.4%	5	62.5%
	I examined the directions of the of the two end segments with respect to each other	23	57.5%	18	46.2%	3	37.5%
	None of the above	0	0.0%	0	0.0%	0	0.0%
MRT Experience	Yes	3	7.5%	7	17.9%	3	37.5%
	No	29	72.5%	28	71.8%	3	37.5%
	I don't remember	5	12.5%	4	10.3%	2	25.0%

MRT Effort	I put effort into my answers the entire time and guessed on very few questions	19	47.5%	11	28.2%	1	12.5%
	I mostly put effort into my answers, but started guessing when time was getting close to running out	10	25.0%	14	35.9%	4	50.0%
	I guessed and put effort into my answers at about equal proportions	4	10.0%	12	30.8%	2	25.0%
	I guessed a good deal of the time, but put effort into a few questions	4	10.0%	2	5.1%	0	0.0%
	I guessed most of the time	0	0.0%	0	0.0%	1	12.5%
MRT Confidence	I think I got 95-100% of the questions I attempted correct	4	10.0%	1	2.6%	0	0.0%
	I think I got 85-95% of the questions I attempted correct	8	20.0%	10	25.6%	1	12.5%
	I think I got 70-85% of the questions I attempted correct	11	27.5%	15	38.5%	2	25.0%
	I think I got 50-70% of the questions I attempted correct	0	0.0%	0	0.0%	0	0.0%
	I think I got 35-50% of the questions I attempted correct	2	5.0%	2	5.1%	1	12.5%
	I think I got fewer than 35% of the questions I attempted correct	2	5.0%	0	0.0%	1	12.5%
MT Strategies	I attempted to memorize each individual set	19	47.5%	26	66.7%	4	50.0%
	I used acronyms to remember sets (e.g., RBT for "Rip, Black, and T")	15	37.5%	11	28.2%	2	25.0%

	I created images in my head of information from the sets (e.g., I pictured a "purple bun-nY" to help remember "purple, bun, and Y" together	13	32.5%	13	33.3%	1	12.5%
	I used mnemonic devices to remember sets (e.g., "16 ripe oranges were left" to help remember "16, Rex, Orange, and L"	8	20.0%	12	30.8%	3	37.5%
	I related sets to my own experiences (e.g., "my favorite coffee place HUB is by the Blue river")	14	35.0%	8	20.5%	2	25.0%
	I memorized some and used the process of elimination for others (e.g., "I know X is not correct, so it must be Y")	15	37.5%	25	64.1%	4	50.0%
	None of the above	0	0.0%	0	0.0%	0	0.0%
MT Effort	I put effort into my answers the entire time and guessed on very few questions	19	47.5%	11	28.2%	4	50.0%
	I mostly put effort into my answers, but started guessing when time was getting close to running out	12	30.0%	12	30.8%	0	0.0%
	I guessed and put effort into my answers at about equal proportions	4	10.0%	9	23.1%	2	25.0%
	I guessed a good deal of the time, but put effort into a few questions	2	5.0%	5	12.8%	1	12.5%
	I guessed most of the time	0	0.0%	2	5.1%	1	12.5%

	4 of the memory tests	0	0.0%	0	0.0%	0	0.0%
	3 of the memory tests	10	25.0%	11	28.2%	3	37.5%
	2 of the memory tests	16	40.0%	16	41.0%	3	37.5%
	1 of the memory tests	5	12.5%	4	10.3%	0	0.0%
	None of the memory tests	2	5.0%	6	15.4%	2	25.0%
Statements seen	There were no performance differences between groups on this test	20	50.0%	18	46.2%	5	62.5%
	Women tend to perform better than men on this test	16	40.0%	24	61.5%	5	62.5%
	Men tend to perform better than women on this test	16	40.0%	20	51.3%	2	25.0%
Majors	Psychology	13	32.5%	6	15.4%	1	12.5%
	Computer Science	3	7.5%	0	0.0%	0	0.0%
	Biology	3	7.5%	3	7.7%	2	25.0%
	Finance	1	2.5%	0	0.0%	0	0.0%
	Nutrition	2	5.0%	1	2.6%	0	0.0%
	Political Science	1	2.5%	0	0.0%	0	0.0%
	Criminal Justice	2	5.0%	2	5.1%	0	0.0%
	Nursing	1	2.5%	3	7.7%	0	0.0%
	Community Health Science	1	2.5%	2	5.1%	0	0.0%
	Business Management	1	2.5%	6	15.4%	0	0.0%
	Undeclared	2	5.0%	3	7.7%	0	0.0%
	Neuroscience	2	5.0%	4	10.3%	0	0.0%
	Journalism	1	2.5%	0	0.0%	0	0.0%
	Math	1	2.5%	1	2.6%	0	0.0%

	Accounting	1	2.5%	0	0.0%	0	0.0%
	HDFS	2	5.0%	0	0.0%	2	25.0%
	Kinesiology	2	5.0%	0	0.0%	1	12.5%
	Public Health	2	5.0%	1	2.6%	0	0.0%
	Sociology	1	2.5%	0	0.0%	0	0.0%
	Communications	1	2.5%	0	0.0%	0	0.0%
	Microbiology	0	0.0%	1	2.6%	0	0.0%
	Speech Pathology	0	0.0%	1	2.6%	0	0.0%
	Theatre	0	0.0%	1	2.6%	0	0.0%
	English	0	0.0%	1	2.6%	0	0.0%
	Biochemistry	0	0.0%	2	5.1%	0	0.0%
	Chemistry	0	0.0%	1	2.6%	0	0.0%
	Music	0	0.0%	1	2.6%	0	0.0%
	Atmospheric Science	0	0.0%	1	2.6%	0	0.0%
	Mechanical Engineering	0	0.0%	1	2.6%	0	0.0%
	Veterinary Science	0	0.0%	1	2.6%	1	12.5%
	Social Work	0	0.0%	0	0.0%	1	12.5%
Minors	Psychology	2	5.0%	1	2.6%	1	12.5%
	HDFS	3	7.5%	2	5.1%	0	0.0%
	Engineering	1	2.5%	0	0.0%	0	0.0%
	Dietetics	1	2.5%	1	2.6%	0	0.0%
	Substance Abuse	1	2.5%	0	0.0%	0	0.0%
	CHS	1	2.5%	0	0.0%	0	0.0%
	Sociology	2	5.0%	0	0.0%	0	0.0%
	Communications	1	2.5%	0	0.0%	0	0.0%
	Women's Studies	1	2.5%	0	0.0%	0	0.0%

	Spanish	1	2.5%	0	0.0%	0	0.0%
	Geology	1	2.5%	0	0.0%	0	0.0%
	Philosophy	1	2.5%	0	0.0%	0	0.0%
	Econ	1	2.5%	0	0.0%	0	0.0%
	Immunology	0	0.0%	1	2.6%	0	0.0%
	Audiology	0	0.0%	1	2.6%	0	0.0%
	Kinesiology	0	0.0%	1	2.6%	0	0.0%
	Art	0	0.0%	1	2.6%	0	0.0%
	Molecular Biology	0	0.0%	1	2.6%	0	0.0%
	Neuroscience	0	0.0%	1	2.6%	0	0.0%
	Pharmacy	0	0.0%	1	2.6%	0	0.0%
	Entrepenuriership	0	0.0%	1	2.6%	0	0.0%
	Math	0	0.0%	0	0.0%	1	12.5%
	Business	0	0.0%	0	0.0%	2	25.0%
	Undeclared (or didn't identify)	19	47.5%	27	69.2%	4	50.0%

Appendix E

Table E1

Participant Demographics for the Final Study

		Final Study		
		Mean	Median	Mode
Age		20.31578947	20	19
		Range:	18 - 29	
		Count		% of Sample
Race	White / Caucasian	37		64.9%
	Chinese	1		1.8%
	Black and/or African American	5		8.8%
	American Indian / Alaskan Native	0		0.0%
	Japanese	1		1.8%
	Filipino	4		7.0%
	Vietnamese	0		0.0%
	Asian Indian	2		3.5%
	Native Hawaiian	2		3.5%
	Korean	2		3.5%
	Others (please indicate):	9		15.8%
Ethnicity	Non-Latinx / Non-Hispanic	39		68.4%
	Mexican / Chicano or Chicana	10		17.5%
	Other (please indicate):	7		12.3%
Gender	Female / Woman	28		49.1%
	Male / Man	29		50.9%
	Genderqueer	0		0.0%
Sex	Female	28		49.1%
	Male	29		50.9%
Year in School	Freshmen	20		35.1%
	Sophomore	10		17.5%
	Junior	11		19.3%
	Senior	14		24.6%
	Graduate Student	1		1.8%

