Title
Measuring Stereotype Threat

Author
Bathia, Shruti

Publication Date
2021

Peer reviewed|Thesis/dissertation

Measuring Stereotype Threat

by

Shruti Bathia

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Education

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark Wilson, Chair
Professor Bruce Fuller
Professor Jason Okonofua

Fall 2021

Abstract

Measuring Stereotype Threat

by

Shruti Bathia

Doctor of Philosophy in Education

University of California, Berkeley

Professor Mark Wilson, Chair

Stereotype threat is a situational experience, in which individuals feel vulnerable to the possibility of being judged because of a negative stereotype associated to their social group. This experience leads to decline in performance, even among highly skilled individuals. The objective of this research is to provide researchers with a comprehensive theoretical framework of measuring stereotype threat and develop and validate the stereotype threat instrument. To date, no instrument has completely been able to explain the amount of stereotype threat experienced by individuals, placed in different situations and belonging to different social groups.

In Chapter 1, we review past research on stereotype threat and discuss various influential moderators of stereotype threat. We identify gaps in the pre-existing measures and explore ways of operationalizing the stereotype threat construct.
In Chapter 2, we establish the stereotype threat balance framework and define a new stereotype threat construct. We develop the stereotype threat instrument using the four building blocks approach.
In Chapter 3, we measure stereotype threat experienced by transfer students studying in four-year universities across the nation. We collect evidence in support of using the proposed instrument as a valid metric for measuring stereotype threat.
In Chapter 4, using differential item functioning, we investigate for any potential item bias in the stereotype threat instrument for the different racial groups in our sample. Once we establish no measurement bias, we analyze differences in outcomes for the five racial groups, Asians, Blacks, Latinos, Whites and Others (American Indians or Native Americans, Bi-racial, Pacific Islanders)

Keywords: stereotype threat, measurement, validity, item response theory.

To the tiny heartbeat who is beating inside me, I know you are waiting and so am I.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I will mention a small subset of many wonderful people who have walked alongside. I thank my advisor, Professor Mark Wilson for constantly pushing me to be my best, Professor Bruce Fuller for nurturing me as an academic researcher, Professor Jason Okonofua for his continuous support and Professor Harish Chaudhury for believing in me in the very early days of my career. I thank Perman Gocchyev for being so generous with his time and advice and Linda Morell for her constant motivation and support.

Many colleagues and friends have contributed their knowledge, time and thought to my work, I thank all of them. I thank my parents for raising me to be strong and independent and my sisters, Nehal and Dharini for their constant support. Most importantly, I thank my husband Vishal for his love, encouragement and understanding.

Chapter One

Understanding the Dimensions of Stereotype Threat: A Review of Literature

"As an Asian girl, I hate it so much. People tend to think that the stereotype of being smart and getting good grades is good and I guess it could be in some ways, but it has been used so many times to undermine my work and effort. Like when I try super hard to get good test scores, a high GPA, etc. people will say it's just because you are Asian and I feel like my hard work isn't acknowledged and they think if you're Asian you don't have to study, practice or try at all. I just wish people knew that it's about how much effort you put in, not about your ethnicity. Not only that, but a model minority is also used against other minorities like Black and Hispanic people which can create tension and undermine their struggles."

- *anonymous high school student*

In this chapter, we review past research on stereotype threat. We uncover various dimensions of stereotype threat. Socially driven, situationally established and uniquely experienced psychological construct, stereotype threat exerts an influence in all our lives. We explore the individual and situational aspect of stereotype threat and why it differs in terms of severity among individuals and what causes these differences in experience. While some individuals either consciously or subconsciously can weaken the influence of stereotypes in their lives, some individuals' choices, personality and course of life are altered irreversibly. We identify some of the key moderators of stereotype threat such as (a) stigma consciousness, which captures individuals' awareness, consciousness and belief in the stereotype, (b) group identification, which captures individuals' sense of belongingness with the stigmatized group under consideration, (c) domain identification which captures the extent to which a domain is viewed as important by the individual and (d) core strength, which captures individuals' strength of mindset and the ability to overcome the deleterious effects of stereotype threat. We attempt to investigate and unravel the challenges of measuring stereotype threat and why past research has failed to establish a consensus with regards to the operationalization of the stereotype threat construct. We uncover the limitations and challenges of preexisting measures of stereotype threat. We use the theoretical framework developed in this chapter as the basis for developing the stereotype threat instrument in the subsequent chapter.

INTRODUCTION

Stereotype threat is a psychological phenomenon that inhibits the performance of individuals in domains where negative ability stereotypes about the group are highlighted (Picho & Brown, 2011). A distinguishing aspect of stereotype threat is that it is theorized to be caused by situational factors (Steele, 1997; Spencer et al., 2002). Studies in the past have revealed that constant experience of stereotype threat may have long term effects on the well-being of an individual (Major et al., 2003). It also results in lower self-esteem (Schmitt et al., 2002). The opposite also holds true, i.e., lower self-esteem may also result in greater stereotype threat (Vass et al., 2015). Stereotype threat is more likely to occur when the task is more demanding (Keller, 2007), when the individual is more conscious about the stigma (Hess et al., 2009), when the individual strongly identifies with the stigmatized social group (Davies et al., 2006) and when the individual values the domain under consideration (Steele, 1997; Pavlova et al., 2014). Pennington and her colleagues tracked 300 experiments that have illustrated the deleterious effects of stereotype threat across many different populations (Steele, 1997; Major et al., 2003; Pennington et al., 2016). Thus, despite well-established negative consequences of stereotype threat and a rich literature on the social psychological phenomena that lead to an individual experiencing more (or less) of the threat, there are currently no measures that successfully operationalize stereotype threat (Xavier et al., 2014). We explore some of the key issues of measuring stereotype threat and the key moderators that lead to differences in stereotype threat experience among individuals.

BACKGROUND

We all agree that people should be treated equally and fairly. Yet, we are inundated by news reports and personal experiences that portray that differences in terms of race, ethnicity, sexual orientation, gender and socioeconomic status continue to lead to discrimination. These media reports reinforce negative stereotypes about stigmatized student's ability to succeed in school. In education, there is a growing body of research that undermines the conventional assumption that genetics or cultural differences lead some students to underperform in academic tests. Instead, it has become clear that these negative stereotypes raise inhibiting doubts and high-pressure anxieties in a test-taker's mind, resulting in the phenomenon of stereotype threat (Steele & Aronson, 1995).

Stereotype threat is defined as a situational predicament in which people are or feel themselves to be at risk of conforming to stereotypes about their social group (Steele & Aronson, 1995). Twenty-five years have passed since the term stereotype threat was first conceived in Steele and Aronson's original article (Steele & Aronson, 1995) now referred to as the modern classic in the field (Devine & Brodish, 2003). In this pioneering experiment, they demonstrated that African American participants underperformed in a verbal reasoning test relative to their

White peers when it was suggested the test was a diagnostic test of their ability. By controlling for ability across all participants, Steele and Aronson could attribute this underperformance to the African American student's vulnerability to judgement about their groups' lower intellectual ability. This is the phenomenon now popularly known as *stereotype threat*. Since then, we have come a long way. The field of stereotype threat research has attracted so much attention, and the term stereotype threat has been used to unwrap many complex issues. It has been found that even passing reminders that someone belongs to one group or another, such as a group stereotyped as inferior in academics, can wreak havoc in test performance (Walton & Spencer, 2009).

Stereotype threat is a social identity threat. It occurs when individuals perceive their social group to be devalued by others (Major & Crocker, 1993). Stereotype threat is situational. It occurs in situations where negative stereotypes can arouse fear of conforming to that stereotype among group members (Steele, 1997). The implication of deeming stereotype threat as situational would mean that it would be feasible to mitigate the threat by altering the situation. Stereotype threat is general, it is experienced by everybody. It is not limited to social groups about whom negative prejudices already exist. For example, in a study it was reported that Caucasian men, a group that has a relatively positive social status, underperformed in math when they were told that their performance will be compared with Asian men (Aronson et al.,1999). Very similarly, it was reported that White men underperform in tasks that require athletic ability when told that Black men have superior athletic ability as compared to White men (Stone, 2002). This threat is especially frustrating because it affects even those who have the right skills and self-confidence to complete a related task. It arises not only from the process of "internationalization of the stereotype" but also from other factors like identification with the domain and the resulting concern of being stereotyped in it (Steele, 1997). This argument that underperformance is not only a result of internal doubts but also a result of external factors like domain identification, has the implication that one should not only focus on correcting internal psychology but also on mitigating the situational threat which arises due to factors not in the control of the individual. This argument that stereotype threat gets activated through a set of social psychological phenomena both at the individual level and at the situational level forms the basis of our study.

The opposite of stereotype threat is stereotype boost which is when people perform better than they otherwise would have because of exposure to positive stereotypes about their social groups (Dijksterhuis et al., 1998; Shih et al., 1999). For example, in United States, women are often stereotyped as having inferior quantitative skills (Hedges & Nowell, 1995) but are often acknowledged for their superior verbal skills (Shih et al., 2006). On the contrary, Asians are known for the superior quantitative skills (Steen, 1987; Trytten et al., 2013) but have been stereotyped for their inferior verbal skills in English (Shih et al., 2006). In line with the above argument, it was found that Asian American women performed better on a math test when their Asian identity was primed (priming is a process of exposing an individual to something that influences their behavior later), but worse when their female identity was primed, when

compared to a control condition where no social identity was primed (Shih et al., 2012). Stereotype boost can result in performance boosts by exposing individuals to positive stereotypes

Stereotype threat theory and stereotype boost theory run parallel to each other. While somewhat similar they are also very different processes. By no means can stereotype boost be used as a method of mitigating stereotype threat. Stereotype boost only occurs when certain boundary conditions are met, and these can hinder performance when they are introduced in ways that elicit social comparison processes (Shih et al., 2002). While stereotype boost is important in its own way, we choose to focus only on stereotype threat for this research because we are more concerned about the negative rather than positive effects of stereotypes. Historically, there has always been more focus on negative stereotypes because psychologists, policy advocates and academicians have always attempted to bridge the achievement gap between minority and non-minority students. In this context, we engage in understanding and acknowledging the severity of stereotype threat and finding ways of measuring and mitigating it. While there have been attempts in the past to understand stereotype threat, there hasn't been any accepted standard for measuring it (Shapiro & Neuberg, 2007; Picho & Brown, 2011).

Moderators of Stereotype Threat

While there has been agreement on the fact that stereotype threat impairs performance, there are a lot of lingering questions with regards to what mechanisms and processes underlie these effects (Schmader et al., 2008). From a methodological perspective, stereotype threat emerges in tasks that are difficult and demanding and the extent to which the task is perceived to be difficult can be moderated by individual characteristics. This makes the task, the individual and the environment in which these two are situated the universe of stereotype threat. Segments and pieces of the stereotype threat universe have been broken down by various researchers and studied extensively.

The fundamental theory behind stereotype threat is that it is unique. It does not have similar effects on every individual of a stigmatized group. Research has identified numerous moderators that make some individuals more susceptible to and some tasks more likely to elicit stereotype threat (Shapiro & Neuberg, 2007). A moderator variable (such as an individual's domain identification) which is characteristic to the individual may influence the strength and direction of stereotype threat. In this section, we will discuss some key moderator variables that have been identified as critical in the context of stereotype threat.

Stigma Consciousness. One criterion that contributes towards differences in the severity of threat among individuals is their awareness, consciousness and belief in the stigma ascribed to the social group (Brown & Pinel, 2003; Schmader et al., 2004). This awareness or consciousness is the belief that one would be seen in the context of the negative stereotypes about one's group rather than being judged just on one's own behavior (Daley & Schlichtmann, 2018). The extent an individual believes and is aware of this stereotype would lead to varying levels of

expectations about how the individual would be perceived by others. This leads to non-uniform levels of stereotype threat among individuals belonging to the same stigmatized group. Stigma consciousness has been defined as the "probability of being stereotyped" (Pinel, 1999). It has also been noted that high levels of stigma consciousness increase an individual's vulnerability towards stereotype threat (Pietri et al., 2017; Brown & Pinel, 2003). For stereotype threat to occur, individuals must experience some concern about being judged stereotypically (Steele, 1997). In a study by Schmader and his colleagues, it was found that women who believe men to be superior in math tend to evaluate their own math ability using only women as a basis of comparison (Schmader et al., 2004). On the other hand, women who do not hold such prejudices tend to take pride in their accomplishments, especially when they disconfirm the stereotype. (Schmader et al, 2004). Spencer and his colleagues have shown that, on average, women underperform relative to men on a math test when told that the test had shown gender differences, but, on average, they perform as well as men when told that the test did not show any gender differences (Spencer et al., 1999). A more recent study by Cadaret and his colleagues highlight stigma consciousness as an important variable to identify vulnerability and influence of stereotype threat on women in engineering (Cadaret et al., 2017). Pinel has documented some of the consequences of being acutely aware of the stigma associated to one's group. For example, people with higher levels of belief in the stereotype about their social group would tend to interpret ambiguous negative feedback more personally than those with lower levels of belief. (Pinel, 1999).

Collective Identity - Stereotypes usually pertain to a group of individuals. For example, the females, or the Americans, or the youth, or professors in academia – such groups are constantly being labeled with preconceived terms and notions. It has been noted that people not only worry that their own behavior could be used to lend credence to a negative stereotype about their group (Steele et al., 2002), but they also worry that the behavior of a fellow group member could affect the way the group is viewed or perceived by others (Cohen & Garcia, 2005). This is defined as collective threat in the context of stereotype threat literature (Cohen & Garcia, 2005)**.**

Hence, at the very fundamental level, for stereotype threat to cause significant effect on the individual – the individual needs to have a sense of belongingness to the group under consideration. This construct is often studied as "group identity" or "collective identity" in stereotype threat literature (Davis et al., 2006; Marx et al., 2005). Collective identity deals with categorical membership (Ashmore et al., 2004). As described by Simon and his colleagues, it is a place in the social world. (Simon & Klandermans, 2001). The association is more about a psychological positioning and less about physical contact (Ashmore et al., 2004).

Cross was among the pioneers of the development of collective identity (Cross, 1991). He proposed the racial identity developmental model and referred to the stages of development as statuses. Statuses are mindsets through which individuals go through (Helms, 1995), starting from the status of denigration (e.g. "I am not Black") to unquestioned acceptance. Collective identity is a powerful concept as it captures not only the peripheral association of an individual

with the group under consideration, but also encompasses the value and emotional significance one attaches to the group (Tajfel, 1981). Collective identity affects behavior, choice of language, choice of peers and significantly shapes our lives. Therefore, it is not surprising that it has been extensively studied in the context of stereotype threat. It offers explanations for why women who indicate gender to be central to their identity tend to perform worse in math than women who do not indicate gender to be crucial to their self-definition (Schmader, 2002). It offers explanation for why higher levels of racial identification buffer African Americans against self-esteem and social threats (Oyserman et al., 2001). But different researchers have looked at different aspects of collective identity and there is a need to capture a broad overview.

The most basic element of collective identity is self-categorization (Ashmore, 2004). At the very least you would expect individuals who associate themselves to a group to place themselves in that group. It is a natural human tendency to place oneself in a group based on commonalities, perceptions and shared interests. While this sounds like a simple process, there are psychological implications of this choice. Group classification can cause favoritism, loyalty and bonding which can be useful as well as detrimental. Even meaningless labeling into groups can cause prejudice and preferentialism among individuals. As Tajfel and his colleagues demonstrated in an experiment, mere classification of participants into two arbitrary groups, through a random assignment process led to in-group favoritism (Tajfel, 1981).

Individuals are constantly choosing groups and categorizing themselves in multiple levels of these groups. As an example, Asians living in United States categorize themselves as Asian Americans. This brings us to Shih's proposition where she highlights that in the real world, individuals carry multiple identities. Stigmatized individuals can draw support from some other identity to protect themselves from the effect of stigma on one identity (Shih, 2004). Through this process of identity switching, individuals who carry multiple identities can protect their psychological well-being. Identity switching also leads to greater life satisfaction (Thoits, 1986; Rydell et al., 2009)

Domain Identification. Another criterion for stereotype threat to occur is that individuals must strongly identify with the domain or task under consideration for stereotype threat to have any effect at all (Steele, 1997). The theory of domain identification is embedded in the symbolic interactionist perspective of the self (James, 1981; Serpe & Stryker, 2011). The interactionist perspective states that individuals perceive and interpret the various forms of feedback they receive from the environment. If they receive domain specific feedback they incorporate that in the domain-specific self-concept. The extent to which the domain is viewed as important by the individual will affect the extent to which this domain self-concept affects their self-esteem. (Osborne & Jones, 2011). For example, for students to do well in school, they must have a sense of belonging and identification with school achievement, i.e. it should be an integral part of their identity. Doing well should matter. The global society is structured in a way that lower expectations from certain groups (such as females) results in disengagement, disidentification and lower self-expectation resulting in a weak sense of belonging. Therefore, there are

differences in performance among groups, such as we see between males and females in math, despite them having similar ability (Smith & White, 2001).

In an experiment by Tesser and Campbell it was found that individuals can fluctuate their self-esteem by increasing and decreasing the importance of domain. (Tesser & Campbell, 1980; Smith & White, 2001) This explains why individuals go through engagement and disengagement with multiple domains. Disengagement occurs when individuals intentionally distance themselves from a domain to protect their self-esteem from potential consequences of performing poorly (Major & Schmader, 1998).

There has been a lot of research to understand school dropout rates. This is because disengagement in school as an outcome is very difficult to measure, hence researchers use dropout rates as a coarse proxy. Dropping out is a process where individuals disidentify with the academic domain, usually in the pursuit of identification with some other domain. But there are studies to show that the decision of drop out is taken long before a student drops out (Finn, 1989; Griffin, 2002). Therefore, studies have found little difference between self-esteem of students who drop out and students who do not drop out as the interest in a second domain offers protection against reduced self-esteem caused by disinterest in the first domain (academic domain). There is no rule regarding the optimum number of domains one should be interested in, but it is good to have multiple domains of interest to maintain a stable self-esteem (Osborne, 2011). Focusing on one singular domain is not healthy from a psychological point of view. Domain identification varies from individual to individual as well as within individual over time (Rosenberg, 1979). There are studies to show that domain identification has similar effects on outcomes in both adults (Tesser, 1988) as well as children (Crocker and Major, 1989).

Thus, we can argue that susceptibility to stereotype threat not only lies in the internalization of the stereotype but also in caring about the domain. Performance of stigmatized individuals in a domain that they care about, under the influence of stereotype threat, declines. Thus, domain identification proves to be a critical moderator in the context of stereotype threat.

Core Strength. Efforts to make the environment more equitable are important to mitigate the harmful effects of stereotype threat, but how individuals respond to prejudices in the environment to protect themselves also affects the extent to which they experience stereotype threat. According to social mentality theory, stigma is a social threat that challenges a stigmatized individual's social ranking, leading them to feel inferior than others (Birchwood et al., 2007). This perception leads to feelings of internalized shame which is a subcomponent of internalized stigma (Barney et al., 2010).

Stereotypes have two sides to them, they are not just a function of the society nor of the individual. A clear distinction is provided by Corrigan and Watson in their research where they describe public stigma as the judgements that society places about the individual and self-stigma as the degree to which the individual internalizes these judgements (Corrigon & Watson, 2002).

Many individuals in society experience chronic stigma, yet they can achieve their goals. Stigmatized individuals can have the resources to handle the psychological stresses

activated by stereotype threat. Despite the harmful effects of stereotype threat, there are many stigmatized individuals who have achieved success in their respective domains of interest (Miller & Kasier, 2001) by developing emotional stability, high self-esteem, resilience and the right mindset. These attitudes have been collectively described by Judge and his colleagues as *core self-evaluation*, defined as a collection of fundamental beliefs that individuals hold about themselves (Judge et al., 1998). For this research, we choose to call this collection of fundamental beliefs as, an individual's *core strength*. It encompasses an individual's inner power to mitigate the harmful effects of stereotype threat. The fundamental aspects of individual's core strength are discussed below.

Self- esteem, defined as the overall value a person places on herself/himself (Harter, 1990) is a central piece of a positive self-evaluation. It encompasses an individuals' self-liking, self-acceptance and self-respect. Self-esteem demonstrates short-term fluctuations but long term stability (Costa & McCrae, 1994). People with high self-esteem have more stable and consistent views about themselves than those with low self-esteem (Baumgardner, 1990). Self-esteem has been proven to mediate the relationship between experienced and perceived stigma and the personal impact of stigma (Vass et al., 2015). Believing the stereotype regarding oneself to be true leads to a decrease in self-esteem which leads to emotional instability and distress (Watson et al., 2007). Sometimes emotional distress can be severe leading to depression, anxiety and psychiatric symptoms.

Self-efficacy is the belief in one's competence. It is defined as the individual's perception regarding his/her ability to perform across various situations (Chen et al., 2000). In a study by Hoyt and Blascovich, it was found that after negative stereotype activation, women who reported high leadership efficacy demonstrated better performance, greater domain identification and increased well-being, relative to those who reported low leadership efficacy (Hoyt & Blascovich, 2007). Self-efficacy and self-esteem differ with respect to their relative emphasis on motivational v/s affective components (Chen et al., 2004). While self-efficacy captures the self-perception regarding task capabilities, self-esteem emphasizes more on feeling of self (Betz & Klein, 1996). Individuals with high self-esteem in general have a positive view about themselves, whereas individuals with low self-esteem hold a negative view about themselves despite having high self-efficacy (Brown, 1998). Self-efficacy and self-esteem are so strongly correlated that it is difficult to distinguish between them (Eden & Aviram,1993).

Resilience is an important resource to mitigate any form of stress (such as stereotype threat) among individuals (Balgiu, 2017). Individuals with low self-esteem are also low in resilience (Brockner, 1979). Self-efficacy is also related to resilience. In a study by Graham: it was found that Black students report lower self-efficacy in environments that are predominantly White, however those with higher resilience could maintain their self-efficacy even in situations typically associated with stereotype threat (Graham, 1994). Shih analyzed social stigma within the framework of resilience. The resilience framework is the key to understanding how people bounce back in life despite challenges (like stereotype threat) experienced during the various stages of life. There have been many studies that look at the association between stigma and

resilience. Crowe and his colleagues describe the relationship between stigma and resilience as bi-directional. Building resilience helps decrease stigma at the same time exposure to stigma decreases an individual's ability to be resilient (Crowe et al., 2016).

Dimensions of Stereotype Threat

Some scholars have taken a multidimensional approach to understanding stereotype threat. According to the multi-threat framework (Shapiro & Neuberg, 2007), stereotype threat has two dimensions, (a) the target of threat and (b) the source of threat. The target of threat could be the individual or the in-group. (i.e. is the stereotype applicable to one's personal or social identity). The source of threat (i.e. who will judge the performance) could be the in group, the out-group or self. This leads to six qualitatively distinct stereotype threats that manifest through the intersection of the above mentioned two dimensions. Individuals can experience both self or group based threat depending on the environment.