Current / Previous Classes	Psychology 101	0	0.0%
	Psychology of Gender	0	0.0%
	Women's Studies	0	0.0%
MT Strategies	I attempted to memorize each individual set	33	57.9%
	I used acronyms to remember sets (e.g., RBT for "Rip, Black, and T")	29	50.9%
	I created images in my head of information from the sets (e.g., I pictured a "purple bun-nY" to help remember "purple, bun, and Y" together	27	47.4%
	I used mnemonic devices to remember sets (e.g., "16 ripe oranges were left" to help remember "16, Rex, Orange, and L"	23	40.4%
	I related sets to my own experiences (e.g., "my favorite coffee place HUB is by the Blue river")	22	38.6%
	I memorized some and used the process of elimination for others (e.g., "I know X is not correct, so it must be Y")	44	77.2%
	None of the above	0	0.0%
MT Effort	I put effort into my answers the entire time and guessed on very few questions	33	57.9%
	I mostly put effort into my answers, but started guessing when time was getting close to running out	9	15.8%
	I guessed and put effort into my answers at about equal proportions	10	17.5%
	I guessed a good deal of the time, but put effort into a few questions	4	7.0%
	I guessed most of the time	0	0.0%
	All 6 of the AMATs	4	7.0%
	5 AMATs	16	28.1%

	4 AMATs	13	22.8%
	3 AMATs	12	21.1%
	2 AMATs	8	14.0%
	1 AMATs	3	5.3%
Statements seen	Men and women perform equally	44	77.2%
	Women tend to perform better than men	21	36.8%
	Men tend to perform better than women	15	26.3%
Majors	Agricultural Science	1	1.8%
	Business	4	7.0%
	Civil Engineering	1	1.8%
	Criminal Justice	2	3.5%
	Exploratory	1	1.8%
	Kinesiology	6	10.5%
	Management	2	3.5%
	Marketing	2	3.5%
	Microbiology	4	7.0%
	Molecular Microbiology	1	1.8%
	Neuroscience	11	19.3%
	Psychology	14	24.6%
	Public Health	4	7.0%
	Social Work	3	5.3%
	Undecided	1	1.8%
	Undeclared	1	1.8%
Minors	American Government	1	1.8%
	Biology	1	1.8%
	CAS	2	3.5%
	CASAT	1	1.8%
	Chemistry	1	1.8%
	Communications	1	1.8%
	English	1	1.8%
	Gerontology	1	1.8%
	HDFS	3	5.3%
	Human Resources	1	1.8%
	Immunology	2	3.5%

	Math	1	1.8%
	Psychology	1	1.8%
	Substance Abuse	2	3.5%
	Substance Abuse and Treatment	1	1.8%
	Women's Studies	1	1.8%
	Undeclared (or didn't identify)	38	66.7%

Appendix F

Program Scripts for all AMATs

AMAT – v. 90, 105, 221, 357, 479, & 503
Script:
Test completed! Select “Start” when you are ready to begin training for the next test
Statement:
Men and women perform equally

AMAT – v. 91, 106, 222, 358, 480, & 504
Script:
Test completed! Select “Start” when you are ready to begin training for the next test
Statement:
Men tend to perform better than women

AMAT – v. 92, 107, 223, 359, 481, & 505
Script:
Test completed! Select “Start” when you are ready to begin training for the next test
Statement:
Women tend to perform better than men

AMAT – v. 93, 108, 224, 360, 482, & 506
Script:
<p>In this study, you are being asked to take the <b> Atlanta Memory Adaption Test (AMAT).</b> Over the last 30 years, this has been one of the most widely used tests on human memory performance. Memory is critical for success in life. AMAT scores consistently predict academic and professional success. Research in cognitive and brain science shows that some people are at an advantage when taking this test compared to others. Certain associations and pairings of different numbers, letters, symbols, and colors, seem to be easier to remember for certain groups of people than others.</p>
<p>
<p> <b> However, research shows that there are no gender differences on the following versions of the AMAT that you are about to take (i.e., men and women perform equally). </b> </p>
<p>
<p>Select “start” when you are ready to begin training for the first test</p>

Statement:
Men and women perform equally

AMAT – v. 94, 109, 225, 361, 483, & 507
Script:
<p>In this study, you are being asked to take the <b>**Atlanta Memory Adaption Test (AMAT)**.</b> Over the last 30 years, this has been one of the most widely used tests on human memory performance. Memory is critical for success in life. AMAT scores consistently predict academic and professional success. Research in cognitive and brain science shows that some people are at an advantage when taking this test compared to others. Certain associations and pairings of different numbers, letters,

symbols, and colors, seem to be easier to remember for certain groups of people than others.</p>

<p>

<p> <b> Research shows that there is a strong male advantage on the following versions of the AMAT that you are about to take (i.e., men tend to perform better than women). </b> </p>

<p>

<p>Select “start” when you are ready to begin training for the first test</p>

Statement:

Men tend to perform better than women

AMAT – v. 95, 110, 226, 362, 484, & 508

Script:

<p>In this study, you are being asked to take the <b> **Atlanta Memory Adaption Test (AMAT)**. </b> Over the last 30 years, this has been one of the most widely used tests on human memory performance. Memory is critical for success in life. AMAT scores consistently predict academic and professional success. Research in cognitive and brain science shows that some people are at an advantage when taking this test compared to others. Certain associations and pairings of different numbers, letters, symbols, and colors, seem to be easier to remember for certain groups of people than others.</p>

<p>

<p> <b> Research shows that there is a strong female advantage on the following versions of the AMAT that you are about to take (i.e., women tend to perform better than men). </b> </p>

<p>

<p>Select “start” when you are ready to begin training for the first test</p>

Statement:

Women tend to perform better than men

AMAT – v. 96, 111, 227, 363, 485, & 509

Script:

<p>In this study, you are being asked to take the <b> **Atlanta Memory Adaption Test (AMAT)**. </b> Over the last 30 years, this has been one of the most widely used tests on human memory performance. Memory is critical for success in life. AMAT scores consistently predict academic and professional success. Research in cognitive and brain science shows that some people are at an advantage when taking this test compared to others. Certain associations and pairings of different numbers, letters, symbols, and colors seem to be easier to remember for certain groups of people than others.</p>

<p>

<p><b> Now we would like you to work on versions of the AMAT that have shown no gender differences (i.e., men and women tend to perform equally). These versions of the AMAT tend to predict life success better than the previous versions. The purpose of this research is to examine why there were no gender

differences on these versions of the AMAT compared to the ones you took earlier. </b><p>
<p>
<p>Select “start” when you are ready to begin training for the next test</p>

Statement:
Men and women perform equally

AMAT – v. 97, 112, 228, 364, 486, & 510
Script:
<p>In this study, you are being asked to take the <b> **Atlanta Memory Adaption Test (AMAT)**. </b> Over the last 30 years, this has been one of the most widely used tests on human memory performance. Memory is critical for success in life. AMAT scores consistently predict academic and professional success. Research in cognitive and brain science shows that some people are at an advantage when taking this test compared to others. Certain associations and pairings of different numbers, letters, symbols, and colors seem to be easier to remember for certain groups of people than others.</p>
<p>
<p>**<b> We would like you to continue working on versions of the AMAT that again have shown no gender differences (i.e., men and women tend to perform equally). This version of the AMAT tends to predict life success better than the previous version. The purpose of this research is to examine why there were no gender differences on these versions of the AMAT. </b>**</p>
<p>
<p>Select “start” when you are ready to begin training for the next test</p>

Statement:
Men and women perform equally

AMAT – v. 98, 113, 229, 365, 487, & 511
Script:
<p>In this study, you are being asked to take the <b> **Atlanta Memory Adaption Test (AMAT)**. </b> Over the last 30 years, this has been one of the most widely used tests on human memory performance. Memory is critical for success in life. AMAT scores consistently predict academic and professional success. Research in cognitive and brain science shows that some people are at an advantage when taking this test compared to others. Certain associations and pairings of different numbers, letters, symbols, and colors seem to be easier to remember for certain groups of people than others.</p>
<p>
<p>**<b> Now we would like you to work on different versions of the AMAT that have shown a strong male advantage (i.e., men tend to perform better than women). These versions of the AMAT tend to predict life success better than the previous versions. The purpose of this research is to examine why men are performing so much better than women on the following versions of the AMAT compared to the ones you took earlier. </b>**</p>
<p>

<p>Select “start” when you are ready to begin training for the next test</p>

Statement:
Men tend to perform better than women

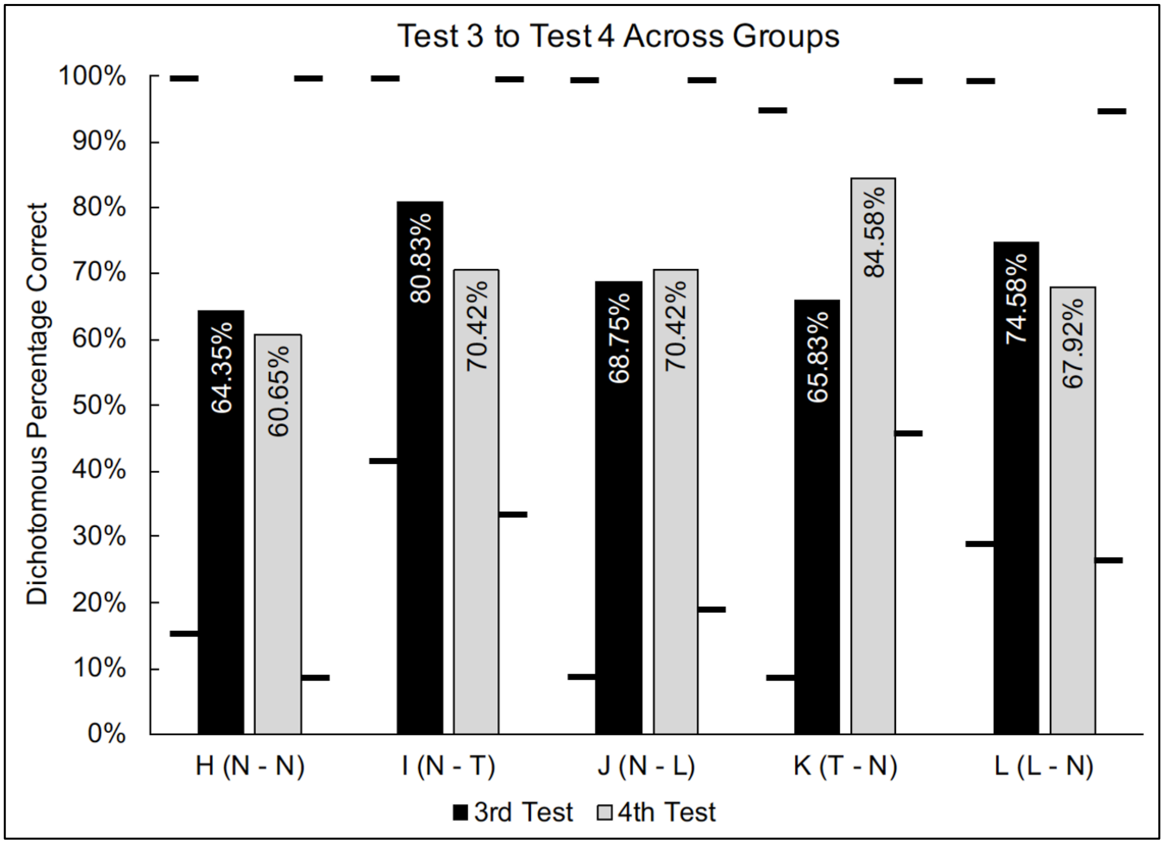
AMAT – v. 99, 114, 230, 366, 488, & 512
Script:
<p>In this study, you are being asked to take the <b> **Atlanta Memory Adaption Test (AMAT)**. </b> Over the last 30 years, this has been one of the most widely used tests on human memory performance. Memory is critical for success in life. AMAT scores consistently predict academic and professional success. Research in cognitive and brain science shows that some people are at an advantage when taking this test compared to others. Certain associations and pairings of different numbers, letters, symbols, and colors seem to be easier to remember for certain groups of people than others.</p>
<p>
<p><b> **Now we would like you to work on different versions of the AMAT that have shown a strong female advantage (i.e., women tend to perform better than men). These versions of the AMAT tend to predict life success better than the previous versions. The purpose of this research is to examine why women are performing so much better than men on the following versions of the AMAT compared to the ones you took earlier.** </b></p>
<p>
<p>Select “start” when you are ready to begin training for the next test</p>

Statement:
Women tend to perform better than men

Appendix G

Figure G1

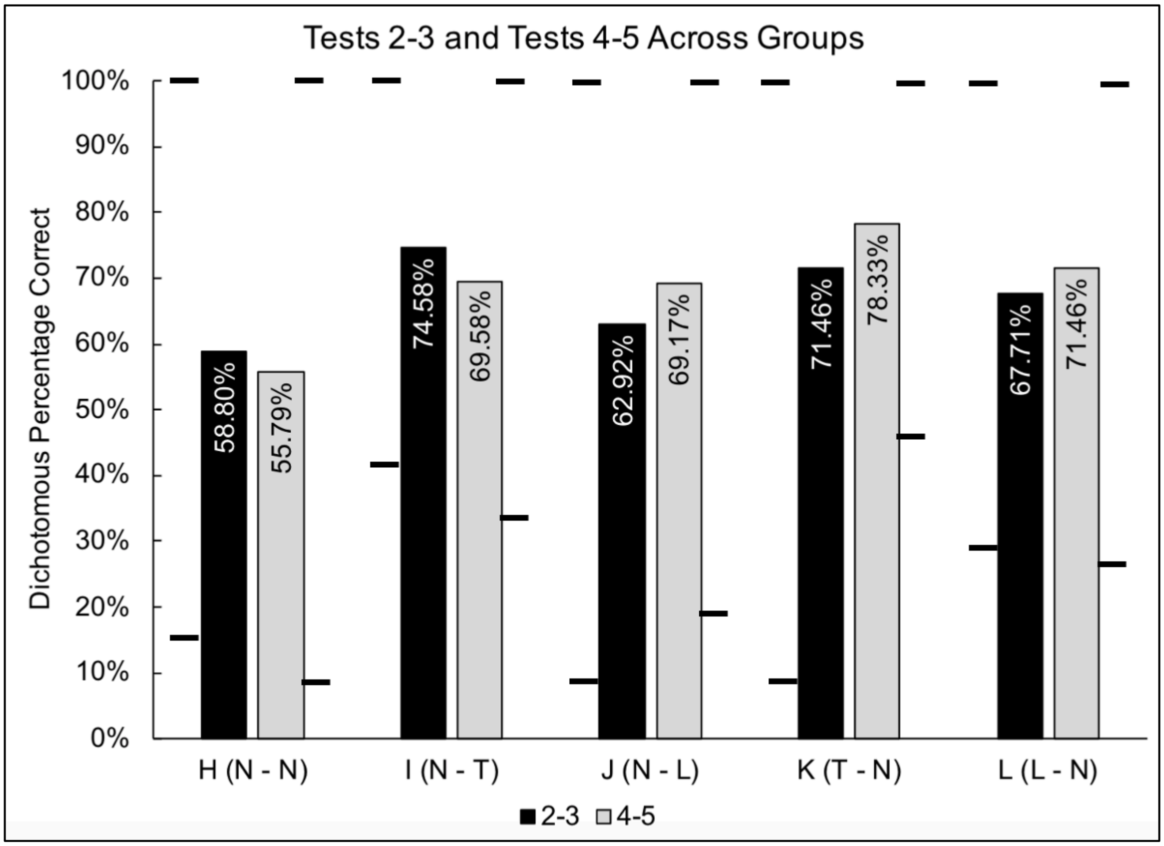
AMAT 3 and AMAT 4 averages across groups (dichotomous)



Note. Black range brackets beside the AMAT 3 and AMAT 4 average score bars show the range in scores across all participants in each group. Ranges are as follows: Group H: AMAT 3 (16.7% - 100%), AMAT 4 (8.3% - 100%); Group I: AMAT 3 (41.7% - 100%), AMAT 4 (33.3% - 100%); Group J: AMAT 3 (8.3% - 100%), AMAT 4 (20.8% - 100%); Group K: AMAT 3 (8.3% - 95.8%), AMAT 4 (45.8% - 100%); Group L: AMAT 3 (29.2% - 100%), AMAT 4 (25% - 95.8%).

Figure G2

AMAT 2-3 and AMAT 4-5 averages across groups (dichotomous)



Note. Black range brackets beside the AMAT 2-3 and AMAT 4-5 average score bars show the range in scores across all participants in each group. Ranges are as follows: Group H: AMATs 2-3 (16.7% - 100%), AMATs 4-5 (8.3% - 100%); Group I: AMATs 2-3 (20.8% - 100%), AMAT 4-5 (0% - 100%); Group J: AMATs 2-3 (8.3% - 100%), AMATs 4-5 (20.8% - 100%); Group K: AMATs 2-3 (8.3% - 100%), AMATs 4-5 (29.2% - 100%); Group L: AMATs 2-3 (4.2% - 100%), AMATs 4-5 (25% - 100%).