One framework proposed by Schamder and his colleagues highlights that stereotype threat involves activation of three core concepts: (a) the concept of one's belonging in group, (b) the concept of the ability domain in question, and (c) one's self-concept (Schmader et al., 2008). Borrowing from Gawronski & Bodenhausen's work on propositional processes, they further highlight that what matters is not just the mere activation of these concepts, rather the activation of a specific propositional relation between these concepts. A positive unit relation would indicate that the concepts are defined in the context with respect to one another. For example – for a Black student taking a math test, a positive unit relation would mean the student believes that he/she has sufficient math ability, identifies himself/herself as Black and believes that Black students can perform well in math. A negative unit relation would mean some of the concepts are defined in opposition to one another. An example of a negative unit relation for the same scenario (a Black student taking a math test), would mean the student believes that he/she has sufficient math ability, identifies himself/herself as Black but does not believe that Black students can perform well in math. This creates an imbalance in the relational structure (Heider, 1958).

Stereotype threat is essentially an outcome of this imbalance. The effort to move from an imbalance state to a balance state, for example, by outperforming in case of a math test, can be very hard. These imbalances are analogous to what has been described in research as stress, frustration, inferior anxiety and even at times depression (Kobrynowicz & Branscombe, 1997; Major et al., 2003). This environment of tension is created due to the simultaneous activation of three above mentioned processes and the links between them. This theory can also explain why some women tend to disassociate themselves with math (Shih, 2004) – to achieve a state of balance. While some experience emotions of "trying too hard" which also interferes with their performance (Steele, 1997). This also explains why there is such a huge variation in the severity of threat experienced by different individuals.

Measuring Stereotype Threat

There is no existing measurement tool that has been accepted as the standard for measuring stereotype threat (Shapiro & Neuberg, 2007; Picho & Brown, 2011). A phenomenon like stereotype threat which is so ubiquitously present needs to be operationalized in a way that is practical. The measurement of stereotype threat is traditionally and most commonly based on an experimental design, usually involving a small sample size, where the treatment involves exposure to statements (prompts) about the negative stereotypes of the stigmatized group which is intended to activate self-doubt and performance anxiety, reducing the ability to perform (Steele & Aronson, 1995). The mean of this group is compared to the mean of a randomly allocated control group. Stereotype threat is then operationalized as the difference in outcome between the treatment group and the control group in an achievement test after student ability has been controlled for. Many studies replicating the experimental situation described above have failed to obtain similar results (Aronson et al., 2002). Relying on experimental outcome data to demonstrate the existence of stereotype threat raises the question of the underlying cause of underperformance (Osborne, 2001; Smith, 2004) and of how the construct can be pragmatically measured. There is a need to move from this experimental based approach of measuring stereotype threat towards defining stereotype threat as a measurable construct.

Moreover, if stereotype threat is ever to be a useful concept in applied areas such as education, it must be operational at the individual level, which is not possible in the traditional approach. The process of defining stereotype threat as a measurable construct is complex. There are ethical issues involved with regards to whether stereotype threat can be intentionally introduced in an environment. But on the other hand, when we think about the environment as it is today, it is not free of stereotypes, biases and inequalities at various levels. The construct needs to be operationalized in a way that it captures or controls for these pre-existing differences rather than inducing them.

There are quite a few measures that have been developed to measure stereotype threat more directly as highlighted by Xavier and his colleagues (Xavier et al., 2014). A careful examination of those reveals that they are targeting a myriad of constructs such as stigma consciousness, knowledge of stereotypes, stereotype endorsement and many more. Stigma consciousness refers to the extent to which a person expects to be stereotyped (Pinel, 1999). The extent to which an individual knows about the negative perception of his group is referred to as knowledge of stereotypes (Casad & Bryant, 2016). Stereotype endorsement is the extent to which someone ascribes certain traits to members of groups (e.g. - women are nurturing and men don't cry) (Schmader et al., 2004). High levels on these constructs are expected to increase one's vulnerability to stereotype threat. Stereotype threat while related to these constructs, is also somewhat distinct. These constructs also have high amount of variability. While some researchers have used 2-3 items per instrument, some have used 8-10 items. The reliability estimates of these measures range from 0.63 to 0.91 (Xavier et al., 2014).

One critical issue with these previous attempts at measuring stereotype threat is the failure to account for item-wise differences. Consider a respondent who experiences very high agreement across two of the survey items compared to another respondent who indicates moderate agreement across two items. Do they both experience a similar amount of stereotype threat? This answer is not addressed by developers of these instruments.

As Shapiro and Neuberg have mentioned, acknowledging different sources of threat leads a multidimensional perspective to stereotype threat (Shapiro & Neuberg, 2007). It is important to be able to differentiate between the target of threat - group and self, and the source of threat - in-group, out-group and self. This approach leads to six possible combinations of threat. Many studies in the past have failed to differentiate between these dimensions. As an example, Marx and his colleagues in their instrument to measure stereotype threat have used items such as "I worry that my ability to perform well on math tests is affected by my gender" which is self-focused and "I worry that if I perform poorly on this test, the experimenter will attribute my poor performance to my gender" which is group focused (Marx et al., 2011).

Shapiro and Neuberg further highlight that it is possible that at any given point of time, an individual may experience multiple threats arising from these various sources (Shapiro & Neuberg, 2007). An instrument is needed that can isolate these different sources of threats in a reliable and valid manner. One way to do that is to focus on a single source and a single target. Another way to do that is to embed the source and target within the item itself. This differentiation hasn't been successfully captured to date.

Most studies that have attempted to measure stereotype threat and estimate its effect on a stigmatized population have looked at academic outcomes. Typically, members of a stigmatized group are exposed to stereotypes about their group and are compared with members of a non-stigmatized group in an academic achievement test. While this research is important, academic outcomes cannot be the only standard by which stereotype threat is evaluated. Not all individuals identify with academic pursuits. Even for those who identify with academic pursuits, the assumption that an achievement test is a valid and reliable measure of the outcome is flawed and needs to be addressed.

As mentioned in the theoretical framework, stereotype threat while a function of the individual also depends on the task and the environment. For example, Steele and his colleagues in their laboratory based experiment used a white man as the experimenter (Steele & Aronson, 1995). Would the results have been different if they had used a black male or a white female as the experimenter? These issues need to be addressed. It is important to develop a clear formulation of the individual, the task and the environment and how these interact with each other.

In the previous sections, we mention various moderators of stereotype threat – the extent to which individuals are (a) aware of the stereotype, (b) identify with the domain under consideration, (c) consider themselves to be members of the stigmatized group (d) and have the internal strength to mitigate stereotype threat. These are all important pieces of the puzzle that need to be conceptualized and measured. While researchers have utilized sections of the

stereotype threat universe, we know of no such research that takes such a comprehensive view on stereotype threat.

Finally, measuring stereotype threat in a pragmatic way inevitably engages the challenges and drawbacks of self-reported measures (Devaux & Sassi, 2016). Using an instrument that mentions the stereotype too explicitly can result in individuals feeling uncomfortable and overtly conscious, which might result in misinformation.

Given the above challenges, the task of developing a reliable, valid and fair measure of stereotype threat is a tall order. Nevertheless, we believe that it is possible to understand and measure stereotype threat through a multidimensional and situational lens by carefully and correctly defining the construct and its constituents. The best way to move forward is to clearly understand the universe of stereotype threat, how the different concepts interact and exert influence and concisely lay down the underlying assumptions. We will attempt to develop this construct in the next chapter.

CONCLUSION

In this chapter, we reviewed and synthesized past research on stereotype threat. Stereotype threat which is generally thought to be omnipresent, can affect almost everybody because of their associations, preferences, choices and membership in groups. It is unique to an individual and unique to a situation. By causing anxiety, task-related worries and self-doubt, stereotype threat, as a short-term outcome, has the capacity to impair performance of an individual in a task and, as a long-term outcome, has the capacity to disengage individuals from a domain completely. While these effects can be life-altering for some individuals, there are some who are able to successfully shield themselves from the harmful impact of stereotype threat with the right mindset, self-confidence and overall ability to bounce back from negative experiences and disturbances. Typically, stereotype threat is triggered during tasks that are difficult and demanding which again is unique to an individual. The threat can be embedded in the environment in which the task and individual are situated. Thus, the task, the individual and the environment and their interaction form the key elements of stereotype threat.

Some key moderators that make individuals susceptible and tasks more likely to elicit stereotype threat have been discussed in literature. It has been established that individuals' level of awareness and belief in a stigma attributed to his/her social group can contribute to differences in severity of threat experienced. However, individuals might or might not have a strong identification with the social group under consideration. This difference in group identification, popularly known as collective identity, also causes differences in severity of threat. Individuals engaging with the task in a domain might not experience the effects of stereotype threat if they do not inherently have the desire to do well in the task or in the domain. This phenomenon is labelled as domain identification. Another variable that causes differences in stereotype threat experience is an individual's mindset, internal strength, confidence and the capacity to recover from difficult situations.

To understand the deleterious effects of stereotype threat, researchers in the past have engaged in the process of operationalizing it, yet, there is a lack of an accepted tool for measuring stereotype threat, possibly because of the complexity of the construct and the lack of a precise definition and clear boundary conditions. In the next chapter, we propose a solution to this situation by bringing all the key variables identified above under one umbrella, making efforts to precisely define and lay out their boundary conditions and conceptualize how they interact in a multidimensional space. This theoretical framework forms the basis of this research.

Chapter Two

Developing the Stereotype Threat Instrument, The Four Building Blocks Approach

In this chapter, we attempt to develop a measure of stereotype threat. We operationalize stereotype threat as the balance between (1) an individual's level of awareness and belief in the stereotype, (2) identification with social group under consideration, (2) identification with domain under consideration and (4) the individual's overall mindset and internal strength to mitigate stereotype threat. This lends a multidimensional aspect to the stereotype threat. We describe a four-phase iterative process of developing a generic stereotype threat measure that can be adapted to any stereotype situation or environment. Finally, we adapt the generic stereotype threat instrument for a specific group (transfer students) and domain (data science). The transfer students stereotype threat instrument will form the basis of data collection and validation processes that we would carry out in the subsequent sections.

INTRODUCTION

Assessment development is the art of designing and gathering evidence about the phenomenon represented in an assessment task. In psychology, the main purpose of using assessments is to gain insight into a person's behavior and attitudes. Assessments play a crucial role in this field and have practically become ubiquitous. Although numerous instruments are being developed constantly, researchers and assessment developers often fail to recognize the importance of sound measurement practice. Before any argument about the respondents can be presented, it is important to ensure the assessments we use to derive those conclusions are sensitive, accurate and meet specific standards. A sound measurement technique is one that uses high quality instruments that are closely aligned to the target construct and integral in capturing individual differences. Therefore, it is important to rely on well-established techniques and frameworks to develop assessments.

One such technique was proposed by Wilson known as the BEAR Assessment System (BAS; Wilson, 2005). In this chapter, we will explore this framework in the context of developing the stereotype threat instrument. We will first operationalize the stereotype threat construct based on the theoretical framework laid down in the previous chapter. Then, we will engage in the development of a generic stereotype threat instrument that can be adapted to different situations, stigmatized groups and domains. The strength of the generic stereotype threat instrument is that it can enable comparisons across domains, across groups or across situations, by controlling for one or more of the varying parameters. For example, if we want to compare the stereotype threat faced by females in a math test and a verbal test, we can successfully do so by using the same instrument and changing just the domain. Of course, the underlying assumption would be that the females taking the math domain stereotype threat instrument and the verbal domain stereotype threat instrument are similar in other aspects like demographics, intellectual ability, etc.

Finally, we will fix the design parameters that have been left unknown in the generic stereotype threat instrument to customize it for a specific group and domain. In this project, for the group, we choose the community college transfer students who transferred from a community college to a four-year university. We want to measure the extent to which these students experience stereotype threat when interacting with data science related courses or activities on campus which forms the domain. Let us now look at the transfer students group and the stereotypes associated with them.

Stereotype Experiences of Transfer Students

When we think about stereotype threat, we often link it with the group that forms the minority in an environment or context. In education, minority groups are more susceptible to stereotype threat which lowers their educational prospects (Espenshade & Walton-Radford, 2009). One such group of minority students are the community college transfer students studying

in a four-year university. In general, community college transfer students are those students who begin their college academic career at a community college, earn credit through completion of coursework and transfer to a four-year university to seek better academic prospects. Transferring from a community college to a four-year university is a complex process (Laanan, 2001). Students go through adjustments at all levels, be it psychological, academic or environmental (Lopez & Jones, 2016). In a study by Packard and his colleagues it was found that female STEM students who were transferring from a community college to a four-year university reported more positive experience when interviewed before transitioning than after the transition. Post transition they reported feeling stressed and not being able to cope with the pace of the courses (Packard et al., 2011).

When it comes to STEM related degrees, many transfer students are stigmatized as "latecomers to science" because they may not have interacted with science-related courses in community college (Jackson & Seiler, 2013). After transferring these students experience a disruption in their social and academic identities. This disruption tends to be more severe for students who are older, coming from marginalized racial and ethnic backgrounds and low income families (Crisp & Nunez, 2014). According to the Community College Research Center (CCRC), in 2018, the ethnicity of community college students in United States was 45% White, 25% Hispanic, 13% Black, and 7% Asian. Although White students form the majority, their population in community colleges has continued to drop over the past decade (CCRC Report, 2018). About one-third of the students in the same year were first-generation students. 67% of the community college student population came from families with household income of less than $50,000 (CCRC Report, 2018).

Hence it can be concluded that transfer students go through a lot emotionally, socially academically and personally as they transition to a completely new campus. While there are many stereotypes about the transfer student community, the most common stereotype mentioned in literature and even felt by so many transfer students in our study, is the feeling of not being smart enough. A transfer student goes through imposter syndrome, which is a feeling of self-doubt, questioning his/her academic abilities. This causes a decrease in academic output especially in science related fields. Throughout this study, we have had interactions with a lot of transfer students who have shared their experiences with us. In the words of a few students –

"We are poor, incarcerated, dropouts, less talented academically. It feels okay, much of this is true! And not all of it is bad. For example, kudos to criminals for going to school! No shame to that."

"Most people assume that being a transfer student is synonymous with being dumb. Many speculate that transfers are people who couldn't get into college as a freshman due to bad brings. This mentality honestly does not phase me as it goes from a place of immaturity and ignorance. Being a transfer is something that should be celebrated, not stigmatized."

We will use our stereotype threat instrument to investigate the effect of the above-mentioned negative stereotypes in the context of the transfer student community (collective identity). We choose data science as our domain (domain identity). The choice of domain is somewhat arbitrary as the instrument is designed to investigate stereotype threat in any domain, however, since data science is a science related field that attracts students across various disciplines, we felt it would be an interesting first context in which to measure stereotype threat of transfer students engaging in a data science course or activity on campus.

Stereotype Threat – The New Construct

The term "construct" derives its name from the word "construction" which means to build something. In the context of psychology, a construct is the ontological form given to an abstract concept or phenomenon. It is the first step towards developing a measure. A construct is always hypothetical and can comprise of multiple sub-constructs or dimensions. These sub-constructs are somewhat distinct but they could be correlated.

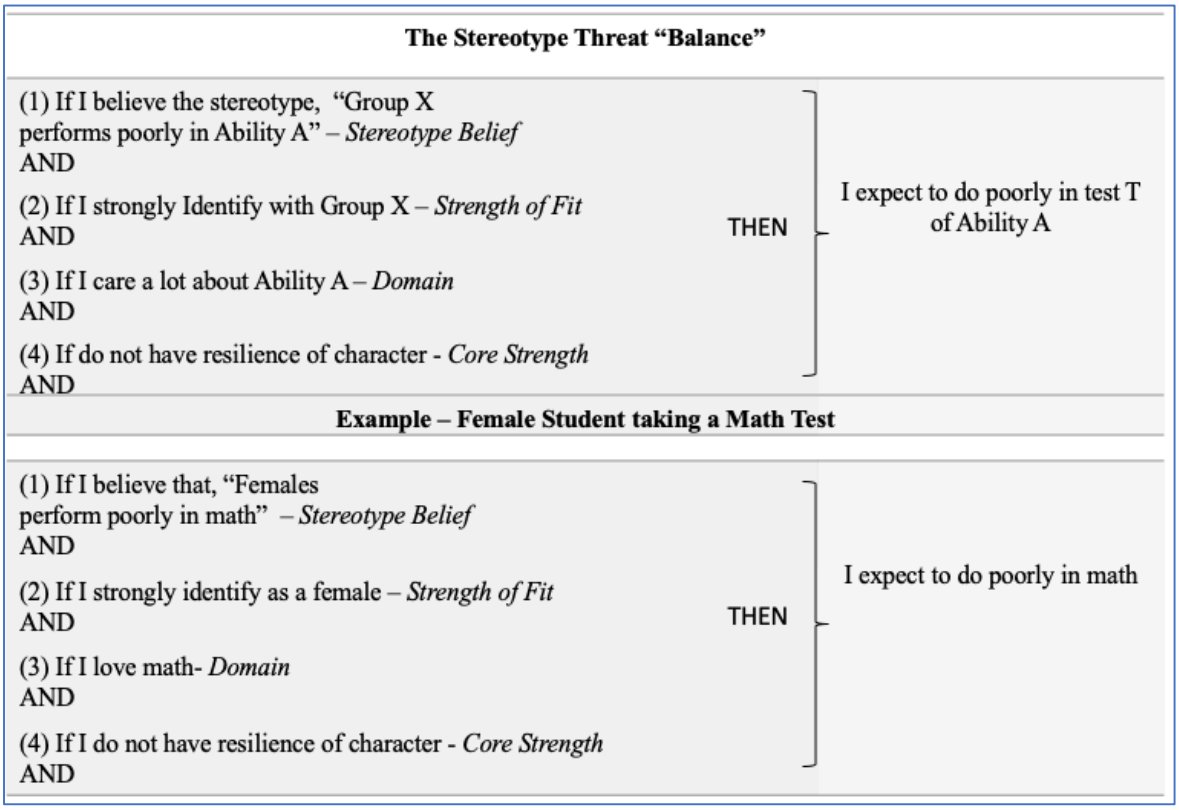
The first step towards measuring stereotype threat is to comprehend the construct that needs to be measured. The construct we are attempting to measure is "stereotype threat." In the previous chapter, we discussed some of the key concepts that are related to stereotype threat. We use those concepts to develop the construct. Every construct needs to begin with a clear definition. We define the construct of stereotype threat as: a measure of imbalance between an individual's group identification, domain identification and self-concept (Schmader, 2002). Along with these three concepts, a fourth factor that contributes towards the extent to which an individual experiences stereotype threat is his/her awareness or consciousness with regards to the stereotype. Thus, we hypothesize stereotype threat to be affected by the following sub-constructs – (1) Belief in the stereotype of the social group under consideration - *Stereotype Belief Construct*, (2) Identification with social group under consideration – *The Strength of Fit Construct*, (3) Identification with domain under consideration – *The Domain Construct*, (4) Evaluation of an individual's self-esteem and inner strength to mitigate the harmful effects of stereotype threat: *The Core Strength Construct*. This makes stereotype threat a multidimensional construct. We will engage in describing the construct more extensively in the subsequent section.

Before we do that, we also need to establish how these latent variables or sub-constructs interact with each other. Let us understand this with the help of an example as depicted in Figure 1. Figure 1 (at top) depicts how the four theorized sub-constructs contribute towards the overarching stereotype threat construct. As an example, we consider an individual who is a member of a group about whom a negative stereotype exists which we hypothesize as directly affecting an individual's ability to perform well in each task. Specifically, we hypothesize that an individual belonging to Group X and having ability A tends to expect to do more poorly in a test T of ability A according to the extent he/she has more of– (a) belief in the stereotype about Group X (b) identification with Group X (c) care about doing well in ability A and (d) lack core strength to combat stereotype threat. We hypothesize that, all else being equal, this would hold

true for any given situation and any given source of stereotype threat. One underlying assumption is that the individual believes the test T is a true measure of ability A.

Figure 1

The Logic of Stereotype Threat

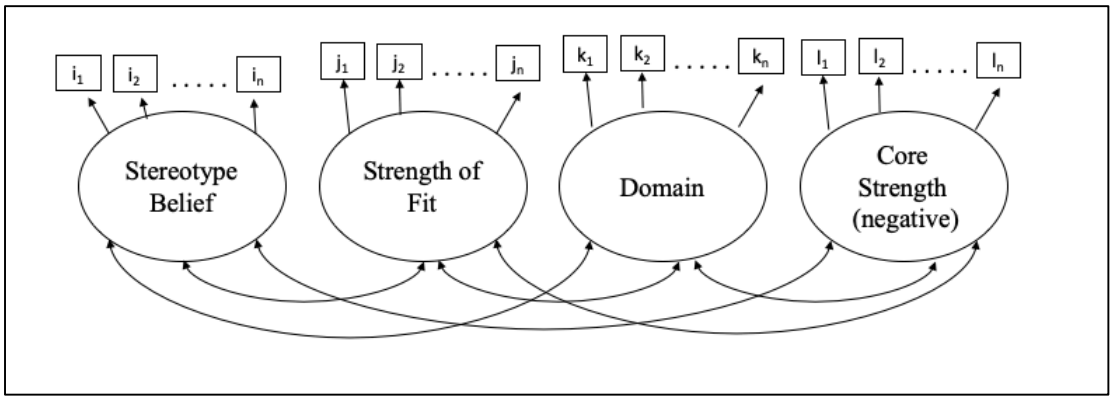


This is exemplified by assigning a specific group, domain and task to an individual (see Figure 1, at bottom). If a female student is taking a math test, she would expect to do more poorly in math more if she (a) believes in the stereotype that "females perform poorly in math in comparison to males" (b) strongly identifies as a female (c) loves math (d) does not have the resilience of character. Again, these parameters should be true for any given domain, group and task.

The features of the construct can be graphically represented as shown in Figure 2. The graphic representation shows the four sub-dimensions as latent variables (in ellipses). The curved lines represent the correlations between them. Each latent variable is mapped onto items (i_1 to i_n , j_1 to j_n , k_1 to k_n , l_1 to l_n for the four dimensions respectively) which aids in the operationalization of the construct, using items (represented in boxes).

Figure 2

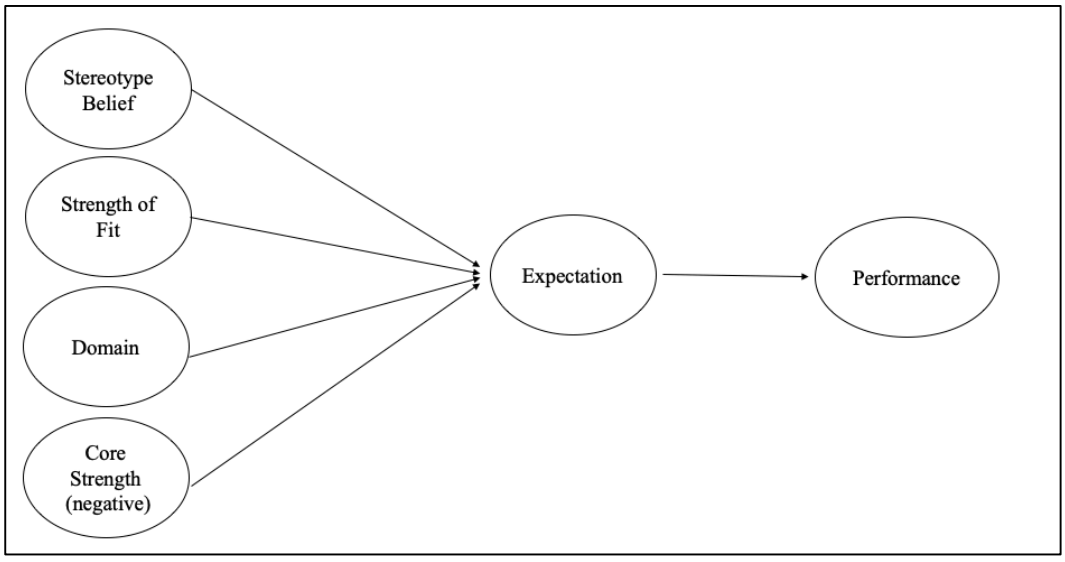
A Psychometric Representation of the Stereotype Threat Construct



The way this relates to the traditional paradigm for detecting stereotype threat is through performance on a test. The hypothesis is that the amount of stereotype threat present in an individual or the respondent will impact his/her performance expectation which in turn impacts the performance on the item representing that latent variable of ability. This relationship between the construct, expectation and performance is represented in Figure 3. This illustrates a composite model (Wilson & Gochyyev, 2020). In this study, we do not intend to study the relationship between expectation and performance but provide the diagram for completeness.