Appendix H

Table H1

Means Table for Within-Subject Variables

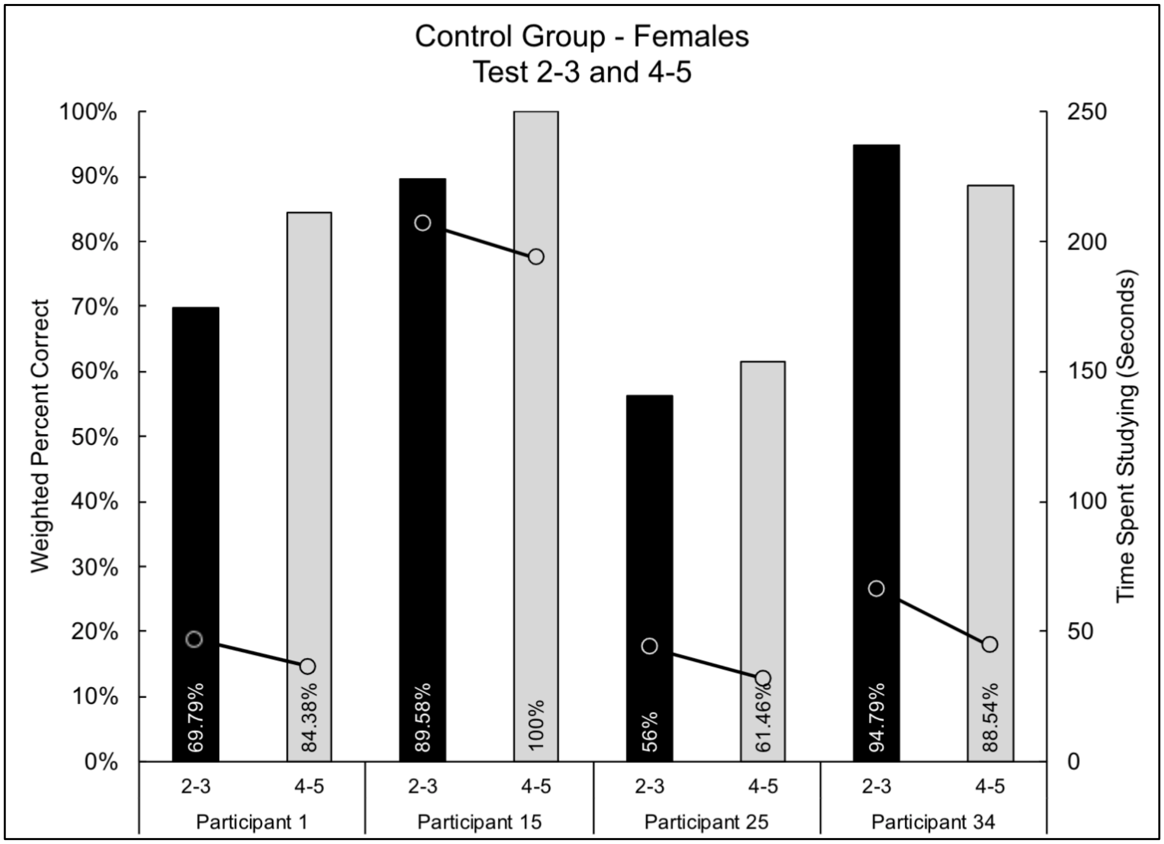
	Group H		Group I		Group J		Group K		Group L	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
AMAT 1	.70	.17	.57	.17	.69	.19	.61	.21	.64	.19
AMAT 2	.72	.18	.68	.33	.72	.19	.86	.12	.76	.15
AMAT 3	.78	.17	.81	.20	.80	.20	.76	.27	.83	.17
AMAT 4	.72	.28	.70	.25	.82	.19	.91	.12	.81	.11
AMAT 5	.69	.22	.69	.29	.79	.22	.83	.17	.85	.19
AMAT 6	.64	.22	.82	.18	.85	.14	.86	.14	.89	.14

Note. n = 9 (Group H); n = 10 (Groups I, J, K, & L). Within Effects: Group I (AMAT 1 significantly less than AMAT 3, AMAT 4 and AMAT 6). Group K (AMAT 1 significantly less than AMAT 4 and AMAT 6).

Appendix I

Figure I1

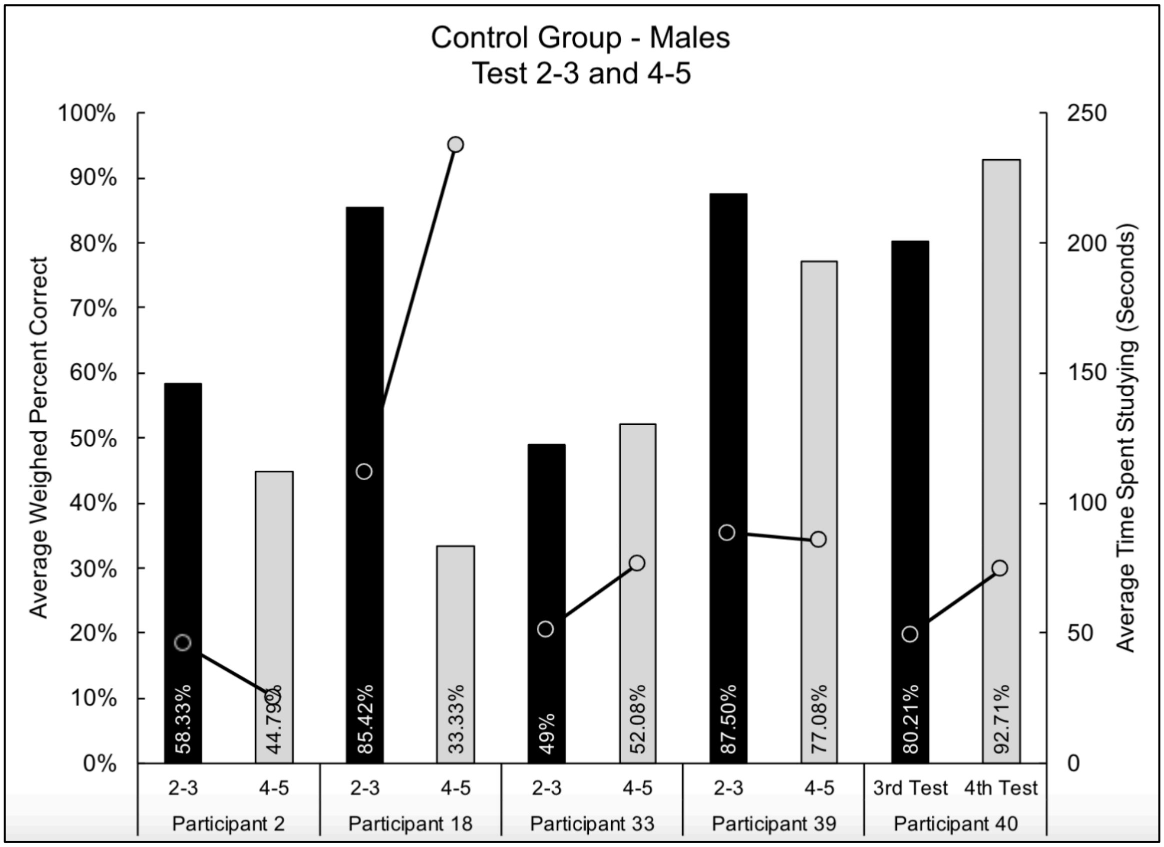
AMATs 2-3 and 4-5 for Group H (Females)



Note. n = 4.

Figure I2

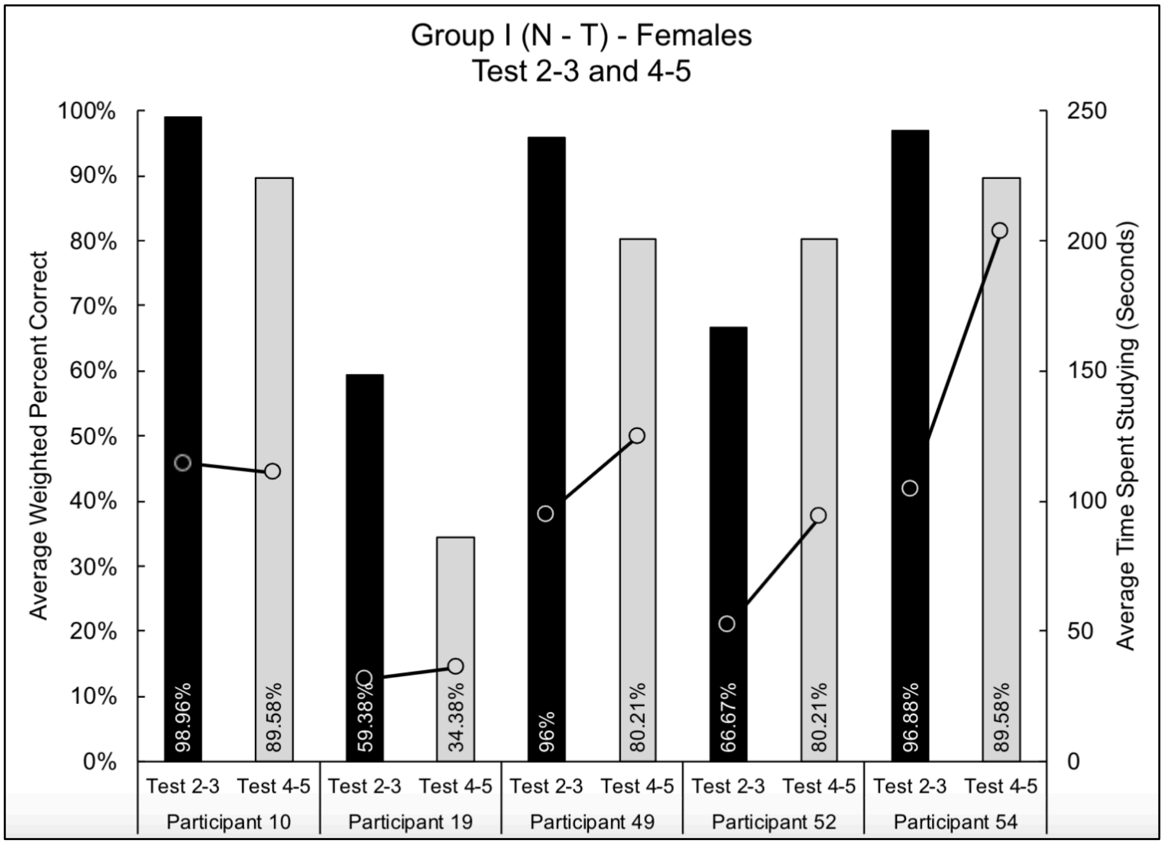
AMATs 2-3 and 4-5 for Group H (Males)



Note. n = 5.

Figure I3

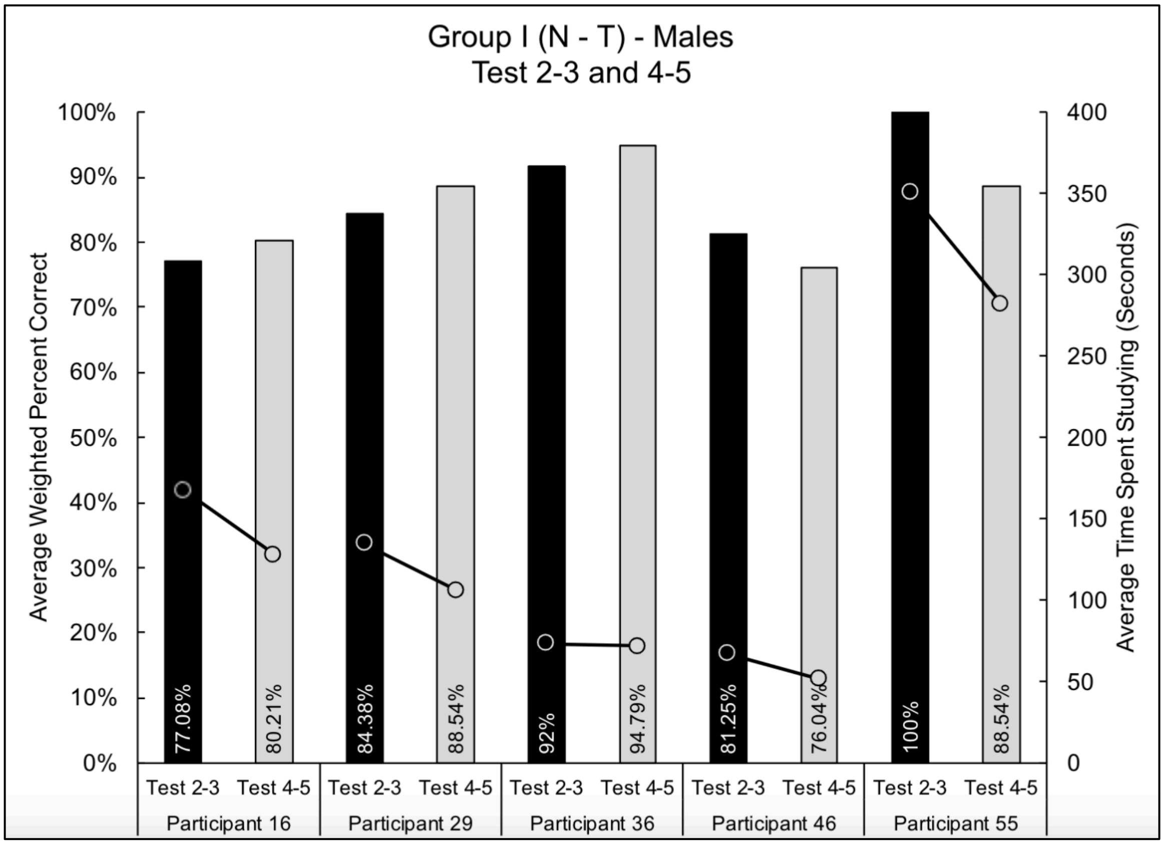
AMATs 2-3 and 4-5 for Group I (Females)



Note. n = 5.

Figure I4

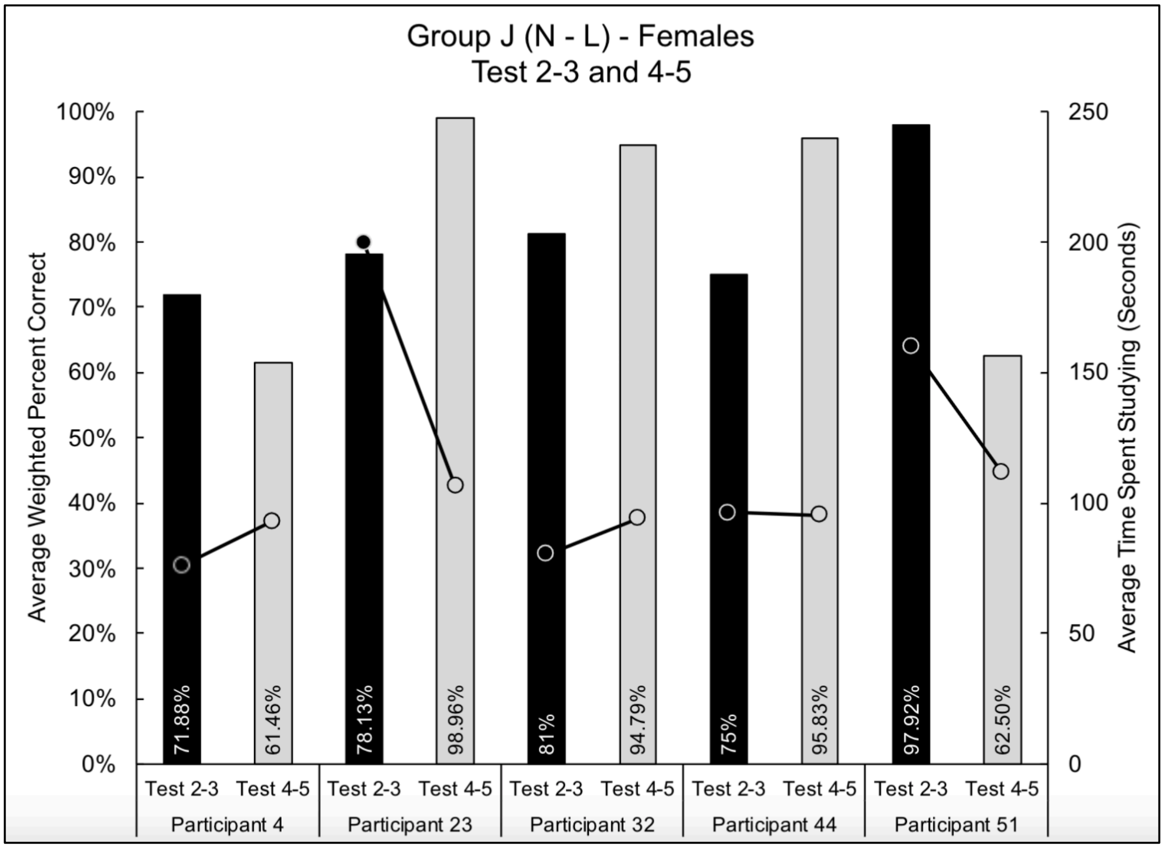
AMATs 2-3 and 4-5 for Group I (Males)



Note. n = 5.

Figure I5

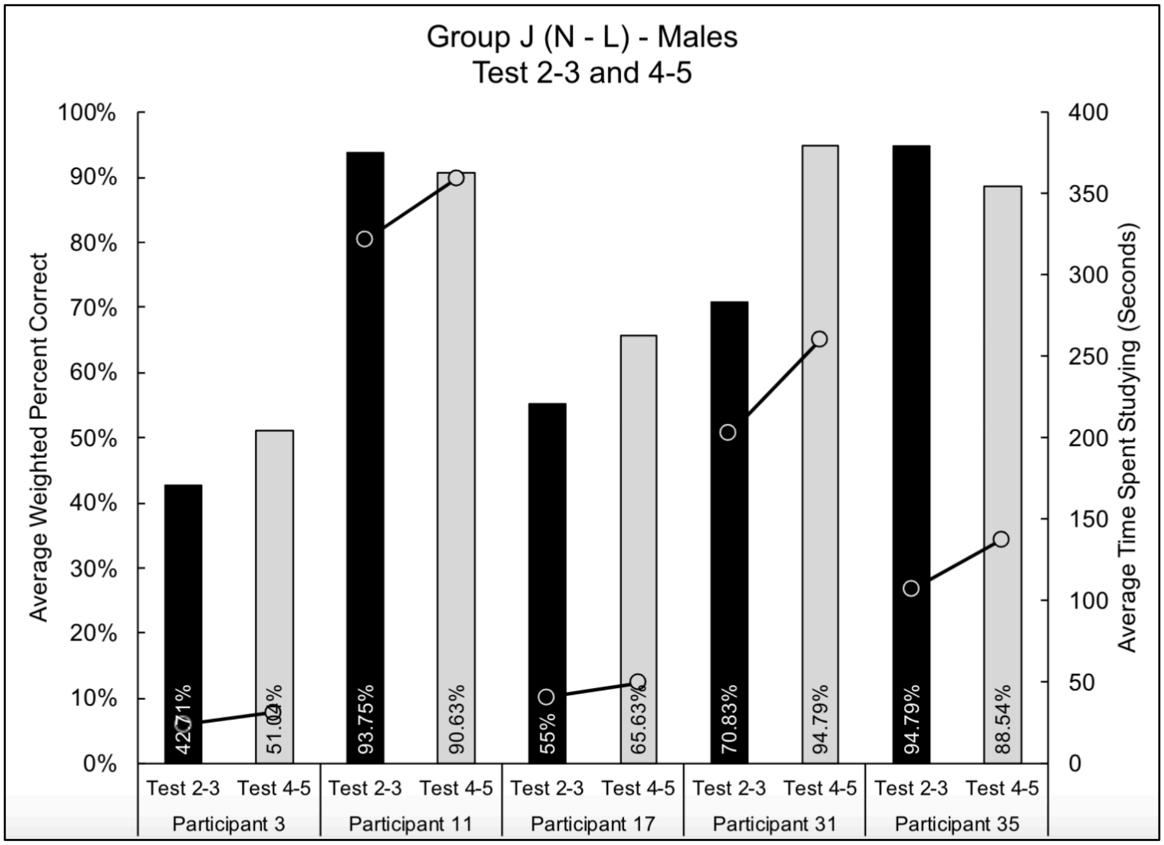
AMATs 2-3 and 4-5 for Group J (Females)



Note. n = 5.