Figure 3

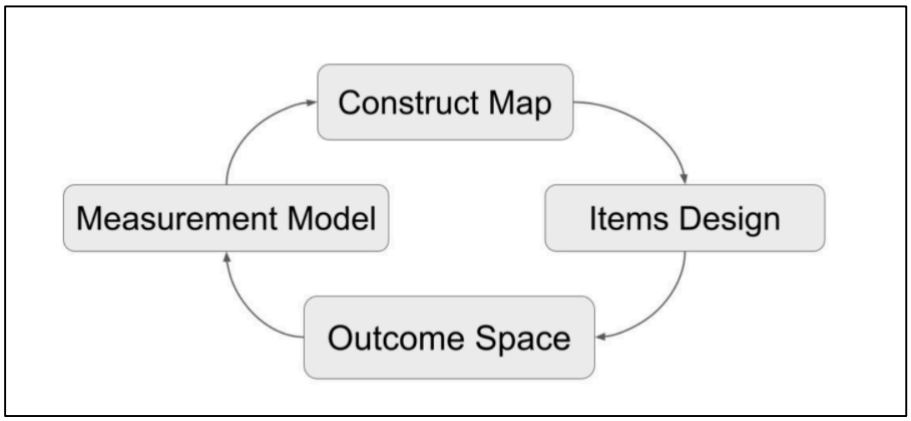
Relating Stereotype Threat to Expectation and Performance



Instrument Development

We use the BEAR Assessment System (BAS) to develop the stereotype threat instrument. The development of BAS builds on more than three decades of research in the design of assessment tasks and measurement techniques to support evidence-based assessment following a construct modeling approach. It uses four key principles to guide assessment development, (a) defining the construct that we are interested in assessing (b) creating items consistent with the construct (c) defining the relation between possible item responses and the construct, using an outcome space and (d) analyzing the resulting scored data using measurement models. As shown in Figure 4, these building blocks represent steps in a cycle of development, which should iterate several times to refine the construct maps, the scoring guides and the items. Let us discuss each of the elements of the BAS framework separately in the context of our instrument.

Figure 4
Four Building Blocks of Measurement



Building Block One: The Construct Map

A construct map is a foundational feature of BAS. It is an explanation of the theory or the construct. It is a graphical representation of how the construct develops as a continuum. It provides an ordering of qualitatively different points along the construct of interest, focusing on one characteristic, derived in part from research and in part from professional judgments about what constitutes higher and lower levels of the behavior, attitude, performance or competence. Thus, generating a construct map requires articulation of what represents various stages along the progression of the construct. We also need to specify the specific indicators at each stage that would enable the assessment developer to place respondents at different locations along the continuum. Construct map also serves as a guide for writing test items and evaluating responses.

We initially develop the stereotype threat construct map based on theory and use empirical evidence to validate the construct map. Developing a construct map is an iterative process and empirical findings can also be used to modify the construct map. The construct or the latent variable that we are attempting to measure is "stereotype threat" which we defined in the previous section as comprising of four sub-constructs. Now let us look at each sub-construct separately.

The Stereotype Belief Construct

We define the Stereotype Belief Construct as the extent to which individuals are aware, conscious and believe in the negative stereotypes about their social group. These differences in belief among individuals would lead to varying levels of expectations about how the individual would be perceived by others. It would also lead to differences in the extent to which individuals engage in self-protecting strategies like blaming the circumstances in the face of failures or setbacks.

We present the stereotype belief construct map in Fig 5. We specifically focus on negative stereotypes because as mentioned in the previous chapter, we are concerned with the measurement of stereotype threat which arises due to negative prejudices. Stereotype boost is not captured in this construct.

While some individuals believe that discrimination is always "out there" and it is hard to change anyone's point of view, others do not attribute their failures and weaknesses to their group membership. The latter is especially true when these individuals remain isolated from their in-group members for a long period of time and have very few occasions to reflect upon their stereotyped status (McGuire & McGuire, 1981). We use the above argument to develop the stereotype belief construct and categorize individuals based on the amount of belief they place in the negative stereotype under consideration. The hypothesis of the construct starts at the lowest level (Disbelief) where we place individuals who do not believe in the negative stereotype and do not interpret their weaknesses in the context of their membership in the stigmatized group. At the intermediate level (Partial Belief) we place those individuals who have some belief that being a part of a group negatively influences how people think about them. They sometimes feel being judged but do not always look upon their group membership as a source of negative discrimination. At the highest level (Belief) individuals are chronically aware of the stigma associated with their group and believe that it strongly influences their life and what people think about them. They believe that discrimination and stigma is always going to be "out there" and it is hard to change the society's view point.

Figure 5

Stereotype Belief Construct

Levels	Description	Example Items
Belief	Individuals are chronically aware of stigma associated with their group and believe it to be true. They believe that stigma is always going to be omnipresent and it is hard to change the society's point of view	I almost always feel myself to be a victim of the stereotypes that are associated to the transfer student community. People from other groups almost always interpret my behavior based on me being a transfer student
Partial Belief	Individuals have some faith and belief that being a part of a group influences how people think about them but do not always look upon their group membership as a source of discrimination	I often feel that I am a victim of the stereotypes that are associated to the transfer student community. Some people judge me based on my transfer student status
Disbelief	Individuals do not interpret their weaknesses in the context of their membership in the stigmatized group.	I never feel the stereotypes that are associated to the transfer student community to be also true about myself My being a transfer student does not influence what people think of me

The Strength of Fit Construct

We define Strength of Fit construct as the extent to which individuals feel a sense of belongingness to the social group under consideration. Individuals differ with regards to the need to affiliate with and be accepted by members of the group. While some individuals want the attention and support from members of their social group and are willing to provide the same attention and commitment to other group members, there are some who feel disconnected or prefer to stay disconnected, either because of negative feelings about the group or because of affiliations to other groups. We capture these differences through this construct.

We present the strength of fit construct map in Figure 6. Research indicates that those individuals who identify strongly with the social group are more susceptible to stereotype threat. Hence, through this dimension we study an individual's collective identity and the extent to which he/she identifies with the social group under consideration. Collective identity deals with categorical membership (Ashmore et al., 2004). It is based on shared ideas, thoughts and characteristics. This association is psychological and does not require any contact - it requires a

psychological positioning (Ashmore et al., 2004) and an acceptance by the person (Deaux, 1996).

Figure 6

Stereotype of Fit Construct

Levels	Description	Example Items
Over Fit	Individual idolizes the group is a way that is detrimental to the society or group or individual.	I feel that everyone should want to be a transfer student
Complete Fit	Individual attaches a high degree of importance to the group, is a proud member and feels emotionally invested in the group. Individual has a strong sense of belonging and engages in responsible action that positively impacts the members of the group.	I am a proud transfer student Being a transfer student is an important part of my self-image
Moderate Fit	Individual has a positive attitude towards the group, acknowledges his association publicly, attaches a certain degree of importance to the group, and is emotionally attached.	I consider myself to be a transfer student I do not care who knows I am a transfer student
Low Fit	Individual at the very least declares to be a member of the group, has somewhat positive feelings for the group, comfortably acknowledges his association in private but not so much in public, is not emotionally attached to the group and does not use his identity frequently in daily activities.	I am not proud about being a transfer student I am not comfortable with anyone knowing I am a transfer student
Negative Fit	Individual is hesitant to call himself a group member, has a negative attitude towards the group and feels embarrassed to acknowledge his association to the group in public.	I wish I wasn't a transfer student I sometimes have negative feelings about being a transfer student

We adapt Ashmore's conceptualization of collective identity for the construct map. According to Ashmore a collective identity should be comprised of the following elements - (a)

self-categorization- to what extent is the individual able to identify with the group (b) *positive association* - once the identification has been established, does the individual have a positive feeling towards the group (c) *importance* - in a world where we carry multiple identities, what is the degree of importance an individual attach to this group? (d) *emotional attachment* - does the individual feel emotionally invested and have a sense of belonging to the group? (e) *social embeddedness* - how often does the individual use the group identity in their everyday life? (f) *behavioral involvement* - to what degree does the individual engage in action that directly implicates the collective identity? (g) *cognitive awareness* - the degree of knowledge that an individual has of a group that directly affects his/her identity. These six elements have an interdependent relationship, for example - a higher sense of belonging will lead to an individual using the group identity more often and it would most definitely mean that he/she acknowledges his/her membership in the group. Thus, rather than assess each separately, they are assessed as parts integrated into a single construct.

The hypothesis of the strength of fit construct starts at the lowest level (Negative Fit) where individuals have a negative attitude towards the group and are hesitant to acknowledge their association with the group in public. At the next level (Low Fit) we place those individuals who at the very least are willing to declare their association with the group in public and have some positive feelings for the group. At the next level (Moderate Fit) are individuals who have a positive attitude and a sense of belonging to the group. At the next level (Complete Fit), individuals are attached to the group, have a very strong sense of belonging and bear a sense of responsibility towards the group. At the highest level (Over Fit), individuals idolize the group in a way that is detrimental to the society, the group or the individual.

The Domain Construct

We define the Domain Construct as the extent to which individuals form a relationship between themselves and the domain under consideration. This relationship is strongly influenced by their self-perceived competence in the domain and the need to feel recognized for that competence. This leads to individual differences in domain identity.

We present the Domain Construct in Figure 7. A sense of identification with the domain is one of the strongest predictors of an individual's performance. This construct encompasses (a) Domain Self Concept defined as individual's perception of his/her competence in a domain (Bong & Skaalvik, 2003) and (b) Domain Self-Esteem defined as how one feels about his/her domain self-concept. A domain self-concept of an individual would be "I love doing math", while a domain self-esteem would be "I am proud of my math skills". Outcomes from a domain would only contribute to stereotype threat if the individual identifies with the domain. Domain identification varies between individuals as well as within individuals over time. (Rosenberg; 1979). Individuals can manipulate their identification by switching on and off their domain identities (Tesser and Campbell, 1980; Crocker and Major, 1989; Shih, 2004). Early signs of student dropout from high schools can be detected by measuring a student's changes in domain

identification (Griffin, 2002). While domain identification is important for success, having a strong identification with only one domain in a lifetime is unhealthy and causes a lot of psychological stress in an individual's mind (Osborne; 1997). Everyone should be encouraged to identify themselves to multiple domains - the optimal number may differ for each person. (Osborne, 1997)

Figure 7

Domain Construct

Levels	Description	Example Items
Passion	Individuals are extremely passionate about doing well in this domain. Domain is strongly tied to their self-esteem and self-concept. They have high regards for their skills in this domain.	I am passionate about doing well in data science I think data science is a totally engaging field.
Attachment	Domain is tied to an individual's sense of self and becomes more centrally integrated to self-concept. The domain provides meaning to the individual.	I am motivated to do well in data science I think data science is very interesting.
Connection	Individual chooses to engage in activities related to the domain, persists to become better in the domain and are not disheartened by failures.	I am willing to put in expected hours to excel in data science I think data science is interesting
Indifference	Individual does not have positive or negative feelings about the domain.	I do not care much about being good in data science I think data science is a little interesting
Dislike	Individual actively distances his/her self-esteem tied to the domain. Individual disengages with the domain and establishes relationship with other opposing domains.	My skills in data science are poor I would rather spend time doing something else than doing data science I think data science is boring

The hypothesis of the Domain construct starts at the lowest level (Dislike) where the individual does not like the domain at all. They would rather spend time doing something else than engaging in any activity related to the domain. At the next level (Indifference), individuals

have a neutral attitude towards the domain. Their abilities in the domain do not form a central part of their identity. At the next level (Connection), individuals engage in activities related to the domain and make a positive attempt to master the domain. At the next level (Attachment) individuals are motivated to do well in the domain. Their domain identity is tied to their sense of self. At the highest level (Passion), individuals are extremely passionate about doing well in the domain. The domain is strongly tied to their self-esteem and self-concept. They have high regards for their skills in this domain.

The Core Strength Construct

We define the Core Strength construct as the extent to which individuals have the inner strength to mitigate the harmful effects of stereotype threat. Their ability to withstand difficulties and bounce back is strongly influenced by their sense of self-worth.

We present the Core Strength Construct in Figure 8. Through this construct, we seek to capture the fundamental beliefs that individual holds about himself/herself. Having a positive sense of self tends to help improve performance, however the opposite also holds true - performing well in a task may reinforce an individual's belief in his/her abilities. (Judge et al., 2007). While some aspects of an individual's performance are situational and contextual we believe that there are some traits that hold true across all situations. We capture those in the construct.

The hypothesis of the Core Strength construct starts at the lowest level (Chaotic). We place those individuals in this category who show signs of persistent depression, anxiety and loss of interest in activities. At the next level (Low), we have individuals who have low evaluation of self and a feeling of inferiority and insecurity. At the next level (Moderate) individuals have mixed feelings about their sense of self. While they can appreciate some aspects of their self, they often feel unworthy and influenced by other people's opinions. At the next level (High) individuals have a sense of self-respect and are willing to stand up for their choices and opinions. The have an overall positive attitude towards themselves and are confident about their abilities. At the highest level (Extremely Positive), individuals have a strong sense of confidence and are proud of their skills and whatever they have achieved so far.

Figure 8

Core Strength Construct

Levels	Description	Items
Extremely Positive	Individuals believe that they are good and worthy and that others view them positively. Individuals are more likely to take responsibility for their actions	I am proud of what I have achieved so far I can excel in any difficult task given to me
High	Individuals have a sense of self-respect for themselves even if others might have a different opinion.	I am confident about my abilities I can complete a difficult task most of the times
Moderate	Individuals have mixed feelings. While they can appreciate their own self most of the time, sometimes they feel unworthy. Individuals tend to be influenced by the opinions of other people	I have several good qualities I make a positive attempt to complete a difficult task
Low	Individuals have low overall evaluation of self, persistent feeling of inferiority and a sense of worthlessness, feelings of insecurity and loneliness	I do not have much to be proud of I start a difficult task but give up too quickly
Chaotic	Individuals have persistently depressed mood or loss of interest in activities, causing significant impairment in daily life	I have nothing to be proud of I never take on a difficult task

Building Block Two: Items Design

The second building block, the item design, includes the questions, performances and other stimulators that provides empirical evidence related to the levels of the construct map. There are a variety of items that can be considered. The most common format is the Likert-type item. This agree-disagree approach to measuring attitudes has been around for decades. (Likert, 1932). The reason why Likert scales are so ubiquitous is because of the ease of developing and scoring such items. However, there are many criticisms of this format as well, such as the inclination of people to choose one response side or the other, or the lack of consensus on what a response option would entail (Willits et al., 2016). Another critical issue with Likert items is that

it makes a very strong assumption that each response option is equally spaced (Wilson et al., 2021) which might not be true in a lot of situations. Hence, we choose an alternative to Likert response format – the Guttman format which has more advantages than disadvantages. The idea of a Guttman format is to provide the respondent with a block of response options that progressively become more difficult to agree with. (DeVellis, 2007). Typically, an individual will only endorse the block of statements up to a critical point which would vary by respondent and this then constitutes the variable to be estimated in the scale.

Table 1

Likert and Guttman Example Items from the four sub-dimensions – (a) Stereotype Belief (b) Strength of Fit (c) Domain (d) Core Strength

Sub Dimension	**Likert Items**	**Guttman Items**
Stereotype Belief	Select one answer for each of the statements below – A) My being a transfer student does not influence what people think of me a) Strongly Agree b) Agree c) Neutral d) Disagree e) Strongly Disagree B) Some people judge me based on my transfer student status a) Strongly Agree b) Agree c) Neutral d) Disagree e) Strongly Disagree	Which is the one statement that best describes you – A) My being a transfer student does not influence what people think of me. B) Some people judge me based on my transfer student status. C) People from other groups almost always interpret my behavior based on me being a transfer student
Strength of Fit	Select one answer for each of the statements below – A) I am not a proud being a transfer student a) Strongly Agree b) Agree c) Neutral d) Disagree e) Strongly Disagree B) I feel that everyone should want to be a transfer student a) Strongly Agree b) Agree c) Neutral d) Disagree e) Strongly Disagree	Which is the one statement that best describes you - A) I wish I wasn't a transfer student B) I am not proud about being a transfer student C) I consider myself a transfer student D) I am a proud transfer student E) I feel that everyone should want to be a transfer student
Domain	Select one answer for each of the statements below –	Which statement best describes you –

	A) My skills in data science are poor a) Strongly Agree b) Agree c) Neutral d) Disagree e) Strongly Disagree B) My skills in data science are excellent a) Strongly Agree b) Agree c) Neutral d) Disagree e) Strongly Disagree	A) My skills in data science are poor. B) My skills in data science are average. C) My skills in data science are above average. D) My skills in data science are excellent.
Core Strength	Select one answer for each of the statements below – A) I am proud of what I have achieved so far a) Strongly Agree b) Agree c) Neutral d) Disagree e) Strongly Disagree B) I have several good qualities a) Strongly Agree b) Agree c) Neutral d) Disagree e) Strongly Disagree	Which is the one statement that best describes you - A) I have nothing to be proud of. B) I do not have much to be proud of. C) I have several good qualities. D) I am confident about my abilities. E) I am proud of what I have achieved so far.

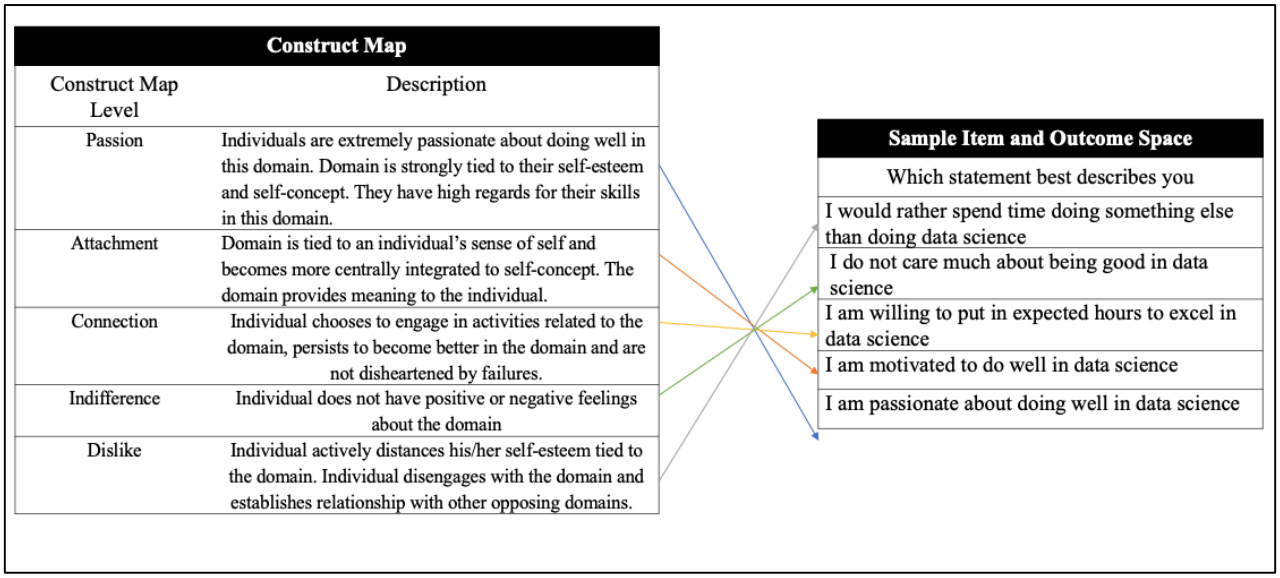
To create the Guttman-style items, we initially developed a collection of Likert-style items that corresponded to different levels of the stereotype threat construct map. After piloting those items with researchers and psychometricians, we chose 25 Likert-style items that seemed to be in line with the construct map. These items encompassed all the four sub-constructs that we define in our construct map. We grouped these items together based on similar content to make several cascading groups or blocks of ordered response options. Finally, we created new response options to fill in gaps so that each block represented an ordered set of statements that were mapped to each level of the construct maps. Through this exercise, we could successfully create 14 Guttman items across the four dimensions. In Table 1, we present example items from each sub-dimension in both Likert format and Guttman format to capture the distinction between the two. As you can see in the Table, Guttman items are a collection of Likert items converted to statement like format and arranged in an order of statements that are most easy to agree with to statements that are the most difficult to agree with. The complete instrument can be found in Appendix.

Building Block Three: The Outcome Space

The third building block or the outcome space for an item refers to a procedure for classifying or categorizing results. It is the process of assigning numerical values to item responses. These values should enable us to link the responses back to the construct map. It is preferred that the outcome space is well-defined with finite and exhaustive categories. In Figure 9, we present a sample item with arrows indicating how the outcome space option maps to the levels of the construct map. Because the instrument is designed in a style with a somewhat Guttman structure, the outcome space is ordered to indicate the cumulative nature of the Guttman items.

Figure 9

Domain Construct with Sample Item and Outcome Space



Since the order of the statements in each Guttman item is designed in a way that it maps to different levels of the construct map, we chose to score the response (selection of a statement) to the item by an individual based on level of the construct map it maps to. For example, in the item depicted in figure 6, if the individual selects "I am motivated to do well in data science", since the item is mapping to level four of the construct map, we would score the response as four. We create these scoring guides for each of our 14 Guttman items.

Building Block Four: Measurement Model

The final building block, the measurement model deals with the process of analyzing the assigned numerical values in a way that it can be related back to the construct map. For decades,

Classical Test Theory (CTT) has been extensively used in the field of psychology. However, the last 50 years have seen a shift away from CTT towards Item Response Theory (IRT) which addresses many of the disadvantages of CTT. IRT represents a family of models called item response models and the measurement theory that they have in common. Unlike the CTT, IRT does not assume that the respondent's true score on the latent ability equals his/her observed score plus error. Instead the model is framed as the probability of a respondent making a certain response, given the underlying position on the construct. IRT has caused major positive changes to psychological test development (Hambleton et al., 1991). The fundamental feature of IRT is that it considers each item individually and so the conclusion of an assessment does not depend on the instrument but on each item within the set of items. Another advantage of IRT is the principle of invariance, i.e. the item parameters do not depend on the respondent's ability and vice versa. The responses given by a group of respondents are used for the estimation of the items and the respondents in that same scale. For an item to be useful, it should be able to differentiate between two persons located in different points along the scale.

The three common types of item response theory models are the one (1PL), two (2PL) and three parameter (3PL) logistic models (Hambleton et al., 1991). The three models differ in terms of the number of parameters estimated. The 1PL model includes only the item difficulty level (which governs the probability that a person will answer the item correctly). The 2PL model, along with the item difficulty, also involves the so-called discrimination (which governs the rate at which the probability of endorsing a correct item changes, given ability levels). The 3PL model, along with the item difficulty and discrimination parameter also involves a pseudo-guessing parameter (which attempts to account for guessing on an item).

The 1PL model is the simplest form of IRT models. Like the other two, it has one parameter that describes the latent trait (ability – θ) of the person responding to the items and another parameter for the item (difficulty). Item difficulty is determined at the point of median probability i.e. the ability at which a respondent would be expected to endorse the correct answer with a probability of 50%.

The following equation represents its mathematical form -

$$P\left(Y_{ij} = 1 \,\middle|\, \theta_j\right) = \frac{e^{(a(\theta_j - \delta_i))}}{1 + e^{(a(\theta_j - \delta_i))}}. \qquad (1)$$

Equation 1 represents the item response function of a 1 PL model, predicting the probability of a correct response given by any respondent 'j' with ability θ_j , on item 'i' with item difficulty δ_i and discrimination parameter 'a', which is constant between items. When we constrain the item discrimination parameter to 1, we get the Rasch model (Rasch, 1960) which is the basis for the family of models that has been used throughout this research. A further elaboration of this model will be carried out in the subsequent chapters.

CONCLUSION

To understand the deleterious effects of stereotype threat, researchers in the past have engaged in the process of operationalizing it, yet, there is a lack of an accepted tool for measuring stereotype threat, possibly because of the complexity of the construct and the lack of a precise definition and clear boundary conditions. We propose a solution to this situation by bringing all the key variables under one umbrella, making efforts to precisely define and lay out their boundary conditions and conceptualize how they interact in a multidimensional space. We develop a generic framework of the stereotype threat instrument that can be adapted to measure stereotype threat under different conditions. We adapt the generic instrument to model how the stereotype experience of a specific stigmatized group, the transfer students, depends on a specific domain, data science. As the next part of the project, we will engage in data collection and test out the psychometric properties of the proposed instrument.

Chapter Three

Validation of the Stereotype Threat Instrument: Measuring Stereotype Threat Experienced by Transfer Students

In this chapter, we test out the psychometric properties of the stereotype threat instrument using a sample of community college transfer students facing negative stereotypes because of their community college status while attending a four-year university. We collect validity and reliability evidence for interpretation and use of scale data. The final calibration sample included 392 transfer students from diverse backgrounds. We used item response theory to analyze the survey data. Overall, we found evidence in support of using the stereotype instrument as a psychometrically valid and reliable metric to measure stereotype threat. Further analysis would be needed to understand how well it functions for other groups, domains and boundary conditions.

.

INTRODUCTION

Validation is the process of collecting validity evidence to evaluate the accuracy, interpretations and proposed uses of tests and instruments based on empirical findings (Cook & Hatala, 2016). Validity can be thought of as a hypothesis and the process of validation as a process of collecting evidence to support or refute this hypothesis. The Standards for Educational and Psychological Testing ("Standards") by American Educational Research Association (AERA), American Psychological Association (APA) and National Council of Measurement in Education (NCME) define validity as a unitary concept in which all accumulated evidence from various sources and processes lead to one single claim and argument about the assessment – is it valid? (Standards, AERA, APA & NCME, 2014). In his paper Kane articulates, "validity is to evaluate the rationale, or argument, for the claims being made, and this in turn requires a clear statement of the proposed interpretations and uses and a critical evaluation of these interpretation and uses" (Kane, 2006).

As a process, validation involves collecting and analyzing data to evaluate the accuracy of an instrument. Therefore, we will fix the design parameters that have been left unknown in the generic stereotype threat instrument to customize it for a specific group and domain. In this project, for the group, we choose the community college transfer students who transferred from a community college to a four-year university. We want to measure the extent to which these students experience stereotype threat when interacting with data science related courses or activities on campus which forms the domain. We will simultaneously examine the psychometric properties of the stereotype threat instrument.

The Sample

For data collection, we followed the protocol laid down by UC Berkeley IRB (study 2021-06-14454). We advertised the stereotype threat instrument in various transfer student groups on social media. These groups comprised of students from four-year universities across the nation. We requested participation from any individual who went through the community college experience and successfully transferred to an undergraduate program in any recognized university. There were 444 students who took the survey across 80 universities nationwide. For our analysis, we retained approximately 88% of our sample (392 students). The remaining 12% of the sample was removed due to missing values. This 12% of the data was chosen to be removed because it had at least 80% of the values missing. The final sample represents a diverse range of students as depicted in Table 2.

Table 2

Demographics of the sample, comparison with the community college student population in United States & California

Demographic Variables	Sample (N=392)	Community College Students in United States (*N= 7.7 million*) **	Community College students in California (*N= 2.1 million*) ***
Gender			
% Female	45.4 %	43.0%	53.6%
% Male	49.2%	57.0%	45.2%
% Other	3.3%		
% Prefer not to answer	1.0%		
% Not Reported	1.0%		1.2%
Race/Ethnicity			
% Asian	11.7%	6%	11.56%
% Black	8.7%	14%	5.9%
% Latino	9.9%	27%	44.4%
% White	47.4%	46%	25.88%
% Others	21.2%	7%	12.26%
% Not Reported	1%		
Age			
18-21. years	48.7%	56%	18 to 24 - 57.7%
22-24 years	30.1%	22 & above- 44%	
25-27 years	10.7%		24 & above – 42.3%
28 and above	8.2%		
Not Reported	2.3%		
First-Generation Student*			
% Yes	45.4%	33%	43%
% No	52.8%	66%	57%
% Not Reported	1.8%		

International Student		
% Yes	38.3%	not available
% No	59.7%	
% Not Reported	2.0%	

Note

***** First-generation student means that your parent(s) did not complete a 4-year college or university degree

** The data for community colleges students nationwide is from National Center for Education Statistics Website, data reported for Year 2019.

*** The student demographic data for community colleges in California is from California Community Colleges Website, data reported for Year 2019.

The first column of Table 2 shows demographic characteristics of the sample. We allowed students to choose whether they wanted to disclose personal information which may have resulted in some missing data. Nearly half of our sample is male (49.2%), 45% of our sample is female. While we have students from all racial groups, nearly half of our sample consists of students who identify as White (47.4%), 11.7% identify as Asian, 8.7% identify as Black, 9.9% identify as Latino and 21.2% identify as belonging to other racial groups or as multi-racial. Community colleges are home to a higher percentage of non-traditional students who are slightly older than the average population. Hence, nearly 8% of our sample consists of students who are 28 and above and 45% of our sample consists of first-generation students which is not surprising given that the percentage of first-generation students in community colleges is very high.

While we had representation of students from universities across the nation, nearly 60% of the students in our sample are from universities in California. Because the students in our sample transferred from community colleges, we wanted to compare their demographics with the overall population of community college students nationwide as well as within the state of California.

In Table 2 column 2, we present the demographics of community college students across the nation. According to the Community College Research Center Report (CCR), over 1/3rd of undergraduate students are made up of community college students (CCR Report, 2019). In 2019, according to National Center for Education Statistic (NCES) website, 43% of the students enrolled in community college undergraduate institutions were males, 57% were females. The ethnicity of community college students can be broken down into 46% White, 27% Hispanic, 14% Black and 7% Asian. In the same year, over 25% of community college students had dependent children and the average age of community college students was 27 years, while the average age of students in full-time undergraduate program was 21.8 years (CCRC, 2018). In Table 2 column 3, we present demographics of community college students in California. We see a 10% increase in female enrollment (53.6%) in California as compared to the nation. Nearly 44.4% of the community college students in California identify as Latino, 25.88% as White, 11.56% as Asians and 5.9% as Blacks.

We compare the demographics of our sample with the population of community college students both nationwide and California. We have nearly equal representation of males and females in our sample unlike what we see at the nation level and the California state level. The racial mix of our sample also shows a deviation. This is not surprising because Latino and Black students have lower transfer rates in comparison to Whites (Crisp & Nunez, 2014). The age distribution of our sample is comparable to the age distribution of community college students nationwide and California. The percentage of first generation learners in our sample differs from the nation, but is representative of the California population.

Overall, while we would not expect our sample to be closely similar to the population of community college students in the country or in California, as it was not recruited as a representative sample, we would prefer it to be approximately representative. Roundly, this is true, except for Latino students. One reason could be that some multi-racial Latino students selected the option "Others" instead of "Latino". In the next iteration, we could capture the racial mix of students who selected the category "Multi-racial", which might help us understand the sample better. Nevertheless, a comparison enables us to understand how demographics plays an influential role in determining which students are successfully able to transfer out of a community college.

METHODS

Analyses

We use Item Response Theory (IRT) methods to analyze the items. Within the family of IRT models, we use the multidimensional random coefficient multinomial logit model (MRCMLM) to calibrate the item parameters and ability estimates. (Adam, Wilson & Wang, 1997). The MRCMLM is a generalized Rasch item response model that uses a scoring function and a design matrix to accommodate the applications of the IRT models used in this study such as the partial credit model and the latent regression model.

First, we want to investigate the dimensionality of the stereotype threat construct. We compare three models to investigate dimensionality: (1) unidimensional model where in a single latent variable (stereotype threat) is assumed to be the underlying cause of all the item responses, (2) a consecutive model (Davey & Hirsch, 1991) where a different latent variable is assumed to be relevant for all the items within each dimension (Stereotype Belief, Strength of Fit, Domain and Core Strength), and (3) between item multidimensional model where a different latent variable is assumed to be relevant for all the items within each dimension but it also allows for correlations among dimensions so that precision of estimates in each dimension can be improved.

To estimate the parameters of the unidimensional model and consecutive model, we use the partial credit item response model (PCM; Wright and Masters, 1982). To estimate the parameters of the multidimensional model we use the multidimensional partial credit model

(MRCML; Adam, Wilson and Wang, 1997) to calibrate the items, estimate student locations and explore the disattenuated correlations between the dimensions. PCM is a unidimensional model used for the analysis of responses recorded in two or more ordered categories. MRCML is the multidimensional version of PCM, where we take the multidimensionality of the construct into consideration.

A logit form of the between-item multidimensional PCM for item *i* with response categories $x=0,\ldots m_i$ can be written as

$$\eta_{pi} = \ln[\frac{P\ (X = x|\theta)}{P\ (X = x - 1\ |\theta)} = \sum_{k=0}^{x} \alpha_d \left(\theta_d - \ \xi_{i(d)k}\right). \quad (2)$$

where by definition

$$P\ (X = 0|\theta_d) = \frac{1}{\sum_{j=0}^{m_i} \exp \sum_{k=0}^{j} (\alpha_d(\theta_d - \ \xi_{i(d)k}))}.$$

$\exp \sum_{k=0}^{0} (\alpha_d\left(\theta_d - \ \xi_{i(d)k}\right)) = 1,\ \theta_d$ is the latent ability on dimension d, $\xi_{i(d)k}$ is the item step parameter for step *k* of item *i* on dimension *d*, and α_d is the steepness ("discrimination") of the response curve for all items on dimension *d*. The α_d parameter is traditionally set equal to 1 for all dimensions and is often omitted from expressions of the PCM. If the model is specified with all $\alpha_d = 1$, the latent variance for each dimension is estimated.

An equivalent expression of the PCM may be written as a function of $\delta_{i(d)}$, the average item step parameter and $\tau_{i(d)k}$, the deviation from the average step parameter for step *k* for dimension *d* such that $\sum_{k=1}^{m} \tau_{i(d)k} = 0$. The two PCM parameterizations are related as follows:

$$\delta_{i(d)} = \frac{1}{m} \sum_{k=1}^{m_1} \xi_{i(d)k.} \quad (3)$$

$$\tau_{i(d)k} = \ \xi_{i(d)k} - \ \delta_{i(d).} \quad (4)$$

and

$$\xi_{i(d)k} = \delta_{i(d)} + \ \tau_{i(d)k.} \quad (5)$$

The $\xi_{i(d)k}$ parameter signifies the θ value at which the probability of responding in category k – 1 equals the probability of responding in category k. The $\delta_{i(d)}$ parameter can be

interpreted either as the average of $\xi_{i(d)} = (\xi_{i(d)1}, \xi_{i(d)2}, \ldots\ldots \xi_{i(d)m})'$ parameters, or as the θ value at which the probability of responding in the lowest category equals the probability of responding in the highest category. Importantly, the PCM does not impose any order restrictions on $\xi_{i(d)}$. In contrast to the formal PCM model parameters, the Thurstone threshold (Masters, 1988), denoted $\lambda_{i(d)k}$ for category k = 1, …m, equals the θ value at which the probability of responding in category k or higher equals 0.5. Thurstone thresholds reflect cumulative response probabilities and are necessarily ordered. For this reason, many researchers often prefer to interpret Thurstone thresholds instead of step parameters. Thurstone thresholds are computable from the formal PCM parameters using Newton-Raphson iteration or other numerical methods. (ACER Conquest, 2021)

Second, we have identified certain predictor variables and additional student characteristics that we want to control for to understand group mean differences on our four dimensions. We use the latent regression model to directly estimate regression coefficients from the item response data. (Adam, Wilson and Wu, 1997). This is advantageous because it avoids problems of misleading differences in means by directly estimating the difference in the achievement of the groups from response data. This model is also known as 'person explanatory' (Wilson & De Boeck, 2004). A latent regression can be performed under the MRCML framework. To expand this model from the one-parameter Rasch model, person ability θ is replaced with a linear regression equation, as seen in Equation 6 (in logit form) :

$$\eta_{pid} = \sum_{j=1}^{J} v_j X_{pj} + \varepsilon_{pd} - \beta_i. \qquad (6)$$

In Equation 6, X_{pj} is the value of person p on characteristic j and v_j is the fixed regression coefficient of person with property j. ε_p represents the remaining person p effect on dimension *d,* left over from the effect of the personal characteristics. $\varepsilon_p \sim N\ (0, \sigma_\varepsilon^2)$ and may be considered as the random effect of X_{p0}, the random intercept.

Finally, we use the Delta Dimensional Alignment technique (Schwartz & Ayers, 2011; Feuerstahler & Wilson, 2019) to compare person abilities across dimensions. For identification purposes, the MRCML model sets each dimension's person ability mean to zero and adjusts the metrics, which consequently makes comparisons across dimensions to be inaccurate without a specific transformation. The DDA technique transforms the initial multidimensional item and step parameter estimates by using the mean and standard deviations of the items under each dimension to be consistent with those for the unidimensional model. Specifically, the means (μ_d) and standard deviations (σ_d) from a unidimensional and multidimensional model are used for the transformation via Equation 7 and Equation 8. δ correspond to item parameters and the τ correspond to step parameters as in Equation 5.

$$\delta_{id(transformed)} = \delta_{id(multi)} \left(\frac{\sigma_{d\,(uni)}}{\sigma_{d(multi)}} \right) + \mu_{d(uni)}. \quad (7)$$

$$\tau_{ikd(transformed)} = \tau_{ikd(multi)} \left(\frac{\sigma_{d\,(uni)}}{\sigma_{d(multi)}} \right). \quad (8)$$

Once the transformed item and step parameters are obtained, a final multidimensional model is run with the new parameters as anchored values. The person estimates obtained from this final multidimensional analysis can be compared across dimensions.

Data Calibration Software and Procedure

For data cleaning, recoding, data manipulation and obtaining the descriptive statistics, we used the software R Studio (R Studio, 2020). We used the package *WrightMap* for Wright maps (Irribarra & Freund, 2014). For item response theory procedures, we used the software Conquest 4.0. (Adam, Wu & Wilson, 2012). The general form of item response model fitted by ConQuest is the multidimensional random coefficient multinomial logit model (MRCML) described by Adams, Wilson and Wang (1997) and it is a generalization of the equation shown above (Equation 2). The model is flexible enough to allow the estimation of different Rasch-type IRT models, including the partial credit model and the latent regression model. The Conquest software can produce marginal maximum likelihood estimation (Bock & Liberman, 1970) with an expectation maximization (EM) algorithm (Bock & Aitkin, 1981) in which the person abilities are assumed to be representative sample from a distribution, usually assumed to be a multivariate normal distribution. However, the EM algorithm is unable to solve the computational problem caused by the exponentially increasing number of quadrature points with a high number of dimensions. The amount of time needed for the estimation increases linearly with the total number of nodes. Hence, we use an adaptive numerical integration method like the Gauss-Hermite quadrature (Volodin & Adams, 1995) for estimation of the item response theory models used in this paper.

RESULTS

Model Fit

We conceptualized the stereotype threat instrument as being comprised of four dimensions. First, we want to empirically test whether stereotype threat functions as a unidimensional concept or a multidimensional concept. To do this, we compared model parameters of the multidimensional PCM with the unidimensional PCM as well as the

consecutive unidimensional approach. In Table 3, we present the comparison of the three models across the following outcome statistics - log likelihood, number of parameters, Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the G-squared statistics. The lower value of AIC, BIC and G^2 indicates the model with better fit. From Table 3, we see that the multidimensional model had the lowest values of AIC, BIC and G^2 parameters. Thus, we conclude that the Multidimensional Model fits better than the Unidimensional Model.

Table 3

Model Fit Statistics for the Three Item Response Models – Unidimensional, Consecutive and Multidimensional

Model	Log Likelihood	# of Parameters	AIC	BIC	G^2	Reliability
Unidimensional	-6681.21	50	13462.41	13492.02	13362.41	0.77
Consecutive – Unidimensional*	-6256.93	53	12619.86	12651.24	12513.86	
Stereotype Belief	-1108.33	8	2232.65	2237.39	2216.65	0.49
Strength of Fit	-1546.91	13	3119.83	3127.53	3093.83	0.67
Domain	-2248.54	20	4537.09	4548.93	4497.09	0.87
Core Strength	-1353.14	12	2730.28	2737.39	2706.28	0.71
Multidimensional	-6129.14	59	12376.28	12411.22	12258.28	0.58, 0.71, 0.89, 0.74

Establishing that multidimensional model fits better than the unidimensional and consecutive unidimensional model is not enough. We need to examine statistical significance of the difference in model fit. Thus, the multidimensional model was compared with the unidimensional model and consecutive unidimensional model using an adjusted likelihood ratio test (Rabe Hesketh & Skrondal, 2012). Results in Table 4 show that the multidimensional model shows significant improved fit over the unidimensional model.

Table 4
Likelihood Ratio Test between Item Response Theory Models

Model Comparison	Chi-Square	Degrees of Freedom	P-Value
Unidimensional v/s Consecutive	848.55	3	p<0.00
Multidimensional v/s Unidimensional	1104.13	9	p<0.00

Thus, in summary, we see that the multidimensional model is preferred. This is what we hypothesized in the previous chapters - this provides empirical evidence in support of our theory. (see later for a detailed discussion of the correlations, which indicate the effect sizes also). In all the subsequent analysis, we will consider only the MRCML model and modifications of it to make claims with regards to our instrument and the sample.

Item Fit

Fit statistics are a summary of the degree to which actual responses to items deviate from their expected values (calculated using the estimated model parameters), summed across the various facets of the data (items, dimensions, etc.). It provides information about how well the data for an individual item or an individual student's performance are represented by the model we have chosen (Wright and Masters, 1982). These statistics may be expressed as t-values, which allow an approximate significance test (misfit is statistically significant if $t > 1.96$ or $t < -1.96$), or they may be expressed as mean-squares, which give a measure of effect size. As in statistics, t-statistics are dependent on sample size, while mean-square statistics are not. Large mean-square statistics (MNSQ) are considered those which are less than 0.75, or greater than 1.33 (based on Adams & Khoo, 1996). Mean-squares less than 0.75, and t-statistics less than –1.96, suggest that there is less variability than expected. T-statistics greater than 1.96, and mean-squares statistics greater than 1.33, suggest that there is more variability than expected. As MRCML is a probabilistic model, some amount of variation is expected. Items that show more local dependency than expected have low fit statistics and are usually of less concern than items that show greater randomness than expected, which may indicate several possible issues with the item. For example, it may be measuring another dimension in addition to the one of interest, or perhaps the wording leads many respondents to misinterpret the item, etc. We use the results from the multidimensional analyses to report fit statistics. There was only one item in the survey that showed misfit and we present that item in Table 5. The mean square fit statistics for this item which belongs to the Stereotype Belief dimension is 1.42 and the t statistics is 2.4.

Table 5

Stereotype Threat Transfer Students Instrument – Misfit Item

Dimension	Item Which is the one statement that best describes you	N	Weighted Fit MNSQ
Stereotype Belief	I almost always feel myself to be a victim of the stereotypes that are associated to the transfer student community	15	**1.42** Expected Range (0.75 – 1.33)
	I often feel that I am a victim of the stereotypes that are associated to the transfer student community	87	
	I do not notice whether people treat me as a victim of the stereotypes that are associated with the transfer student community	189	
	I never feel the stereotypes that are associated to the transfer student community to be also true about myself	95	

Through exit interviews and responses to the open-ended item "*Are you aware of any stereotypes about transfer students? How does it make you feel?"* we tried to investigate why this item showed greater randomness than expected. Some respondents indicated not being aware, not wanting to accept or not using their transfer student identity in their daily interactions with their peers. In the words of a few students -

"I am not aware and I also never mention that I am a transfer student in class or people don't remember"

"Don't want to accept"

Some respondents indicated being aware but not getting affected by the stereotypes. In the words of a student –

"Yes, but I am completely unaffected. It is illogical to discount someone's intelligence or academic capabilities based on being a transfer student or not. For example, personally, I waited a few years before going back to school to figure myself out. I am glad I did because I am

now 100% happy with the career path I chose; I did not rush into it. It was also much more economically beneficial."

Some respondents indicated being acutely aware and affected by the negative stereotypes. In the words of a few students -

"Being a transfer student often makes it obvious to others that I come from a low-income background. I am uncomfortable about people knowing this."

"Transfer students are people not good enough to get in as freshmen," "are more likely to fail classes," "inexperienced," "behind on others." It can be hurtful, but some of it is true, so I feel motivated to tear the stats apart"

Some respondents had mixed feelings -

"I am actually not aware of the stereotypes about transfer students besides the fact that they tend to be older than most students and I think this is sometimes true"

On the other hand, some respondents also thought about positive stereotypes when responding to this item –

"Honestly it sounds like most people think transfers are cooler and have more of a social life because they've had a chance to develop themselves outside of school"

"Transfer students work harder - makes me feel encouraged."

We also looked at the response choices of the sample as indicated in Table 3. Nearly 49% of responses indicated not noticing whether they are victims of the stereotypes associated with the transfer student community, and nearly 25% do not feel the stereotypes associated to transfer student community to be true about themselves.

We notice a difference in how students are responding to the item in the survey versus in the interviews. Overall negative stereotype awareness rate was 40% (N=158) in the survey (captured through response to the open item, *Are you aware of any stereotypes about transfer students? How does it make you feel?"*), and 85% in the exit interviews (*N=33*).