Figure I6

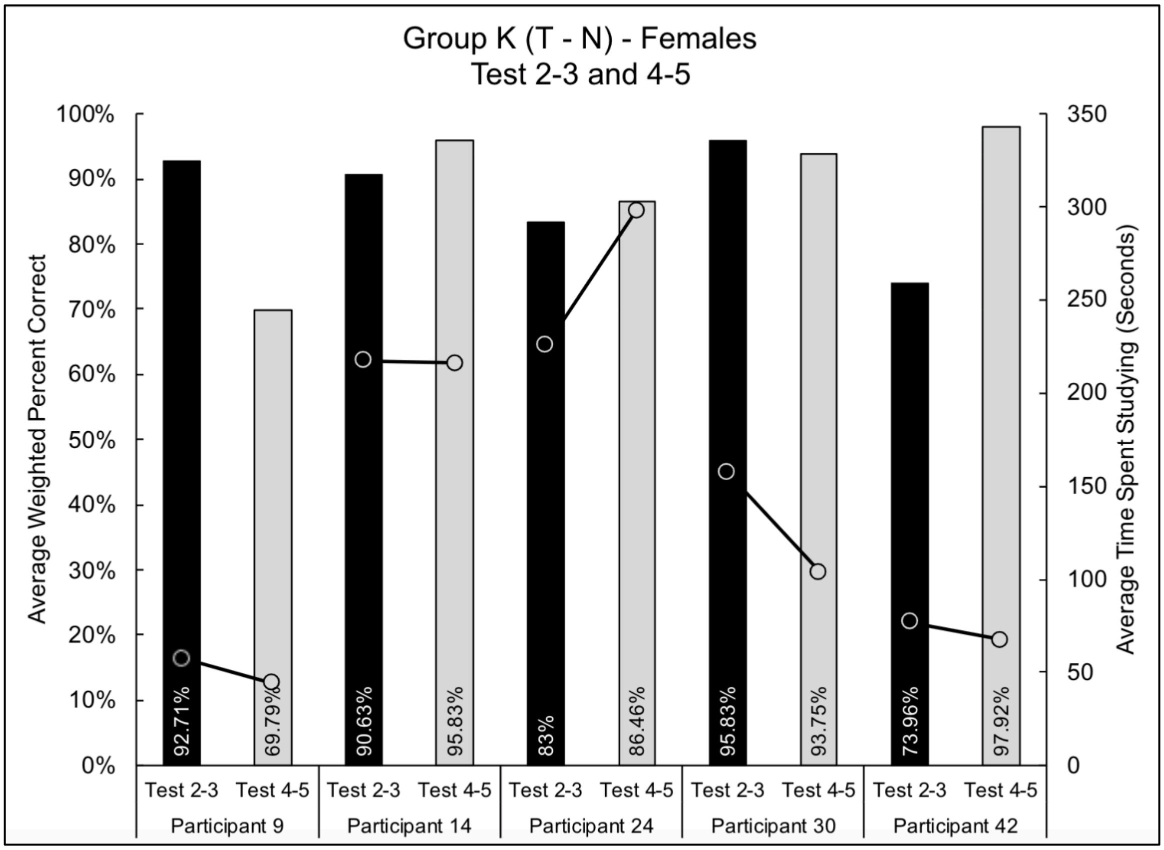
AMATs 2-3 and 4-5 for Group J (Males)



Note. n = 5.

Figure I7

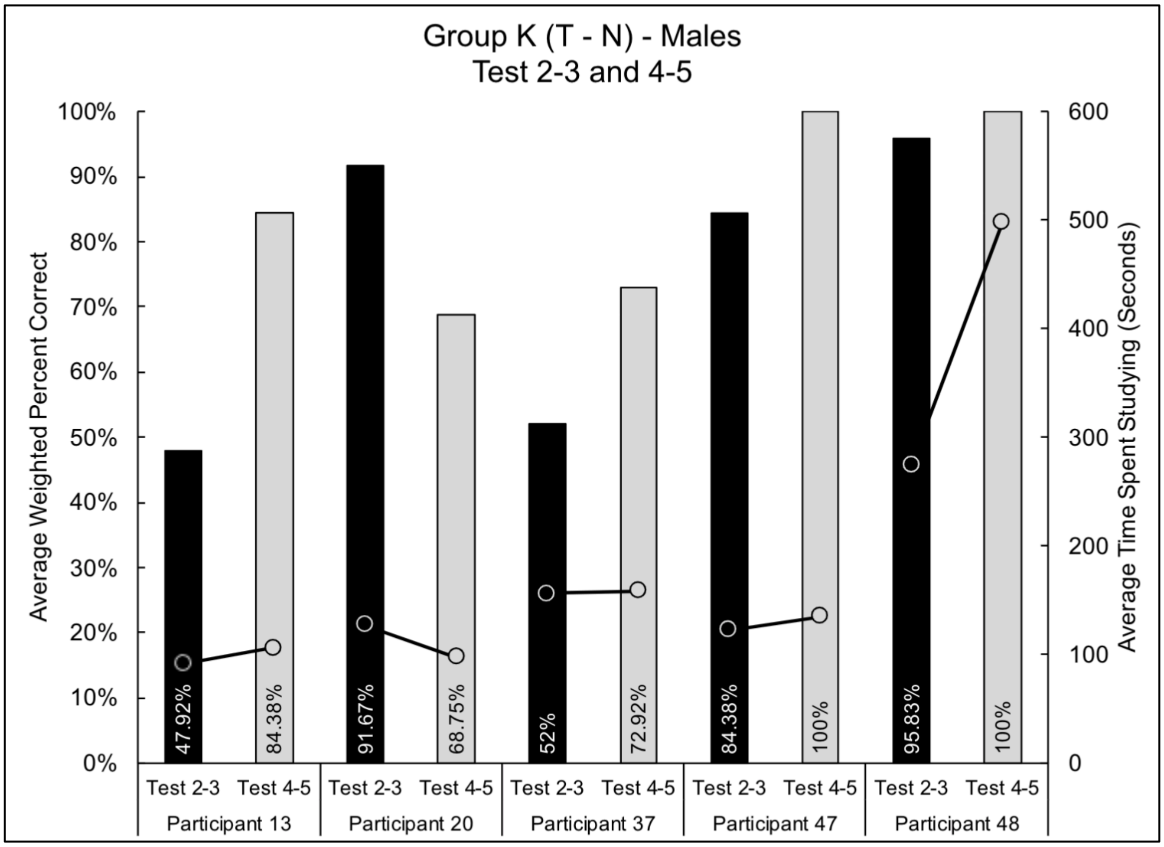
AMATs 2-3 and 4-5 for Group K (Females)



Note. n = 5.

Figure I8

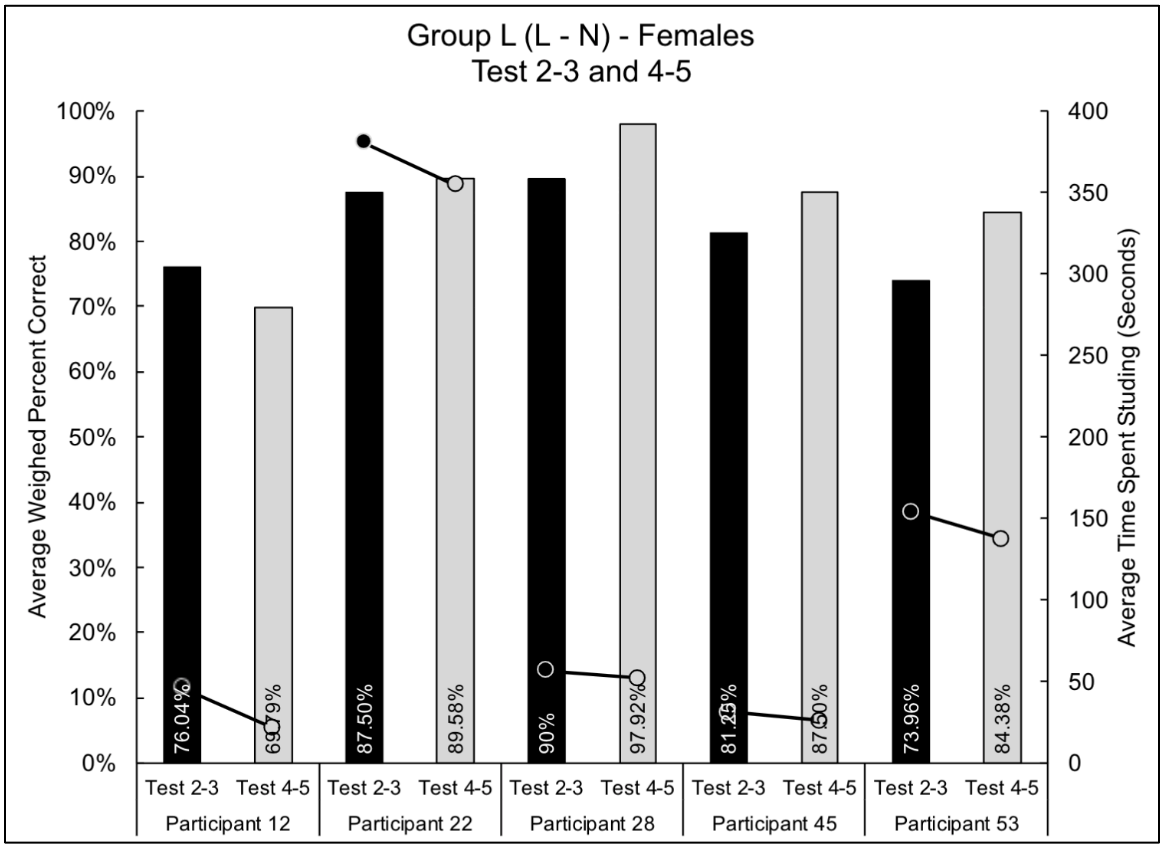
AMATs 2-3 and 4-5 for Group K (Males)