We also looked at how some of the individuals who have been quoted above scored in the four dimensions. In Table 6, we report the weighted likelihood estimates (WLEs) in each dimension for select individuals.

Table 6

Student Responses to Misfit Stereotype Belief Item (WLE Estimates)

	Stereotype Belief WLE estimates (in logits)	Domain WLE estimates (in logits)	Strength of Fit WLE estimates (in logits)	Core Strength WLE estimates (in logits)
"I am not aware and I also never mention that I am a transfer student in class or people don't remember"	-0.36	-1.50	-2.31	-0.71
"Don't want to accept"	-0.35	3.77	4.17	1.63
"Yes, but I am completely unaffected. It is illogical to discount someone's intelligence…"	-1.29	-0.89	-0.70	0.02
"Being a transfer student often makes it obvious to others that I come from a low-income background. I am uncomfortable about people knowing this"	1.38	-5.1	-2.30	0.02
"Transfer students are people not good enough to get in as freshmen," "are more likely to fail classes," "inexperienced," "behind on others."	1.38	-2.60	1.70	0.02
"Honestly it sounds like most people think transfers are cooler and have more of a social life because they've had a chance to develop themselves outside of school"	0.52	-1.82	1.70	0.02
"Transfer students work harder - makes me feel encouraged"	-0.35	-2.18	-0.33	0.79

We can notice some inconsistent patterns, like the student who mentioned "not wanting to accept" scored lower on the Stereotype Belief dimension (-0.35) and the highest in the other three dimensions. It could be that not wanting to accept is analogous to not believing in the stereotype as well. As another example, the student who mentioned a positive stereotype regarding transfer students also scored relatively lower on the Stereotype Belief construct (-0.35). Further analysis reveals that the student who mentioned "not wanting to accept" and both (and the only two among the selected) the students who reported positive stereotypes about transfer students all identified themselves as White. Everyone else in this group belonged to non-Asian minority racial groups. Thus, there could be a race-intersectionality interfering with this item. Historically, in this nation, students have more strongly felt victims of race based stereotypes than any other stereotypes (Smith & Hung, 2012). Thus, we need to explore these racial differences as well. We will do this in the subsequent chapter. However, from the evidences presented above, we can conclude that the negative stereotypes regarding the respondent's social group (transfer students) are so subtle and engrained that it is hard for transfer students to articulate what they feel, believe and go through in the survey. During interviews, students feel more free and open to accepting the negative stereotypes.

We acknowledge that this item and overall the Stereotype Belief construct is hard for students to respond to, hence we see a high degree of randomness. At this stage of our research, we still recommend keeping this item and testing it out with other groups and situations. A careful inquiry during the one on one interviews can throw light on whether this is a pattern across all groups and situations or just particularly for the transfer student's group.

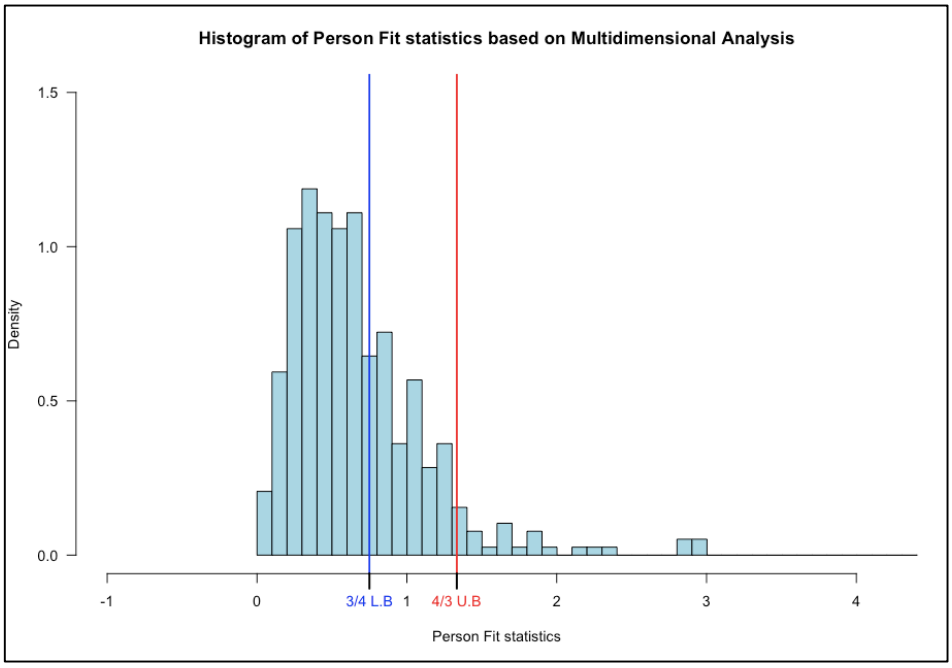
Person Fit Statistics

Just as fit statistics can be calculated for each item, they can also be calculated for each person. The interpretation of such statistics is like those for items; low fit statistics indicate sets of responses that are very regular; high fit statistics indicate more random variation than expected. In general, we are not concerned about student's who respond more regularly than others—instead our focus is on students with higher random variation We present the distribution of person fit statistics in Figure 10.

In the Figure, students with fit statistics below the left blue line (the lower bound acceptable range) show very regular responses. It can be seen that a large proportion of students 66.67% fall in this range. The items were designed to perform this way and this is not a cause of concern. 27.13% of the students fall between the left blue line and the right red line (the upper bound acceptable range). Again, these students are also within the permissible range, so not a cause of concern. Only 6.2% (*N=24*) of the students fall above the right red line.

Figure 10

Person Fit Statistics



One would expect certain percentage of randomness in a probabilistic model like MRCML. Nevertheless, to describe the results comprehensively, we show some examples of students in this upper tail, using kidmaps. Kidmaps (shown in Figure 11 and Figure 12) are modified versions of Wright Map (Adams & Khoo, 1996). Each map represents only one respondent and depicts the response pattern for that respondent. The left-hand side of the map shows the item responses that were not achieved by the respondent and the right-hand side shows items that were achieved. The symbol "XXX" denotes the location of the respondent. The "surprise lines are indicated as two sets of dots: ".....". Responses outside the surprise lines are considered "surprises," and the further they are away from the surprise lines, the more surprising they are.

Figure 11
Kidmap for Case No 25 on Strength of Fit Dimension (MNSQ = 2.96)

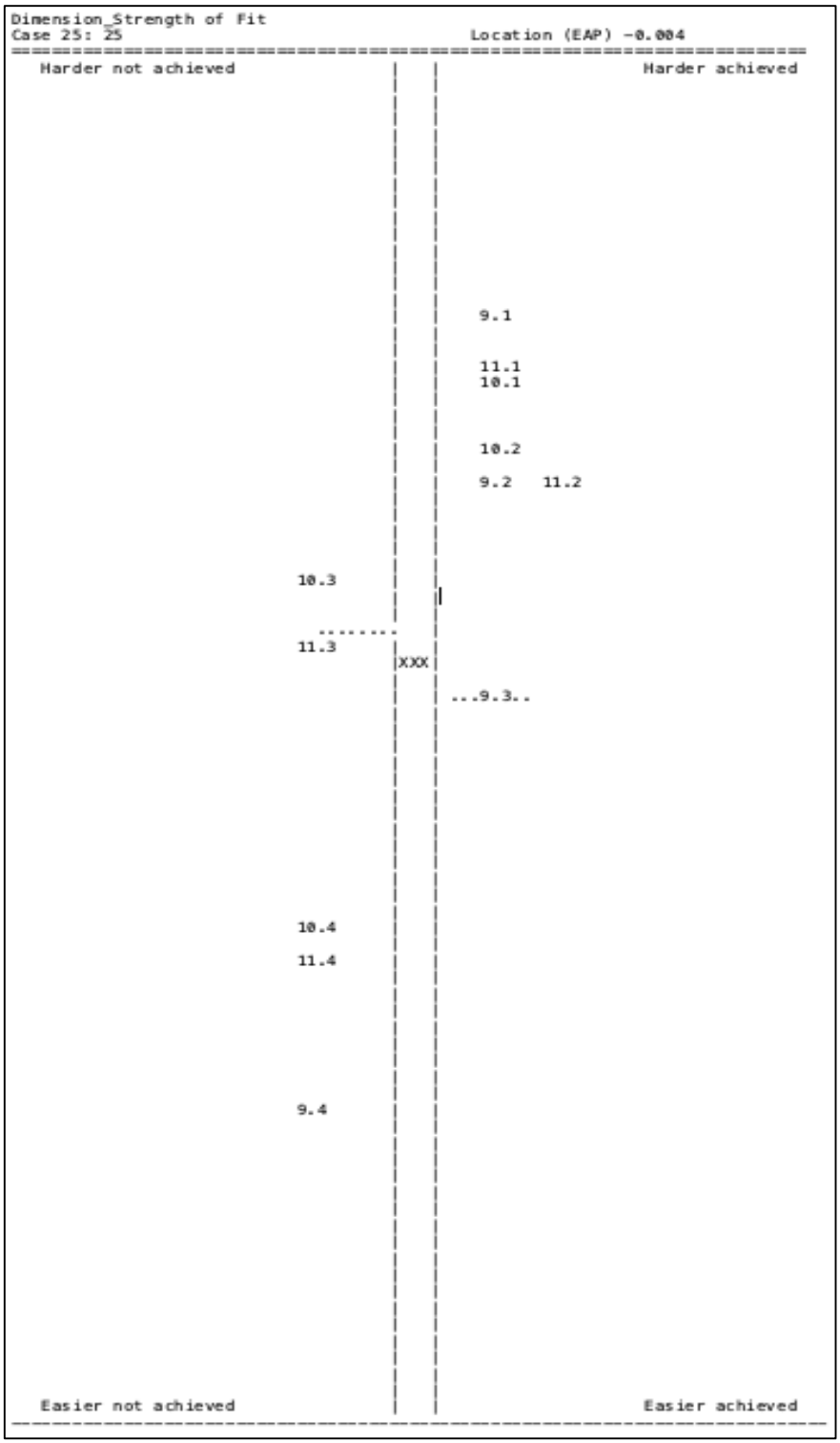
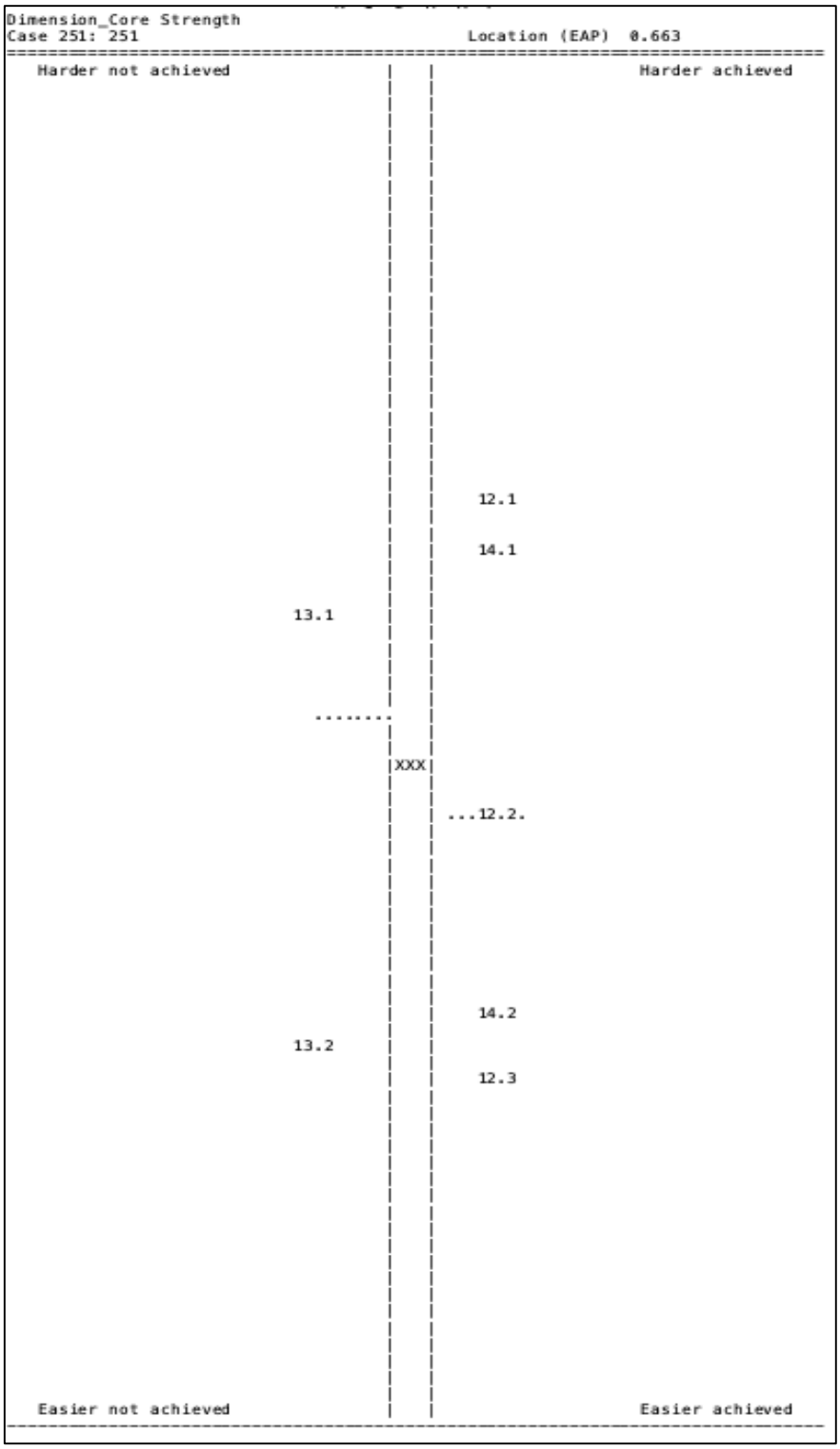


Figure 12
Kidmap for Case No 251 on Core Strength Dimension (MNSQ = 5.06)



In Figure 11, we present the kidmap for Person 25 on the Strength of Fit dimension – the MNSQ fit statistic for this student is 2.96, considerably above the usual limit of 1.33. We notice that Person 25 has one item options that are easier but not achieved (9.3=item 9, response 3) and one item response options that was harder and achieved (10.3 = item 10, response 3). This pattern of being able to agree with a few more difficult items and not with a few easier items, indicates a student who is responding fairly consistently with the overall tendency of most students, but that has some important differences for a few items. In a clinical situation, one would want to investigate why the student differed from typical for these few items.

Similarly, in Figure 12, we present the kidmap for student 251 on the Core Strength dimension—the MNSQ fit statistic for this student is 5.06, a great deal above the usual limit of 1.33. We see that the responses 12.3 and 14.2 are much more surprising than any responses by student 25. This student has more "surprising" results than expected ones. These specific surprising responses need to be considered for this specific student, but the pattern is so strong, and relatively uniform, that one might wonder whether this student experiences Core Strength in a different way than other students.

While we present only two example kidmaps, we investigated kidmaps for all the 24 students who were above the acceptable range, and found some inconsistent response patterns across dimensions. Having a small percentage of students whose responses are more random than expected is not a great concern—this will still occur even in data simulated to fit the psychometric models exactly. Thus, we can conclude that overall the student responses are consistent to the items, although, in a clinical situation, one would want to look again at the students with poor fit to make sure whether reporting overall outcomes such as locations on the flagged dimensions, was proper.

In Table 7, we compare the WLE estimates of the students showing high person misfit with those of the entire sample. We notice that these 24 students that are showing high person misfit have higher Stereotype Belief, low Domain Identification, low Strength of Fit and low Core Strength in comparison to the average respondent. The fact that these students come with a relatively strong belief in the negative stereotype and relatively lower identification with the domain and transfer student community in comparison to the sample, seems like they are probably disengaged with the survey also. Nevertheless, since the number is very low, we can go ahead with the rest of the analysis.

Table 7

WLE Estimates of selected students showing high person misfit (N=24)

	Stereotype Belief WLE estimates (in logits)	Domain WLE estimates (in logits)	Strength of Fit WLE estimates (in logits)	Core Strength WLE estimates (in logits
Respondents showing high person misfit (*N=24)*	0.55	-0.44	-0.10	-0.46
Average scores of all respondents (*N=392)*	-0.00	0.12	0.05	0.04

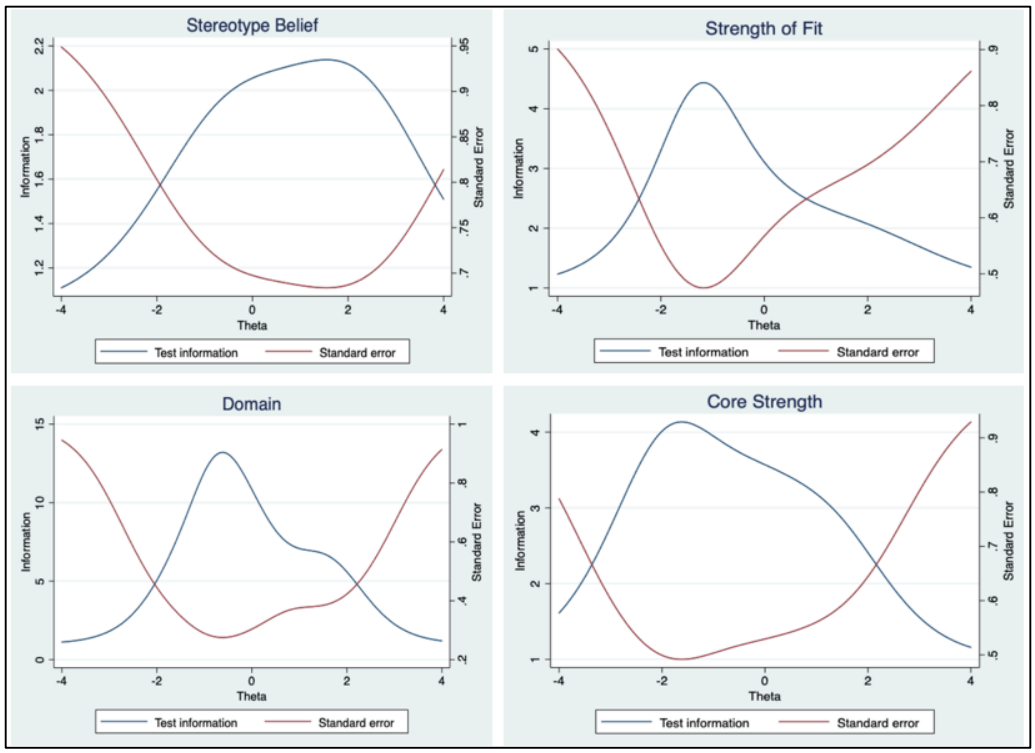
Standard Error Measurement

We also analyzed the standard error of measurement and test information curves for all the four dimensions (see Figure 13). The test information function is shown in blue and it values may be read off the left-hand axis. The SEM is shown in red and its values may be read off the right-hand axis. The test information function tells us how well is the instrument doing in estimating ability over the whole range of ability scores. The SEM function tells us the precision of our instrument. Higher SEM leads to less precision.

As shown in Figure 13, for the Stereotype Belief dimension, we notice that the SEM is higher for students who scored lower in this dimension. This means there is less information about students with lower Stereotype Belief. SEM usually increases for outliers and this is expected. For the domain dimension, we notice that the SEM is higher for students both at the lower and upper range of estimates. For the strength of fit dimension, we observe a similar pattern but also notice that the SEM curve is steeper than the other dimensions. This means that maximum information is available for students within the ability range of -2.5 to 0.75 logits. For the items in the core strength dimension, the SEM is higher for students who scored higher, which means the information is less precise for students with higher estimates in this dimension.

Figure 13

Standard Error of Measurement (SEM) and Test Information Function for the Four Dimensions – Stereotype Belief, Domain, Strength of Fit and Core Strength



Reliability

In addition to fit statistics, we also look at the EAP/PV reliability which gives the precision of person ability estimates. It is measured as the model explained variance divided by the total person variance. Table 6 lists down the reliability of unidimensional and multidimensional analysis. Overall reliability of the instrument when treated as a unidimensional construct is 0.763. In the multidimensional format of the instrument, we could achieve high levels of reliability for all but one dimension – the Stereotype Belief dimension with a reliability of 0.587. The reason we see a weak reliability for the Stereotype Belief dimension is because probably the instrument is not sensitive enough to differentiate between individuals with high Stereotype Belief and low Stereotype Belief. Thus, we might want to include more statements in the Guttman block of items. However, this was something we also found when we were interviewing the transfer students – they were confused with regards to their overall beliefs with regards to the stereotypes about their group.

Table 8
Reliability of Stereotype Threat Transfer Students Instrument

Dimension	EAP/PV Reliability (Unidimensional)	EAP/PV Reliability (Consecutive)	EAP/PV Reliability (Multidimensional)
Stereotype Belief	0.77	0.492	0.581
Strength of Fit		0.667	0.708
Domain Identification		0.874	0.887
Core Strength		0.709	0.740

Validity Evidence

There is a scientific as well as a social need to validate an assessment (Kane, 2006) and we use both these lenses as we adapt the six strands of validity proposed in the Standards of Educational and Psychological Testing report, evidence based on (a) instrument content, (b) internal structure, (c) response processes, (d) relation to other variables, (e) consequences of use and (f) fairness. (Standards, AERA, APA & NCME, 2014). We will present the argument of validity for each dimension separately.

Stereotype Belief

Evidence based on Instrument Content: To compile evidence based on instrument content, we must analyze the relationship between the instrument content and the construct it is intended to measure (AERA/APA/NCME, 2014, p. 11).

The Stereotype Belief construct was created not only based on social psychological theory but also based on many conversations and discussions we had with the transfer students during the initial stages of the project. We conducted 20 one-on-one interviews with the transfer students to capture their views and experiences. We asked students about their awareness regarding the stereotypes associated with transfer students. Most of the students, especially those enrolled in STEM courses, indicated feeling as “underdogs” and going through “imposter syndrome” especially in their first semester of transfer. In the words of a student –

"Transfer students are people not good enough to get in as freshmen, are more likely to fail classes, inexperienced, behind others. It can be hurtful, but some of it is true, so I feel motivated to tear the stats apart. Mixed feelings I guess"

The challenges of transferring from community college were not only academic in nature but also cultural, as indicated by a few respondents–

"It is hard to make friends because groups have already been formed"

"*It is hard to get into clubs because they pool you with freshman."*

While it is more severe during the first year of transfer, there were respondents who indicated that the effects are long lasting –

"I fear discrimination by companies, hence I do not like to put my transfer student status on resumes when applying for jobs."

We didn't associate the term "positive" or "negative" when we asked them regarding their familiarity with stereotypes, yet, most of the respondents spoke to us only about the negative stereotypes. Only two respondents mentioned positive stereotypes, such as –

"transfer students are more hardworking than the rest"

"transfer students have more life experience than others"

We realized that these negative stereotypes are not a manifestation of the external community (such as professors, other non-transfer students, etc.), rather they are a manifestation of their own beliefs. This clearly came out during our conversations as indicated by a respondent-

"In general professors are nice and the education faculty cares."

Hence, the source of threat is clearly embedded within the individual. This threat stems from the fact that going to community college was not the first choice for most of the respondents. As indicated by a student -

"In community college, there is no social aspect, people are there because they got rejected so their goal is to get out of there"

Hence this feeling of "others got it right in the first shot" adds an additional layer of negative beliefs to an otherwise already stressful process of transferring and getting adjusted to a new environment. We also realized that these respondents were at varying levels of awareness and belief with regards to the stereotypes. There were a few who also indicated being aware but not being affected and we wanted to capture that aspect in our construct as well. In the words of one student –

"I never feel the stereotypes that are associated to the transfer student community to be also true about myself"

Many of these aspects that we covered through interviews are in sync with what we found in the literature as well. In summary, students have varied opinions about the stereotypes associated to transfer students and thus, also vary in the extent to which it affects them.

We use this evidence to develop our construct map. We attempt to categorize students based on the extent to which they believe in the negative stereotype – *disbelief* category comprises of students who do not believe in the stereotype, *partial belief* category comprises of students who believe in the stereotype to some extent and *belief* category comprises of students who completely believe the stereotypes regarding the transfer student. We capture two aspects of their belief/awareness, (1) the extent to which students feel the negative stereotype to be true about themselves and (2) the extent to which students feel the negative stereotype to be true about their group. The items are developed in correspondence to the construct map, so that they can successfully capture the differences in belief among respondents. Each statement of the Guttman item is scored based on the level of the construct map it is aligned to. We carefully align the construct map, items and scoring guide in a meaningful way, thus establishing content validity evidence of the instrument.

***Evidence based on Response Processes*:** To compile evidence based on response processes, we engaged in analyzing student's thought processes while taking the instrument using the think-aloud procedure. We captured feedback from 13 participants.

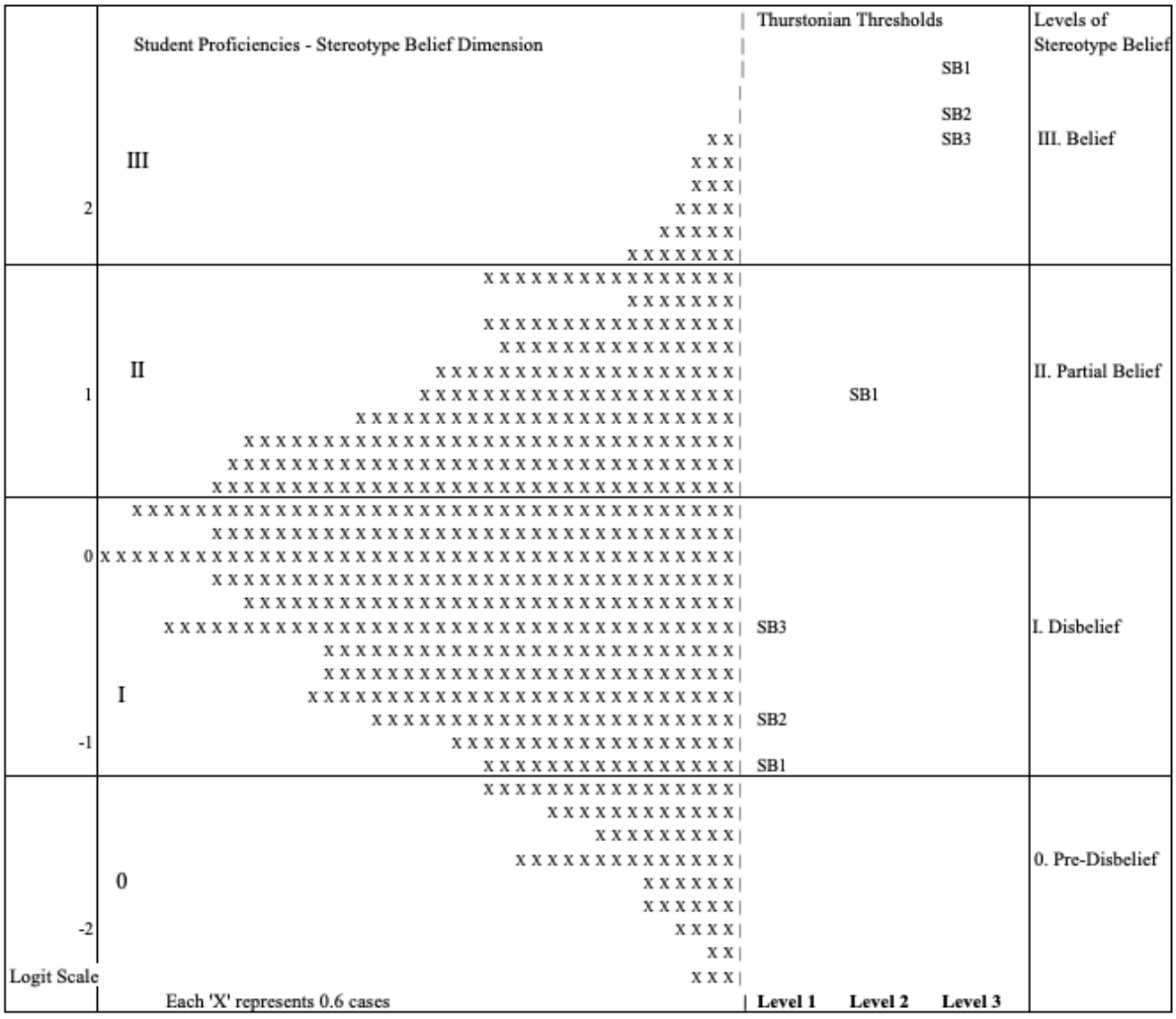
Overall, students found the survey items clear and easy to understand. While filling out the demographic section, some students felt like the negative stereotypes have a more severe impact among transfer students who are females, first generation learners, undocumented students and other minority groups. We will empirically test some of these aspects in the subsequent chapter. We used their feedback to make changes to the instrument, especially with regards to the ordering of the items. For example, we reordered the Stereotype Belief items to appear last in the survey so that responding to these does not influence their responses to the questions in the other dimensions.

Evidence based on Internal Structure: *Wright Map.* The Wright Map provides insight into how empirical evidence corresponds to theory. In the Rasch framework, the probability of endorsing the most valued statement is modeled as a function of the respondent's attitude and item difficulty (Wright and Masters, 1982). A Wright map depicts the visual representation of this relationship. We present the Wright map of the Stereotype Belief construct based of the consecutive unidimensional analysis in Figure 14. The left side of the map shows the distribution of the respondents based on their estimates on the Stereotype Belief construct. In accordance with the model assumptions, the respondents are distributed normally between a range of around

-2 logits to +2 logits, we observe most of our respondents lying between -1 and +1 logits. This means that most students are tending to pick answer options at the middle of the scale. Hence, we can conclude that the range of items are appropriate for this group of respondents. In other words, the items are not too easy or not too difficult to agree with.

Figure 14

Wright Map of Stereotype Belief Dimension from Consecutive Unidimensional Analysis



The right side of the map is ordered so that each column represents a different level of the Stereotype Belief construct. This representation helps to see whether items are behaving as expected with respect to the hypothesized construct map levels. The right side also displays thresholds for each item step that have been separated into columns, which correspond with the

levels of the construct map. Since each column represents a threshold that corresponds to a given level of the construct map, if the construct map levels have a reasonably constant meaning across items, the thresholds in each column should be in a similar level of difficulty, and the difficulties should increase as the levels of the construct map increase. We see that all items in the Stereotype Belief construct follow this pattern.

Further as we notice in the Figure, two of the items have only two thresholds since they had only three response choices. Given that there are three levels of the construct map that they can map onto, we had a choice of placing the two thresholds onto any two levels of the construct map. We chose to place them in Level 1 (Disbelief) and Level 3 (Belief) respectively. This is because of two reasons – (a) theoretically, both the items were designed in a way that they would successfully differentiate students who believed the negative stereotypes from those students who did not, and this is most strongly formed at the bottommost and topmost levels of our construct map, and (b) empirically, it was observed that the two thresholds were closer to the median item thresholds at level 1 and level 3.

We also notice that the item thresholds representing the different levels of the construct occupy different "bands" of the scale. We arrive at this banding structure by drawing a line between the sets of thresholds for each level. We notice that these bands are distinct and ordered consistently with the hypothesis in the construct map (Figure 5). These bandings provide internal structure evidence by demonstrating how consistently the items engage the levels.

In addition, by following the lines across from the item side to the student side, we can use the bands to associate the students with different levels of the stereotype belief dimension. In fact, we can now report the numbers (and percentages) of students in each level in our sample as shown in Table 12.

Evidence based on Relation to Other Variables*:* Along with the 14 Guttman items, we requested additional information from the respondents which we would use as external variables. We list the external variables in Table 9.

Table 9

Response Variables

Response Variable	Category of Responses	Label
What type of influence does being a transfer student have in your life? (influence)	1. Positive Influence (Reference Category) 2. Negative Influence	positive_influence
How many data science related courses/activities are you currently involved in? (num_courses)	1. 0-5 courses/activities 2. 6 – 10 courses/activities (Reference Category)	num_courses
Some people say that a lot can be overcome with the right mindset and will power. Do you think so? (mindset)	1. Yes (Reference Category) 2. No or Neutral	mindset

As part of this study, we used a Likert-type item (as shown in Table 9), "What type of influence does being a transfer student have in your life?" to ascertain of how students feel about their transfer student status. The item responses were judged into three responses categories (positive, negative and neutral) and we scored the judged categories into two final categories– (a) Positive influence (scored as 1), (b) Neutral or negative influence (scored as 0). We also asked students to indicate, as a numeric input, the number of data science related courses/activities they were currently involved in to understand the extent to which they are engaging with the data science domain. The responses to this item ranged from 0 to 10. We categorize these responses into two groups – (a) 0 to 5 courses/activities (scored as 0), (b) 6- 10 courses/activities (scored as 1). Finally, as an open response item, we asked the students whether they feel a positive mindset can help overcome a lot of the challenges in life. All students who indicated that a positive mindset can help overcome challenges were scored into category "Yes" (scored as 1). Students who completely or partially negated the belief that a positive mindset can help overcome challenges were scored into category "No" (scored as 0).

We use a multidimensional latent regression model to analyze the effect of these predictor variables on the overall estimates in each dimension. We present the regression results in Table 10.

Table 10
Latent Regression Coefficients

Regression Variable	Stereotype Belief	Strength of Fit	Domain	Core Strength
positive_influence	-0.674 (0.097) ***	1.054 (0.124) ***	0.328 (0.230)	0.592 (0.148) **
num_courses	0.426 (0.097) ***	0.241 (0.123)	1.137 (0.229) ***	-0.154 (0.148)
mindset	-0.501 (0.139) ***	0.278 (0.176)	1.152 (0.329) ***	0.847 (0.212) ***
constant	-0.122 (0.130)	-0.513 (0.165)	-1.713 (0.308)	-0.638 (0.198)

**** p<0.01, ** p<0.05, * p<0.1*

As we notice in Table 10, the overall Stereotype Belief of students who indicated their transfer student status as having a positive influence in their lives is 0.674 logits lower than students who indicated otherwise. This difference is statistically significant with p value less than 0.01. Similarly, we notice that students who took a higher number of data science courses/activities scored significantly lower on Stereotype Belief than those who were involved in fewer courses/activities. Finally, we notice that students who indicated a positive mindset as the key driving factor to overcome obstacles scored significantly lower on Stereotype Belief than those who indicated otherwise. These differences are consistent with what we were hoping to expect (see Figure 1, Chapter 2), thus lending further support in favor of our instrument.

Evidence based on Consequences of Use: Regardless of the evidence gathered in the above categories, if an instrument is found to have negative consequences then careful consideration is required when determining its intended use (Wilson, 2005). The main purpose of this instrument is to make positive efforts to measure, understand and mitigate the harmful effects of stereotype threat and that would be the best use of it. Given the newness of the instrument, there are no actual consequences of use yet. However, asking students questions regarding their stereotype belief might, in a way, reinforce the stereotype in their minds again. This is something we should be careful about. We attempted to solve this issue by using the term "stereotype" instead of "negative stereotype" in our survey, thus allowing respondents the liberty to think about both negative as well as positive stereotypes.

Fairness: The Standards define fairness using two key concepts, the first focuses on fair and equitable treatment of all assessment takers (Standards, AERA, APA & NCME, 2014). Overall, by using a common platform like social media, we ensured that the survey was made accessible to all individuals from diverse subgroups (as those defined by race, gender, etc). The interviewer took special considerations to ensure that the interview process was unbiased and all

sensitive information was handled with care and anonymity. On the Stereotype Belief construct, the focus was solely on the stereotypes of the transfer students. We refrained from asking items which focused on understanding these beliefs for a sub-group and thus, ensured a universal design of the instrument. The second concept of fairness emphasizes issues of fairness in measurement quality. (Standards, AERA, APA & NCME, 2014). An important aspect of this concept includes testing for any measurement bias within items that might be favoring a sub-group. We will cover that aspect in the subsequent chapter.

Strength of Fit

Evidence based on Instrument Content: During the interviews, we asked students regarding the extent to which they feel associated to the transfer student community. Some respondents indicated a positive association –

"it brought me to my group of friends",

There were some students who felt they are being looked down. As a result, they felt uncomfortable disclosing their transfer student status. In the words of a respondent –

"Although I am proud of being a transfer I still get weird looks from people when I tell them I am a transfer."

Some students claimed that they receive differential treatment because of their transfer student status. In the previous section, we indicated a respondent not feeling comfortable with putting his transfer student status in his resume because of fear of discrimination. All these instances indicate students' varying levels of associations with the transfer community. In the words of a few respondents –

"People seem to treat me differently at the university when they realize I am a transfer student."

"My transfer friends are a lot more mature and focused in our studies. We all came here for a reason, and we have wealth of stories about our zigzaggy paths. Some fun ones, some painful ones. At the same time, many others, especially clubs, exclude us, and many of us have had very limited opportunities before coming here. A lot of first-years judge us negatively as having taken the easy path."

We also encountered students who found it hard to articulate and understand their identification with the transfer students group -

"It's hard to summarize the transfer experience, especially for older re-entry students. There's just a lot to it"

Thus, we realize that students varied greatly with regards to their level of comfort in disclosing and acknowledging their transfer student identity. While some claimed to be proud group members, there were some who were not so comfortable with their transfer student identity and there were some who were embarrassed to acknowledge their association. We try to capture these differences in sense of belongingness with the stereotyped group (transfer students in this case) through the Strength of Fit construct map. We used five levels to categorize students on strength of fit - *negative fit, low fit, moderate fit, complete fit, over fit* (see "instrument development" section for detailed definitions) The hypothesis is that higher Strength of Fit would lead to higher levels of stereotype threat experience. The items were developed such that each statement in the Guttman block targeted one of the levels mentioned above. The scoring of the Guttman items was based on the level of the construct map the selected statement (of the block) was aligned to. Thus, by aligning the construct map, items and scoring guide through this process, we could establish content validity evidence of the construct.

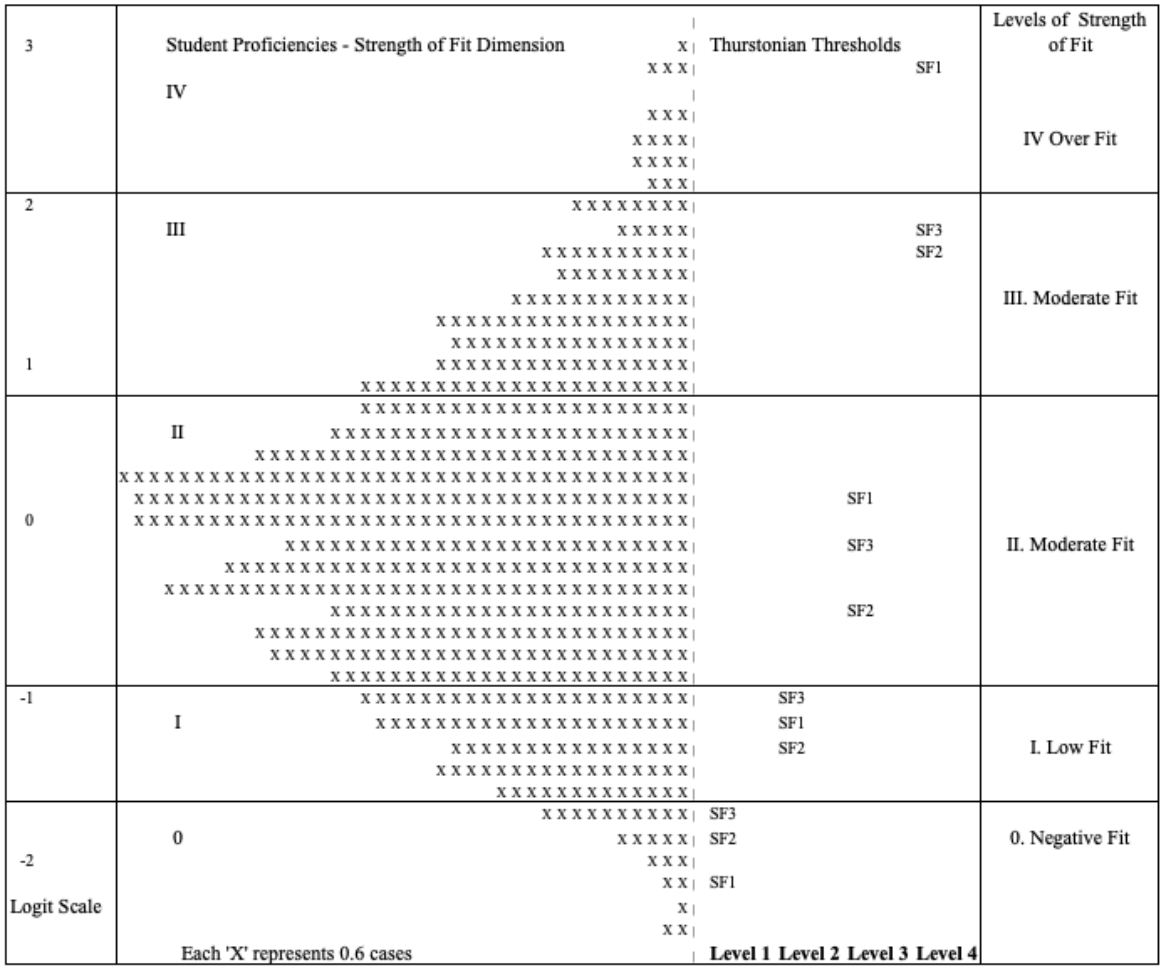
***Evidence based on Response Processes*:** Overall, students in think-aloud interviews indicated that they were comfortable answering the items in the Strength of Fit construct. There were few students who mentioned that they haven't met anyone who was obsessed with their transfer student status which is currently the highest level of our Strength of Fit construct map (Over Fit). We believe, while this may be true for the transfer students community, we need to test it out with some other group (age, race, ethnicity, etc.) before making further inferences. Students also indicated that their intersecting identities like race and gender have a strong impact on their identity as a transfer student as well. We will test this out in the subsequent chapter.

Evidence based on Internal Structure: We present the Wright map of the Strength of Fit construct based of the consecutive unidimensional analysis in Figure 15. On the left side of the map, the respondents have a roughly normal distribution between around -2 logits to +3 logits. We observe that most of the respondents lying between -1.5 and +1.5 logits. Hence, we can conclude that the range of items is appropriate for this group of respondents.

We notice that the bands are distinct and ordered except for the topmost threshold of item SF1 (refer to Appendix for complete item). During think-aloud interviews, students mentioned that the response option that maps to Level 4 of item SF1 (I feel everyone should want to be a transfer student) was one of the most difficult statements to agree with throughout the instrument. Hence, we see that evidence in the Wright Map as well. We extend the lines from the item side to associate students with different levels of the strength of fit dimension. We report the numbers (and percentages) of students in each level in our sample in Table 12.

Figure 15

Wright Map of Strength of Fit Dimension from Consecutive Unidimensional Analysis



Evidence based on Relation to Other Variables. From Table 10, we notice that, on average, the Strength of Fit estimates of students who indicated their transfer student status as having a positive influence in their lives is 1.054 logits higher than students who indicated otherwise. This difference is statistically significant with p value less than 0.01. Similarly, on average, students who took a higher number of data science courses/activities also scored higher on Strength of Fit (0.278 logits), but this difference was not statistically significant. Finally, we notice that on average, students who indicated a positive mindset as the key driving factor to overcome obstacles scored 0.278 logits higher on Strength of Fit than those who indicated otherwise. But this difference was not significant.

Overall, the differences observed among various subgroups are consistent with theory where we found that students who view their group positively would also be more strongly attached to their group. (see Figure 1, Chapter 1). We did not observe any significant association of the outcome variable (*positive_influence*) with the number of data science courses taken and

believing in a positive mindset. This is because viewing the transfer student's group positively does not necessarily lead to higher domain identification or higher core strength.

Evidence based on Consequences of Use: As mentioned previously, the instrument is still very new and we would like to test it on a larger group before any consequence of use can be determined. There are no negative consequences specific to the Strength of Fit dimension. One thing we should be careful about is that the items have been designed to capture situation specific stereotype threat. Group identification is a dynamic construct and it may change over time which would also lead to changes in stereotype threat. This is especially true in our situation, since the students on average spend up to 2-3 years as transfers. Thus, we will need to recalibrate the individuals' stereotype threat experience again should the situation/conditions change.

Fairness: We attempt to ensure that the items of the Strength of Fit construct are focused on the transfer student community overall and that there is no influence or bias towards a sub-group. By allowing the students to fill the survey at a location and time of their preference, we ensure they are in a comfortable space and do not feel influenced in anyway. In the pilot stage, we tested out the instrument and no concerns were reported. We allow students to skip/miss a question if they do not feel like answering it. Overall, we attempt to treat each respondent and their responses in an equitable way, thus addressing any concerns with fairness.

Domain

Evidence based on Instrument Content: When we started our study, we did not have a preference of domain. Through conversations we realized that data science as a field attracts students from various disciplines and hence we decided to test data science as our domain first. Overall students felt that the coursework changes drastically when moving from a community college to a four-year university. In the words of a respondent –

"Community college courses are slower paced and student find them easier to complete."

We realize this shift in pace strongly influences their interest in data science also. Respondents indicated that despite being involved in data science courses in a four-year university, they were not exposed to data science courses in the community college. Thus, one would expect transfer students to be lower in domain identification because of not getting enough exposure to data science courses at the community college. But, some respondents indicated actively participating in data science related clubs and activities in the absence of formal courses.

We use this evidence to develop the Domain construct map. We created the construct map levels to capture the varying levels of interest in the data science Domain, from *dislike* to *passion,* the five construct levels were designed with careful consideration and discussion with experts and students themselves. We use the construct map levels to guide the development of the items. While developing the items, we decided to incorporate both participation in formal data science courses as well as involvement in data science related activities as indicators of domain interest. All the items solely focus on individuals' self-perception and not their actual abilities. The scoring of the items was aligned to the construct map levels. By careful articulation of what domain identification means, we could establish a strong alignment of the construct map, the items and the scoring guide and thus establish content validity evidence of the domain construct.