Note. n = 5.

Figure I9

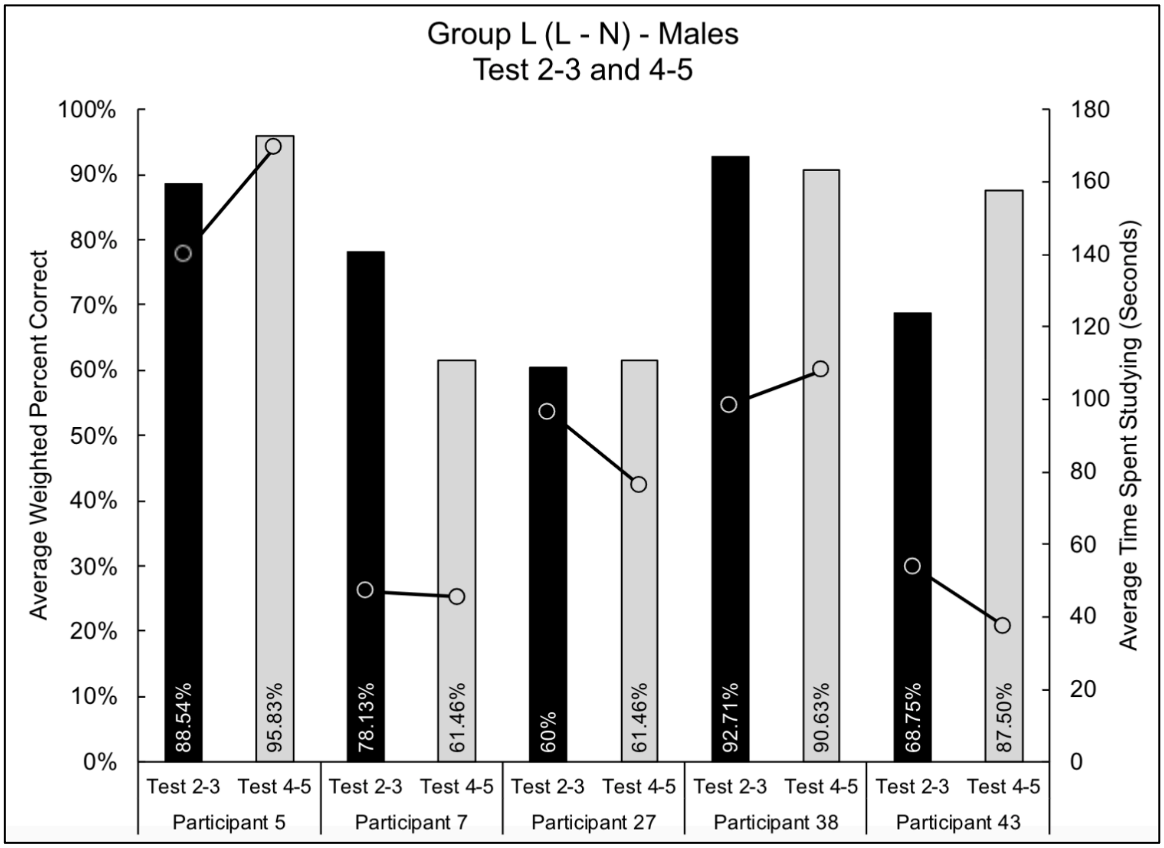
AMATs 2-3 and 4-5 for Group L (Females)



Note. n = 5.

Figure I10

AMATs 2-3 and 4-5 for Group L (Males)

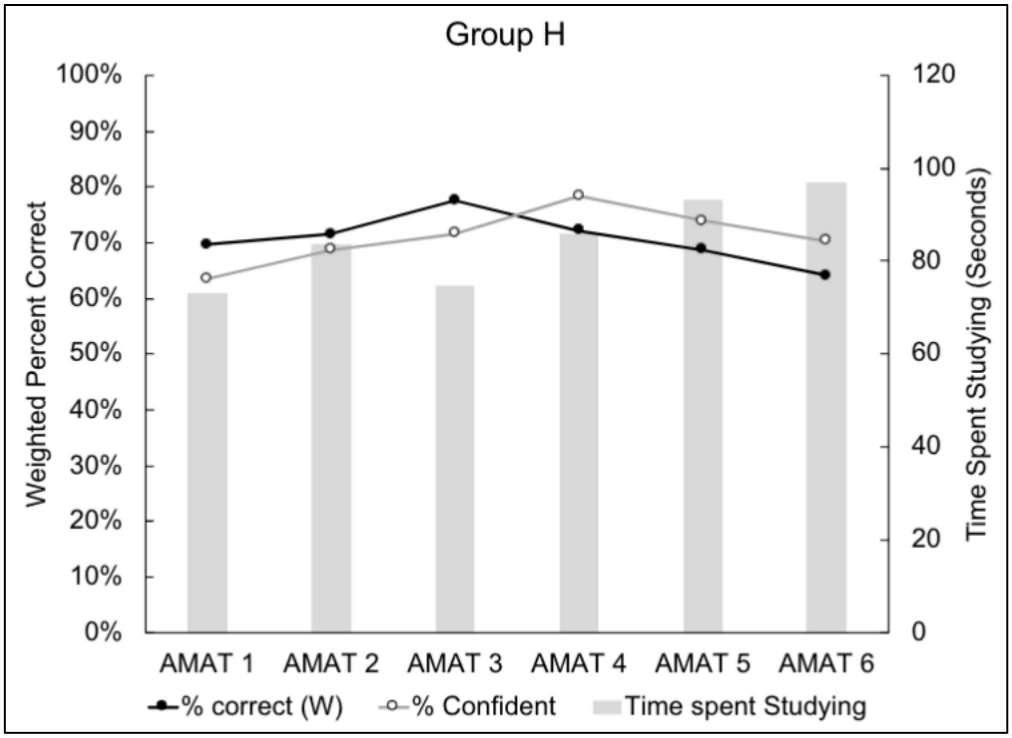


Note. n = 5.

Appendix J

Figure J1

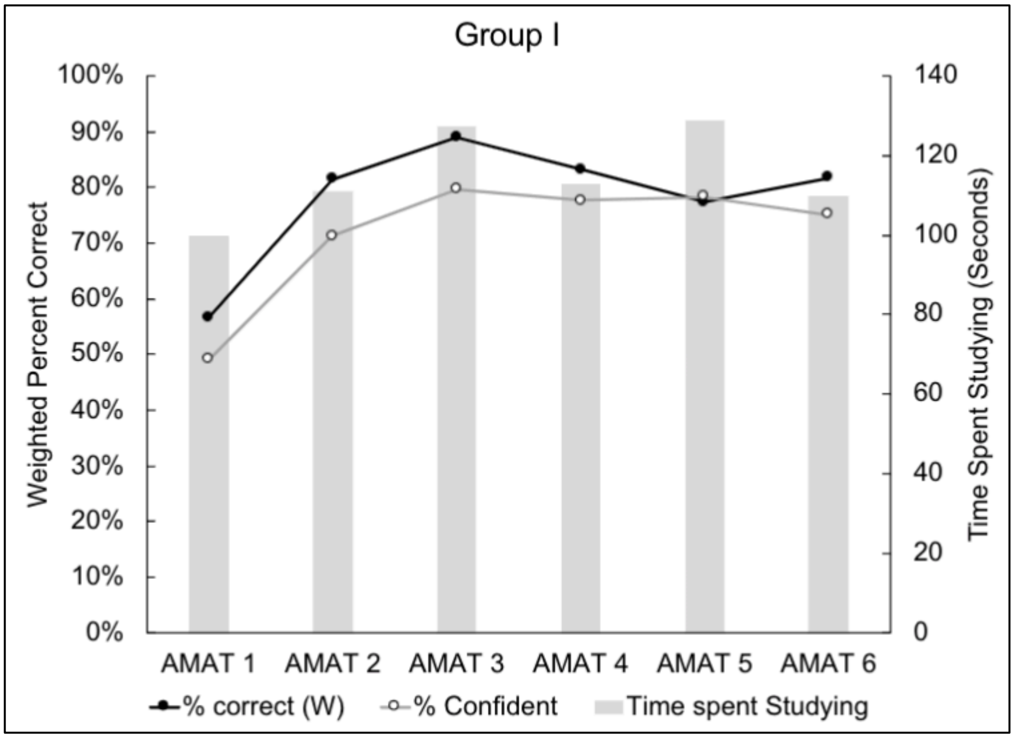
Averages across AMATs 1-6 (Group H)



Note. n = 9.