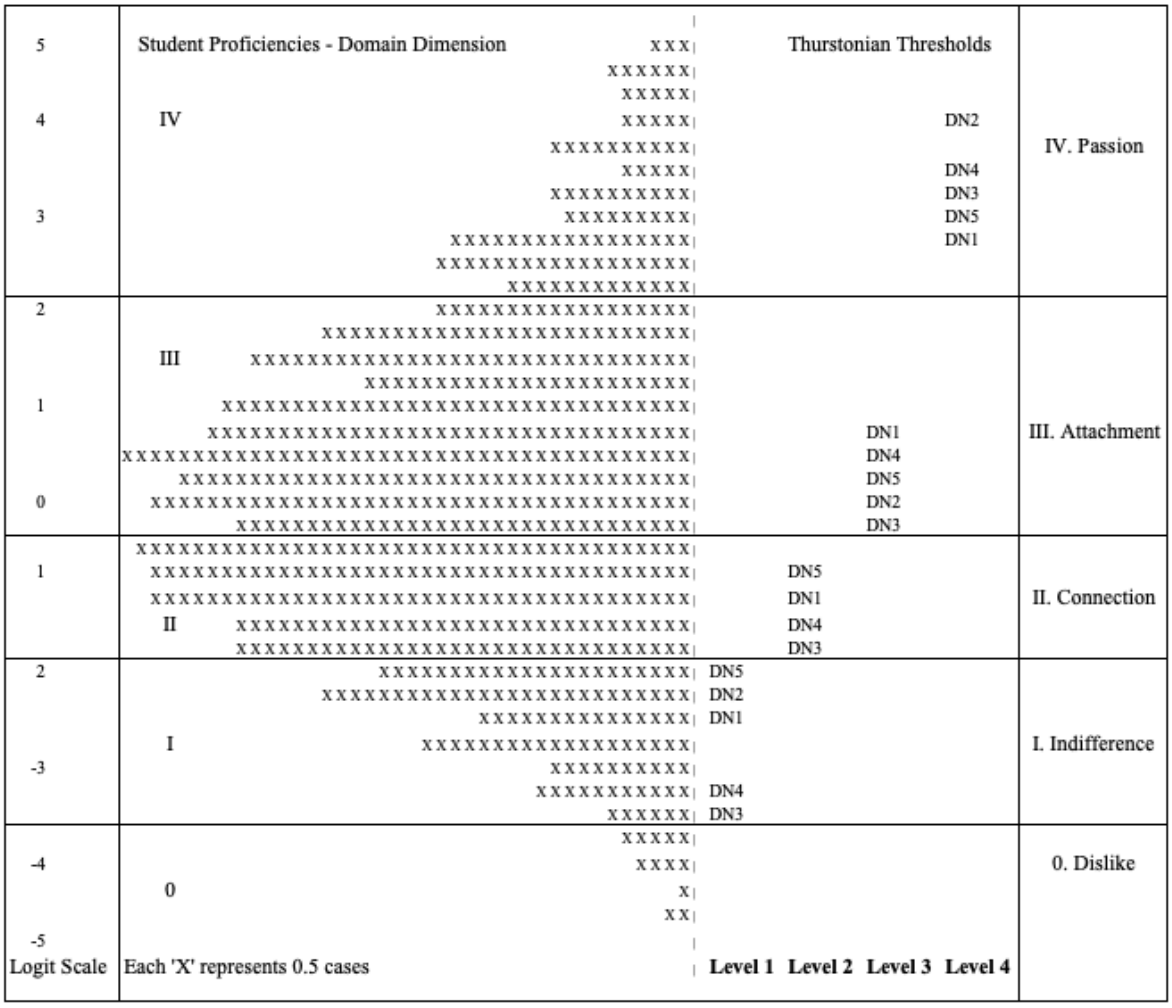
***Evidence based on Response Processes*:** Overall, students were comfortable with the items in the Domain construct. Students mentioned that further clarification was needed with regards to what constitutes a data science course – whether computer science, econometrics, etc. fall in the data science bracket. We clarify this in the survey clearly. We also asked students to indicate their interest in activities (such as clubs, quizzes, online courses, etc.) related to data science to determine their overall interest in the domain. Since our choice of domain was academic in nature, we were planning to include overall GPA scores of students in our study. But students in our think aloud interviews mentioned that grade point average or GPA scores are often misreported and should not be a part of the study. GPA scores also greatly vary depending on the course, department and university. Hence, we decided not to include them in the study.

Evidence based on Internal Structure: *Wright Map.* We present the Wright map of the Domain Construct based of the consecutive unidimensional analysis in Figure 16. The respondents show a roughly normal distribution between around -5 logits to +5 logits. We observe most of our respondents lying between -3 and +2 logits. Hence, we can conclude that the range of items are appropriate for this group of respondents.

We see that the thresholds in each column are arranged around similar difficulty level. We notice that the banding structure is distinct and ordered, consistent with the hypothesis of the construct map (Figure 4). This provides internal structure evidence. We report the frequency distribution of students in each level in our sample in Table 12.

Figure 16

Wright Map of Domain Dimension from Consecutive Unidimensional Analysis



Evidence based on Relation to Other Variables. The overall domain identification of students who indicated a positive mindset as the key driving factor to overcome obstacles was 1.152 logits higher than students who indicated otherwise (Table 10). This difference is statistically significant with p value less than 0.01. Similarly, on average, students who took higher number of data science courses/activities, also scored significantly better in this dimension. This provides a strong evidence in support of our instrument. Indeed, more data science courses a student takes, more will be his/her identification with data science.

Evidence based on Consequences of Use. This instrument is solely for measuring stereotype threat among transfer students and we recommend usage only for that purposes. Since the items are based on individuals' self-perception which does not capture data science ability, we should be careful with regards to the conclusions that we can draw about the individual.

Thus, domain identification items should not be used to make decisions about an individual's overall ability in data science.

Fairness. Domain is the only academic focused dimension in our instrument. We need to be careful so that students from different academic backgrounds can approach the items without any discomfort or apprehensions. We avoid asking any questions with regards to their data science ability (such as scores in a recent data science course). We only capture their self-perception. With that we can achieve universality of our construct. Students from different academic backgrounds (economics, math, geography, political science, etc.) can approach the items with equal comfort. Some of the students indicated lack of opportunity to take data science courses despite having interest in pursuing a data science course or degree. Keeping this in mind, in the items of the Domain construct, we do not differentiate between students who were involved in data science activities and students who were enrolled in formal data science courses, thus ensuring equality and standardization of our instrument.

Core Strength

Evidence based on Instrument Content: Almost all respondents indicated that the right mindset plays an important role in overcoming stereotypes associated to transfer students. In the words of a respondent –

"I think it is harder to make friends but there are so many resources and people to help!"

When we asked them to name one key driving factor that helps them navigate through the challenging environment created by the negative stereotypes, a lot of students mentioned "resilience" as that key factor, some mentioned "self-belief" as well. Some students also felt that that the transfer student experience makes them more hardworking, passionate and helps them develop a stronger mindset. In the words of a respondent –

"Professors like transfer students because they are more hardworking and focused than the rest of the students."

During our conversations, some students also indicated that self-confidence and self-esteem play an important role to overcome barriers. In the words of a few respondents –

"I am much older and more mature than I would have been if I started this program out of high school. I know what I want, how to achieve it, and I have the tools to do so"

"If I hadn't gone to a Community College, I would not have seen a part of myself that I never knew existed. I also could really make a better decision about what I wanted to study and make an impact on my community back home. Lastly, I was able to set an example for my younger siblings and other students that have similar backgrounds to mine that we are able to make it out and make something of ourselves."

Yet, developing a core strength is not enough and there are other factors, both situational and contextual that are difficult to overcome. In the words of a respondent –

"Obviously, there is more to just will power, like racism and other biases. However, I am the child of an incarcerated person, ACOA and daughter of a heroin addict, I have worked to support myself since I was 17 and I have gotten straight A's in school and never paid rent late because I work really hard and used strategies that most savvy poor people have had to use to get ahead."

Through these conversations, we realize that a stronger mindset is definitely very important, but not enough. Not every individual has the will power and strength to overcome the deleterious effects of stereotype threat. There are also some individuals who also strongly believe nothing can be done. We would most likely expect the latter individuals to feel the threat more intensely. We use these concepts to develop the Core Strength construct map. The five construct map levels ranging from a *chaotic mindset* to an *extremely positive mindset* attempt to place everyone at a level where their core strength is directly aligned to. We develop items that target some key principles such as self-esteem, self-efficacy and resilience which have a strong influence on the core strength. The scoring of the items was aligned to the construct map levels. Thus, with careful alignment of the construct map, items and scoring guide, we could establish content validity evidence for this dimension.

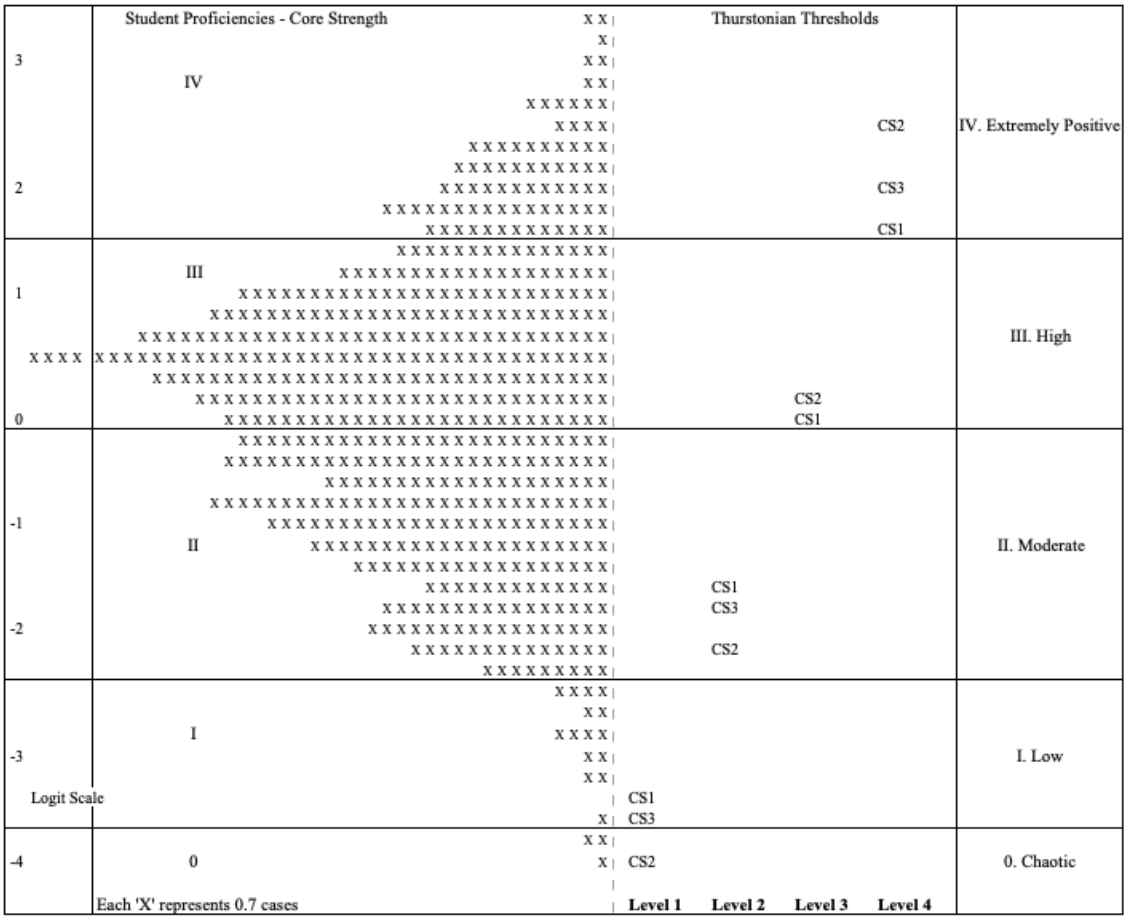
***Evidence based on Response Processes*:** Overall, students found the survey items easy to comprehend. Based on the feedback from the think-aloud interviews, we decided to place the core strength items at the start of the survey. This is because we did not want their answers in any other dimension to influence their responses to the items of this construct. For example, we suspected that a poor domain identification might lead to some lower responses in the Core Strength construct which is something we wanted to avoid. This decision of placing Core Strength items to appear first in the sequence of items was made after incorporating feedback from students on the correct ordering of the items and dimensions.

Evidence based on Internal Structure: We present the Wright map of the Core Strength Dimension in Figure 17. Overall the banding is clear and distinct providing internal structure evidence. We used theoretical and empirical reasoning as mentioned previously to place one of the items with only three thresholds (four response options) at Level 1, Level 2 and Level 4 of the construct map. On the person side, the respondents show a roughly normal distribution

between a range of around -4 logits to +3 logits. We observe most of our respondents lying between -2 and +2 logits. Hence, we can conclude that the range of items are appropriate for this group of respondents. We report the frequency distribution of students in each level in our sample in Table 12.

Figure 17

Wright Map of Core Strength Dimension from Consecutive Unidimensional Analysis



Evidence based on Relation to Other Variables: On average, the strength of mindset of students who indicated their transfer student status as having a positive influence in their lives is 0.592 logits higher than students who indicated otherwise (Table 10). This difference is statistically significant with p-value less than 0.01. Similarly, students who indicated a positive mindset as the key driving factor to overcome obstacles scored 0.847 logits higher than students who indicated otherwise. This difference is statistically significant with p-value less than 0.01. This is consistent with our hypothesis (see Figure 1, Chapter 2) that students with weaker strength of mindset would feel more threatened from the stereotype. However, we find no significant association between Core Strength and the number of courses taken which is again

consistent with what we hypothesized in the stereotype threat balance framework. Thus, these differences lend credibility to our instrument.

Evidence based on Consequences of Use: Given the newness of the instrument, we have found no actual consequences of use yet. We do not recommend usage of this instrument on respondents below eighteen years of age. We do not recommend usage of the instrument to make any characteristic judgements about individuals. The purpose of the core strength dimension is to gauge an individual's ability to overcome stereotype threat.

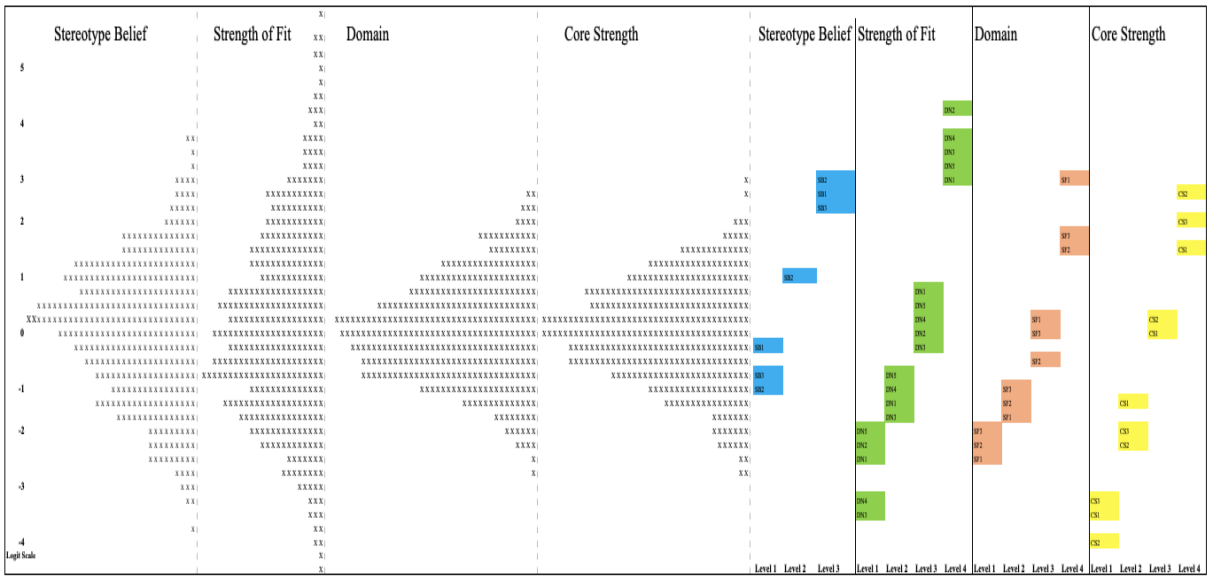
Fairness: We attempt to ensure that the items in this dimension are unbiased and standardized, not only for different demographic subgroups, but also for different stereotype threat conditions. This is the only dimension that can be used for different stereotype threat situations. By ordering the items of the Core Strength dimension to appear first in the survey, we attempt to ensure that the responses are unbiased and not influenced by the responses to items of other dimensions. We will cover measurement related fairness concerns in the subsequent chapter.

Internal Structure Across the Four Dimensions

We present Wright map from the multidimensional analysis in Figure 18. The four dimensions are aligned using the delta dimensional alignment technique enabling comparison of students across dimensions. The respondents are distributed between a range of -4.5 to +6 logits on the Strength of Fit dimension. This range is maximum when we compare across dimensions. This means that overall our sample of students have highly varied levels of group identification. We noted, for all dimensions, the variation in person distribution is greater for the multidimensional model than for the consecutive unidimensional model. This is what we theorized in the earlier chapter when we mention the stereotype threat balance framework. The boundary conditions like having stereotype belief, domain identification, group identification and a weak mindset would increase the chances of experiencing the threat, whereas, it is possible for individuals to guard themselves from the harmful effects of stereotype threat by switching off one of the other variables. This leads credibility to our framework and the fact that stereotype threat must be treated as a multidimensional concept, something which a lot of researchers in the past have failed to recognize.

Figure 18

Multidimensional Wright Map of Four Dimensions – Stereotype Belief, Strength of Fit, Domain, Core Strength



In Table 11, we present the disattenuated correlation matrix[1] which stands as effect size evidence of stereotype threat being a multidimensional construct. We see moderate negative correlation of stereotype belief dimension with the other three dimensions. This can be interpreted to mean that: individuals with high belief in stereotype threat would tend to have a relatively weaker core strength and a relatively weaker mindset to overcome stigma, and will also be inclined to try to disassociate themselves from their group identity and domain identity. Alternatively, this could also be interpreted as: a healthy mindset tends to lead to lower belief in the negative stereotype, and to stronger group and domain identification. We see moderate negative correlation of strength of fit with the domain dimension. We also notice that strength of fit strongly correlates with the core strength dimension (0.71). This means that individuals with a stronger core strength will also tend to have higher group identification and/or vice versa. We see moderate positive correlation of domain dimension with core strength domain. This means that individuals with higher domain identification tend to have higher overall core strength. This is what we theorized previously and the empirical evidence is in support of this theory.

[1] Note that the multidimensional model directly estimates these correlations, incorporating measurement error into that estimation—hence, the correlations are disattenuated.

Table 11
Correlation Matrix (Disattenuated)

	Stereotype Belief	Strength of Fit	Domain	Core Strength
Stereotype Belief	1			
Strength of Fit	-0.562	1		
Domain	-0.360	0.487	1	
Core Strength	-0.629	0.710	0.461	1

In Table 12, we present the frequency distribution of the sample in each level of the construct map for the four dimensions. We can successfully determine the number and percentage of students in each level using the Wright Map. The strength of this representation lies in the fact that every respondent can be placed in a level of each dimension based on estimates in the stereotype threat instrument[2]. This helps in determining the extent to which an individual experiences stereotype threat. It also enables useful comparisons across individuals and dimensions.

As we notice in the Table, 53.3% of the sample lies in Level 1 of the Stereotype Belief construct and 65.6% of the sample lies in Level 2 of the Strength of Fit construct. Whereas, 42.9% of the sample lies in Level 3 of the Domain construct and 43% of the sample lies in Level 3 of the Core Strength construct. Thus, an average individual in our sample is low on stereotype belief and strength of fit as compared to domain identification and overall strength of mindset.

[2] Except, potentially, students with large amounts of misfit.

Table 12
Frequency Distribution of Sample in each level of the Construct Map for the four dimensions, Stereotype Belief, Strength of Fit, Domain and Core Strength

	Stereotype Belief N (%)	Strength of Fit N (%)	Domain N (%)	Core Strength N (%)
Level 0	43 (11.0%)	14 (3.9%)	6 (1.7%)	2 (0.5%)
Level 1	208 (53.3%)	53 (15.2%)	55 (15.5%)	11 (2.5%)
Level 2	125 (32.0%)	231 (65.6%)	90 (25.4%)	163 (40.2%)
Level 3	14 (3.6%)	43 (12.1%)	151 (42.9%)	174 (43.0%)
Level 4		11 (3.1%)	51 (14.4%)	55 (13.6%)

CONCLUSION

In the previous chapters, we created a 14-item Guttman scale to measure stereotype threat. We evaluated the instrument on the student group who transfer from a community college to a four-year university seeking better academic prospects. These students are often stereotyped as "not being smart enough" and tend to experience imposter syndrome while in college. These stereotypes may have longterm effects as a few students indicated in their responses. Overall, we surveyed 392 students and conducted 33 interviews and think-alouds. We first tested the psychometric properties of the instrument and concluded that the multidimensional model fit best among those we calibrated. We found one misfit item in the Stereotype Belief dimension. We judge that the misfit could possibly be due to students not being consciously aware of the stereotype. We recommend testing the item with more groups and situations before making decisions regarding keeping, deleting or replacing the item. We found that only a small percentage of students (6.1%) showed misfitting responses indicating some form of untypical response pattern. Since it is only a small percentage, we are not concerned and can conclude that overall, student responses are consistent with the calibration of the items.

Using the six strands of validity evidence (instrument content, internal structure, response processes, relation to other variables, consequences of use and fairness), we presented strong validity evidence in support of our instrument. The Wright Map also shows greater variation of student estimates in the multidimensional model than the consecutive unidimensional model.

Thus, our conceptualization that stereotype threat is a multidimensional construct is additionally supported. The correlation matrix showed moderate associations between our four dimensions, with the stereotype belief dimension being negatively correlated with the rest. This is what we theorized previously– higher stereotype belief will tend to lead to disidentification with domain and social group. And the opposite may also hold true. While the validity evidence is in favor of using the developed instrument to measure stereotype threat, we recognize the need to conduct further research and test the instrument on other groups, situations and larger sample sizes.

Chapter Four

Racial Variation in stereotype threat: true differences or differential item functioning

Racial stereotypes and the threat arising from them have been at the center of stereotype threat research. Before we draw conclusions regarding racial differences in stereotype threat in our transfer students' subgroup, as a final step of validation, we investigate for any potential item bias using differential item functioning. We use IRT methods for DIF detection because they are more sensitive to detecting DIF in a small sample per group and we used IRT methods to obtain sample estimates. We found evidence of DIF in all the three items in the stereotype belief dimension (two in favor of Whites and one in favor of Blacks and Latinos) and one item in the domain construct (in favor of Whites). Overall, we do not feel the instrument is unfair or biased towards a racial group. We feel that a few secondary latent traits (such as self-perception, belief in racial stereotypes, etc.), related to stereotype threat maybe causing racial differences in the way an item is perceived. Finally, we analyze differences in outcomes for the five racial groups, Asians, Blacks, Latinos, Whites and Others (American Indians or Native Americans, Bi-racial, Pacific Islanders). We find that each group has their own strengths and weaknesses with regards to their ability to experience and mitigate the effects of stereotype threat.

INTRODUCTION

In their controversial book, Herrnstein & Murray argued that intelligence and race were genetically linked (Herrnstein & Murray, 1996). Although these claims and the research behind them have been debunked, we still have a long way to go in eliminating the stereotypes and perceptions of academic inferiority that exists with regards to the minority students in United States (Fischer, 2010; Torres and Charles, 2004).