Figure J2

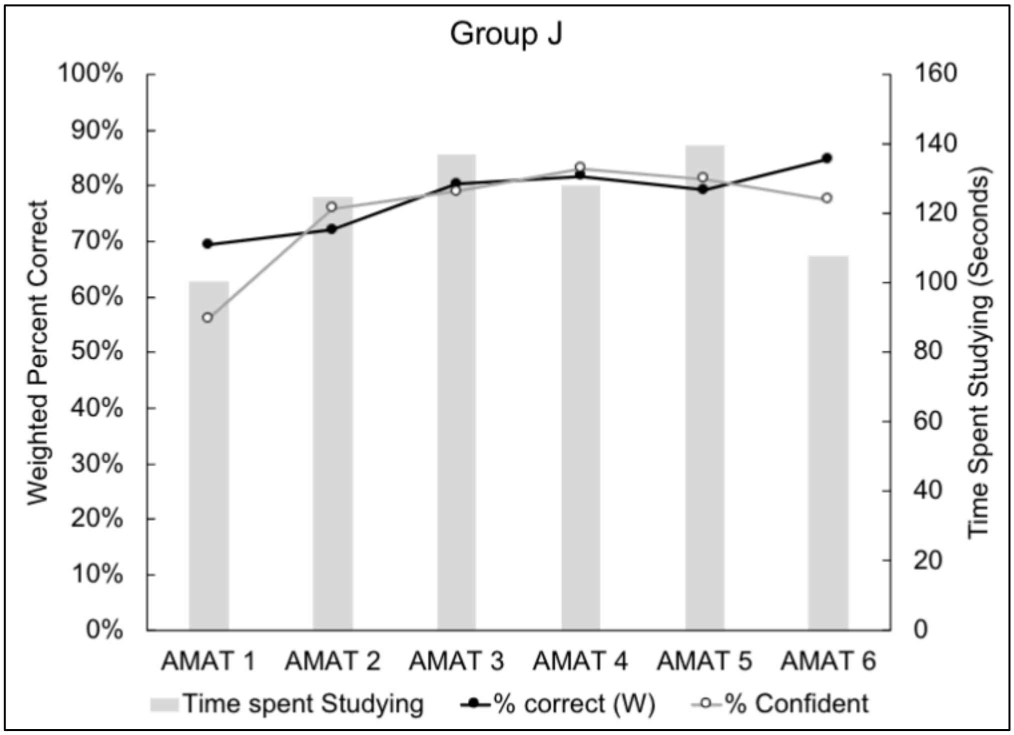
Averages across AMATs 1-6 (Group I)



Note. n = 10.

Figure J3

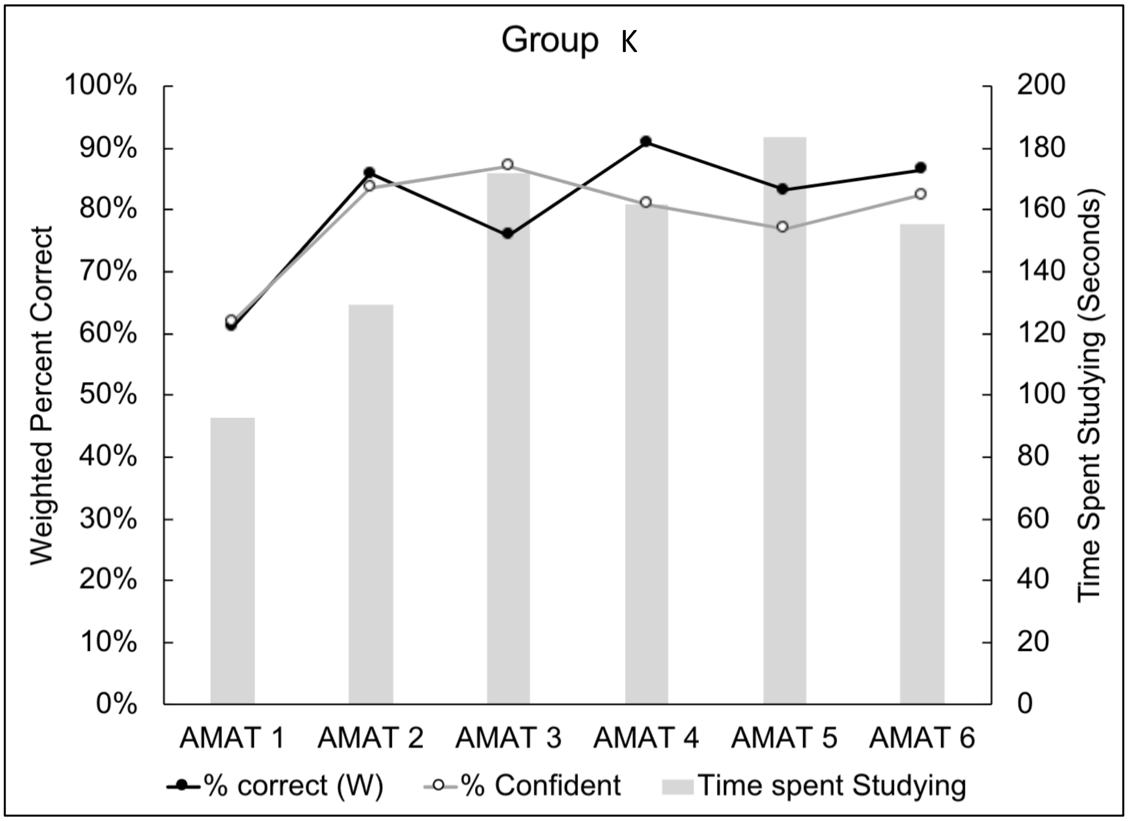
Averages across AMATs 1-6 (Group J)



Note. n = 10.

Figure J4

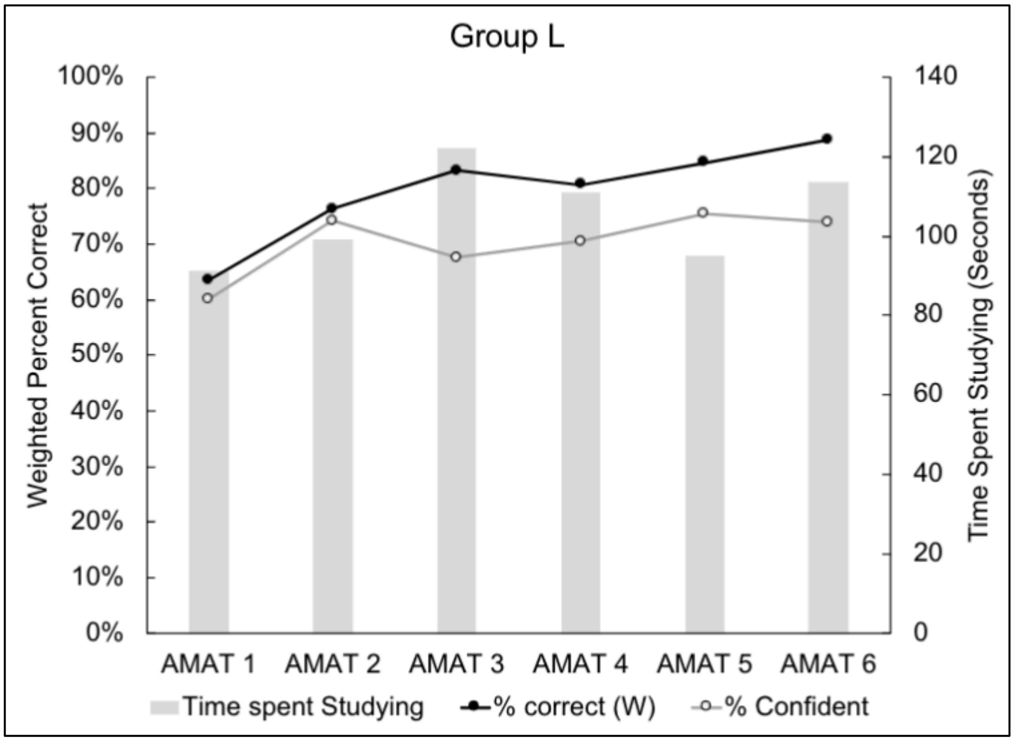
Averages across AMATs 1-6 (Group K)



Note. n = 10.

Figure J5

Averages across AMATs 1-6 (Group I)



Note. n = 10.