Thus, it is not surprising that the influence of race on academic outcomes is also felt among community college students. As Thayer and Olivo have rightly said, "Our higher education system puts the burden on students to deconstruct the inner workings of a disjointed transfer system" (Thayer & Olivo, 2021). Almost all community college enrolled students indicate interest in pursuing a bachelor's degree in future (NCES, 2008). But not all of them can make it out of the community college. It has been found that, there is a strong association between transfer rate and the race of the student (Schulock & Moore, 2007). Blacks and Latinos have recorded lower rate of transfer nationwide, with Latino community groups experiencing the lowest transfer rate among all racial groups (Shulock & Moore, 2007).

Thus, in this context of racial disparities, our study will be incomplete without looking at racial differences in outcomes of transfer students on stereotype threat instrument. While, we do feel the need to investigate how different racial groups have performed in our four dimensions, the assumption that the instrument is devoid of any inbuilt biases is also flawed and needs to be validated before any comparisons across groups can be made. Thus, we would like to first conduct a thorough investigation of the structure of the instrument and whether each item is invariant across different racial groups. The Standards of Educational and Psychological Testing demands "all assessment takers should be treated equally" (Standards, AERA, APA & NCME, 2014). While fairness is a fundamental thread across all strands of validity, addressing measurement bias is central to fairness in assessment and testing. With the above motivation, in this chapter, we test the hypothesis that our stereotype threat instrument is devoid of measurement bias. We use differential item functioning methods to test the above hypothesis.

Differential Item Functioning

Differential item functioning or DIF occurs when groups of examinees with the same level of proficiency in a domain have different expected performance on an item. DIF studies have gained prominence since the 1980s, perhaps because it is the only singular tool known till date that aids in evaluating tests for fairness and equity (Mapuranga et al, 2008). An early definition of DIF was given by Shepard in her book Handbook of Methods for Detecting Test Bias (Shepard, 1982). She defines DIF (which was called item bias back then) as, psychometric features of an item that can misrepresent the competence of one group (Shepard, 1982). She further articulates, "An item is unbiased if, for all individuals having the same score on a homogenous subtest containing the item, the proportion of individuals getting the item correct is the same for each population group being considered" (Shepard, 1982).

According to Lord's famous definition of DIF- "If each test item in a test had the same item response function in every group, then people of the same ability or skill would have the same chance of getting the item right, regardless of the group membership. Such a test would be completely unbiased" (Lord, 1980).

This is popularly known as the no DIF condition. Lord defines the No DIF condition for a dichotomously scored item as -

$$P(X = 1, G|\theta) = P(X = 1|\theta)P(G|\theta)$$

or equivalently,

$$P(X = 1, G|\theta) = P(X = 1|\theta), \forall\theta,$$

where P(.) is a probability, X is an item indicator variable (1 for correct and 0 for incorrect), G is a group indicator variable, and θ is a latent ability. The No DIF definition above states that the probability of a correct response at the same ability should be the same regardless of G.

In his research, Garfield highlights the most infamous example of cultural bias in an item used in the Scholastic Aptitude Test (SAT) (Garfield, 2006). The analogy item is presented below:

RUNNER: MARATHON

(A) Envoy: embassy
(B) Martyr: Massacre
(C) Oarsman: Regatta
(D) Horse: Stable

Approximately 53% of White students chose C, the correct answer, but only 22% of African-Americans students chose C. (Garfield, 2006). The criticism was that students from lower income families were not exposed to the word "regatta" and thus failed to properly answer the question.

For many years, SATs have been embroiled in controversies with critics claiming that SAT items have been sexually, culturally and racially biased (Rattani, 2016). The purpose of including this information in this paper is not to discredit the SATs, but to highlight the importance of understanding and investigating item bias or DIF.

Test makers, be it in the field of education or psychology, assume that when an assessment is created, and the person proficiencies are obtained, the differences in estimates observed among various sub-groups is only a function of everyone's aptitudes or abilities. To some extent, these inherent differences do contribute towards inequality in scores observed, but as highlighted in the SAT example, bias can be embedded within the item itself causing additional, unexpected

differences in test scores among various sub-groups. The two phenomena mentioned above are often confused and misrepresented. While the former is known as differential impact, the latter is the DIF. Differential impact is a property of the individual (Dorans & Holland, 1992) It is present in an item, as well as the overall test because individuals and groups (such as those defined by race, gender, etc. differ with respect to the developed abilities measured by the items. However, this impact on any given item should be consistent with impact on other items of the same type. Thus, differential impact at the item level is frequently explained by impact across all items of similar type or impact at the total score level (Santelices & Wilson, 2010). In contrast to impact, which can be explained by stable, consistent and expected differences in ability distribution across groups, DIF is unexpected. An item does not display DIF if people from different groups have a different probability to give a certain response. The item displays DIF if and only if people from different groups with the same underlying true ability have a different probability of giving a certain response. Hence individuals with the groups need to be matched with respect to ability or the attribute that the item measures so that comparisons can be made and score equivalence can be established (Mapuranga et al., 2008). To match groups based on ability, researchers either use observed score or unobserved latent variable (Mapuranga et al., 2008).

The goal of an assessment developer is to create assessments free from DIF. DIF is a property of an item and it is possible for two items to depict DIF in opposite directions, the first being in favor of one group and the second being in favor of the other group. Thus, their overall effect may cancel out at the test level. Hence, it is important to study DIF at the item level. In a typical DIF study, we investigate item-wise DIF for two groups, the reference group and the focus group. Typically, reference group is defined as the group which is suspected to have an advantage, while the focal group refers to the group anticipated to be disadvantaged by the test. However, the grouping variable can also be multi-categorical. (Wang, 2008).

Stereotype Threat and DIF

Stereotype threat, treated as a psychological bias, can also be investigated using the properties of DIF. Although the effects of stereotype threat have been termed as a measurement problem by past researchers (Walton & Spencer, 2009), very little work has been done to understand stereotype threat from a measurement perspective, more so from an item's perspective. As DIF is the property of an item, item analyses can make a valuable contribution to our understanding of how stereotype threat affects performance on individual items. Most stereotype threat researchers focus on average performance (Flore, 2018) and a lot of interesting information gets lost. As an example of information loss, in a study on a stereotype threat dataset by O'Brien and Crandall, it was found that women in the control group outperformed women in stereotype threat conditions for difficult items whereas the difference virtually disappeared for easy items (O'Brien & Crandall, 2003). But if one only looks at the average differences among the two groups in overall test, it is possible that we would find non-

significant differences. This argument is like what Steele and his colleagues put forth in their work – "stereotype threat appears on a difficult test" (Steele & Aronson, 1995). But difficulty is more a function of an item than of the test. When we look at a typical test such as GRE, items are usually created in a way that an average student would be able to score 50% of the items correctly. Not all items are equal in terms of difficulty. Thus, when we look at item wise performance of students, we would observe that the stigmatized group would perform relatively poor in difficult items at the same time performing quite well on easier items (O'Brien & Crandall, 2003). A DIF study will help highlight item-specific issues of stereotype threat and help in creation of tests that are less susceptible to stereotype threat and thus fairer.

DIF Detection Methods

There are various processes, both parametric and non-parametric that have been used to test for DIF. Some popular ones include the Maentel Haenzel Procedure (Holland & Thayer, 1988), the standardized p-difference index (Dorans & Holland, 1993), logistic regression (Swaminathan & Rogers, 1990), and item response theory (IRT) models. Within the IRT framework popular models include the Rasch. (Rasch, 1980), the two-parameter logistic (2PL), and the three-parameter logistic (3PL) models (Birnbaum, 1968).

Among these methods, the Mantel-Haenszel (MH) procedure, standardized p-difference index and logistic regression are based on observed scores whereas item response theory models assume an unobserved latent variable. There are several advantages and disadvantages for each of these methods. MH methods, which provides a chi-square test of significance is conceptually simple but it not designed to detect non-uniform DIF. It also uses total scores as a substitute for latent traits (Meredith & Millsap, 1992). Zwick mentioned it is unwise to use the MH approach to short tests because the low reliability of the short test would lead to larger Type 1 error (Zwick, 1990). The standardized p-difference approach is like the MH procedure with similar issues.
The logistic regression while sensitive to both uniform and non-uniform DIF, faces similar issues as the MH model – the use of total scores as a proxy for latent trait.

For this research, we use an IRT method of DIF detection. Item response theory models use a logistic function that estimates the group-by-item interaction and tests the null hypothesis that the interaction is zero. We choose IRT models for various reasons. Firstly, in IRT the performance of a respondent is a manifestation of an underlying latent trait which is more powerful than using observed score methods. Using the latent trait method, IRT models can test the null hypothesis of no DIF using item response function differences. Secondly, in the IRT framework item parameters are less confounded with sample characteristics (Hambleton et al., 1991) than in observed score DIF methods, thus allowing for better control for differences in the mean ability levels. Thirdly, IRT methods are more sensitive to detecting statistical significance

when we have a small sample per group (Paek & Wilson, 2011). Finally, we have also used IRT methods to obtain estimates of our sample, thus using IRT for DIF detection ensures coherence with the previous chapters.

Sample

We derive the sample from our previous study consisting of 392 community college transfer students to investigate DIF in our stereotype threat instrument across various racial groups. Although not mandatory, students were asked to provide information with regards to their race. We use this information to test for DIF. The five racial categories we consider for this study are, (1) students who identify as White, (2) students who identify as Asians, (3) students who identify as Blacks (4) students who identify as Latino, and because of low representation all other racial groups are categorized into, (5) students who identify as Others. Within the "Others" category, we include, American Indian or Native American students, Bi-racial students, Pacific Islander students and all those who chose the category "Other". Within the race category, we use students who identify as Whites as the reference group or the comparison group. We investigate DIF for Latinos, Blacks, Asians and Other racial groups, using each group as the focal group in turn.

METHODS

We extend the multidimensional partial credit model discussed in the previous chapter, to include the DIF parameter. The equation is as follows -

$$\log\left(\frac{P\left(Y_{pi}=r\right)}{P\left(Y_{pi}=r-1\right)}\right)=(\theta_p-\delta_{ir}-\gamma_i G) \qquad (1)$$

where Y_{pi} is the score of person p on item i, θ_p is the ability parameter, δ_{ir} is the item difficulty parameter for step r, γ_i is the DIF index parameter for item i and G indicates either the reference group or the focus group. G=1 if g is the reference group and G=0 if g is the focal group.

Under this model representation γ is called the "DIF parameter" which is the difference between the item difficulty of the focal group and the reference group. (i.e $\gamma = \delta_F - \delta_R$). Note that this population modeling takes care of the overall group ability differences, called "impact", which should not be confounded with DIF. This impact modeling can also be specified by latent regression of θ onto the grouping variable.

$$\theta = \beta_0 + \beta_1 Y + e \qquad (2)$$

To compare person estimates across dimensions, we use the delta dimensional alignment technique (See Equation 7 & Equation 8, Chapter 3). Using this technique, we obtain the transformed item and step parameters and run the final multidimensional DIF model with new parameters as anchored values. We use this final model for DIF detection.

We use the Conquest software to estimate the DIF model. Conquest can estimate DIF when grouping variable is polytomous as is the case in our analysis. Hence, we obtained DIF estimates for each group through a single analysis using White as the reference category. We look at two statistics from the DIF results. The first is the effect size which indicates the practical significance of the difference between two groups. For a Rasch model, DIF effect size is the DIF parameter (Paek & Wilson, 2011). Paek and Wilson adopted the ETS DIF criteria (Longford et al., 1993) If the DIF parameter is less than 0.426, it is labelled as negligible DIF, if the DIF parameter is less than 0.638 but greater than 0.426, it is labelled as moderate DIF, if the DIF parameter is greater than 0.638, it is labelled as high DIF. Second, we also look at the statistical significance which is obtained by using a z-test. For the z-test, by incorporating the DIF paramter γ in the item response function, the Rasch DIF model estimates γ and its SE directly from the data, so the z-test can be carried out with ease (Paek & Wilson, 2011).

RESULTS

General Psychometric Properties

Before we dive into DIF analysis, it is important to review the general psychometric properties of the instrument. We specifically examine the reliabilities, person and item distribution, person and item fit statistics. These statistics reveal how well the assessment functions within each group.

Reliability – The reliabilities obtained using the MRCML model for DIF detection in a polytomous grouping variable for each of the dimensions are given in Table 13.

Table 13
Reliability

	Stereotype Belief	Domain	Strength of Fit	Core Strength
Reliability	0.591	0.890	0.709	0.742

We obtained high reliability values (>0.7) for all dimensions except for stereotype belief. This pattern is like what we found in our previous analysis (Chapter 3). One interpretation for the lower reliability for the Stereotype Belief construct could be because the latent variable itself is a difficult one for people to think about with respect to themselves.

Item and Person Fit – None of the items in our model showed misfit. Consistent with our findings in Chapter 3, we found that nearly 6% of the respondents showed high fit, 25% of the respondents show moderate fit and 69% of the respondents showed low fit. These values are within accepted range and not a cause of concern.

Item Distribution Statistics – We examine the item distribution by racial groups to see whether the item difficulty levels match the respondent's ability levels. In Table 14 we present the item distribution statistics by race. The estimates for item difficulties are obtained from the MRCML model by including an additional parameter denoting the grouping variable. We used the delta dimensional alignment technique to transform item and person parameters to obtain comparable estimates across dimensions. We look at the most difficult item to agree with and the easiest item to agree with for all groups across all dimensions. We see maximum variation in item difficulty by groups in the Domain dimension: the most difficult item experienced by Latinos was at a difficulty level of 1.39 logits while for the Blacks it was at 0.32 logits. The least variation in item difficulty by groups is observed in the Strength of Fit dimension, where we see a nearly uniform range of item difficulties.

Table 14

Item Distribution Statistics

Race/ Ethnicity	Stereotype Belief			Domain			Strength of Fit			Core Strength		
	Easiest Item	Most Difficult Item	Δ	Easiest Item	Most Difficult Item	Δ	Easiest Item	Most Difficult Item	Δ	Easiest Item	Most Difficult Item	Δ
Asian	-0.42	0.45	0.87	-0.85	1.16	2.01	-0.09	0.12	0.21	-0.19	0.12	0.31
Black	-0.52	0.45	0.97	-0.26	0.32	0.58	-0.08	0.11	0.19	-0.21	0.14	0.35
Latino	-0.51	0.65	1.16	-0.76	1.39	2.15	-0.36	0.26	0.62	-0.38	0.43	0.81
White	-0.41	0.40	0.81	-0.44	0.54	0.98	-0.25	0.14	0.39	-0.06	0.04	0.10
Other Racial Groups	-0.22	0.21	0.43	-0.76	0.67	1.43	-0.30	0.52	0.82	-0.53	0.45	0.98

Person Distribution Statistics - In Table 15 we report the person distribution statistics by race. We observe, on average, students who identified as Black, have the highest stereotype belief (0.30 logits), followed by Latinos (0.03 logits), Whites (-0.01 logits) and other racial

groups (-0.01 logits). Asians have the lowest stereotype belief (-0.30 logits). This pattern is not surprising. Historically, Blacks and Latinos have faced more severe consequences of negative stereotypes with regards to their race than the rest of the groups. The fact that this pattern also emerges in our analysis, reconfirms the fact that for most individuals, race is central to their identity and it is often difficult to isolate it from other aspects of the identity, in this case, the transfer student's identity (Schulock & Moore, 2007).

Whites and other racial groups have the highest domain identification (0.45 logits and 0.17 logits respectively). Asians and Latinos have the lowest domain identification. (-0.40 logits and -0.67 logits respectively).

On strength of fit, Blacks reported the highest mean estimates (0.30 logits) in comparison to other groups and Asians reported the lowest (-0.40 logits).

The Latinos in our sample had a higher core strength (0.63), followed by Asians (0.44), Blacks scored lowest on core strength (-0.18)

Thus, we see interesting differences in estimates across all four dimensions based on race. Each of these groups discussed above, have their own strengths and weaknesses and ways to mitigate the effects of stigma. Looking at stereotype threat through a multidimensional lens allows us to appreciate the different patterns.

Table 15
Person Distribution Statistics

Race/Ethnicity	N	Percentage	Stereotype Belief	Domain	Strength of Fit	Core Strength
Asian	46	11.8%	-0.30	-0.40	-0.30	0.44
			(1.39)	(1.81)	(1.40)	(1.59)
Black	34	8.7%	0.30	0.07	0.30	-0.18
			(1.19)	(1.46)	(1.18)	(1.73)
Latino	39	10.0%	0.03	-0.67	0.03	0.63
			(1.45)	(2.02)	(1.45)	(1.33)
White	186	47.9%	-0.01	0.45	-0.01	0.01
			(1.40)	(2.15)	(1.40)	(1.70)
Others	83	21.4%	-0.01	0.17	-0.1	-0.27
			(1.27)	(2.24)	(1.27)	(1.41)

Differential Item Functioning

We conducted a DIF analysis using "White" as the reference category. We report and discuss the results separately for each dimension.

Stereotype Belief

We found all items in the stereotype belief dimension exhibiting DIF. While the items SB1 and SB2 (see Table 16 for complete item) exhibited moderate DIF in favor of Whites in comparison to Latinos and Blacks, SB3 depicted large DIF in favor of Latinos and Blacks. The "Other racial group" category did not depict any significant DIF in any of the dimensions. The DIF effect size on all the three items was found to be significant.

We do notice a pattern here. The first two items SB1 and SB2 which exhibit DIF in favor of Whites are focusing on individual's feelings with regards to being judged by others. The third item SB3 which exhibits DIF in favor of Latinos and Blacks focuses on individual's views with regards to the stereotypes associated to the transfer student community. While the first set of items are self-focused, the third item is group-focused. SB3 is relatively more difficult than the other two items SB1 and SB2. It is the fourth-most difficult item of the instrument. Thus, we see that the more difficult item favors Latinos and Blacks, while the easier items favor Whites. Through our conversations with individuals, we did get an indication that the non-Asian minority transfer students are more conscious about their race than the Whites, resulting in less stereotype belief with regards to the transfer student community, but overall, a stronger feeling of being victimized and judged by others. In the words of a Latino student who transferred to an esteemed university ABC (removed to maintain anonymity),

"I feel that ABC isn't for a Brown woman from a Latino community like me. I already feel burned out and burdened by financial responsibilities."

In the words of a White transfer student from the same university,

"I am transfer student, White, gay, from lower income background and proud of all my identities. I am also a recipient of a scholarship (thankfully) that alleviated the financial stress. But I know not all students have been as lucky as me"

Table 16

List of items exhibiting DIF for different groups across four dimensions – Stereotype Belief, Domain, Strength of Fit and Core Strength.

Items exhibiting DIF	Dimension	Group	Effect Size	Statistical Significance
SB1. Which is the one statement that best describes you – A. My being a transfer student does not influence what people think of me. B. Some people judge me based on my transfer student status. C. People from other groups almost always interpret my behavior based on me being a transfer student.	Stereotype Belief	White-Latino (in favor of White)	0.55 (Moderate)	Significant
		White-Black (in favor of White)	0.92 (High)	Significant
SB2. Which is the one statement that best describes you – A. I almost always feel myself to be a victim of the stereotypes that are associated to the transfer student community. B. I often feel that I am a victim of the stereotypes that are associated to the transfer student community. C. I do not notice whether people treat me as a victim of the stereotypes that are associated with the transfer student community D. I never feel the stereotypes that are associated to the transfer student community to be also true about myself.	Stereotype Belief	White-Latino (in favor of White)	0.52 (moderate)	Significant
SB3. Which statement best describes your views - A. Most of the stereotypes about transfer students are true. B. Some of the stereotypes about transfer students are true. C. None of the stereotypes about transfer students are true.	Stereotype Belief	White-Latino (in favor of Latino)	1.05 (High)	Significant
		White-Black (in favor of Black)	0.87 (High)	Significant

DN2. Which is the one statement that best describes you - A. My skills in data science are poor. B. My skills in data science are average. C. My skills in data science are above average. D. My skills in data science are excellent.	Domain	White-Asian (in favor of White)	0.62 (Moderate)	Not Significant
		White-Latino (in favor of White)	0.82 (High)	Significant
SF3. Which is the one statement that best describes you - A. Being a transfer student is not a part of who I am. B. I am not sure if being a transfer student is a part of my identity. C. Being a transfer student might be a small part of my identity. D. Being a transfer student is a part of my identity. E. Being a transfer student is a big part of my identity.	Strength of Fit	White-Latino (in favor of Latino)	0.5 (Moderate)	Not Significant
CS3. Which is the one statement that best describes you - A. I am unable to snap back when something bad happens. B. I have a hard time making it through stressful events. C. It takes time, but eventually I do get over set-backs D. It does not take me long to recover from failures.	Core Strength	White-Latino (in favor of Latino)	0.5 (Moderate)	Not Significant

Domain

Only one item exhibited DIF (see Table 16 for complete item). Item DN2 depicted significantly high DIF in favor of Whites when comparing Whites and Latinos and non-significant moderate DIF in favor of Whites, when comparing Whites and Asians. This item is the most difficult item of the instrument. The WLE estimates for Latinos in this dimension is the lowest among all groups (see Table 15).

While, the domain dimension captures individual's overall interest and engagement with data science, this is the only item that particularly focusses on how individual feels about his/her skills in data science (self-perception). Influence of race on self-perception has been discussed widely in literature (Davis et al., 2006). In an exploratory study by Eckberg, it was reported that students of color are significantly more anxious in research methods courses than their White classmates possibly because of a more aggravated feeling of imposter syndrome (Eckberg, 2015). This study was conducted using only two racial groups – Whites and non-Whites. Overall, it could be that the pride in achievement associated with data science is counterbalanced with self-doubt, more so for the Latino students. This pattern was also observed in the stereotype belief dimension.

Strength of Fit

Item SF3 depicted moderate DIF in favor of Latinos when comparing Whites and Latinos. However, the difference was not statistically significant. Since the effect was only borderline moderate (0.5 logits) and not statistically significant, we do not feel this item is problematic from a DIF perspective.

Core Strength

One item CS3, depicted moderate DIF in favor of Latinos when comparing Whites and Latinos. This difference is not statistically significant. We present the item in Table 4. Since the effect was only borderline moderate (0.5 logits) and not statistically significant, we do not feel this item is problematic from a DIF perspective.

Fairness

Although the presence of DIF is a signal that an item may be biased, it does not guarantee that the item is unfair. One interpretation is that the presence of DIF indicates the existence of a latent trait (secondary) besides the one of primary interest (Martinkova et al., 2016). It is possible that such a secondary latent trait is required or is important for the instrument, even if the reference and focal groups perform differently.

As an example, it was found recently, in a biology admission test for a medical school, one item on childhood illness depicted DIF in favor of women (Martinkova et al., 2016). The explanation given for this difference was that women spend more time with children than men. The faculty still considered the item to be fair because medical experts need to be familiar with childhood illness. Thus, the secondary latent trait, "knowledge of childhood illness", was related to the primary concept being tested "knowledge of biology in medicine"

In a similar vein, we conclude that the stereotype threat instrument is fair. There are many secondary traits pertaining to different groups that are driving DIF. We discussed some of

these, self-perception, a stronger belief in racial stereotypes, etc. These are all related to stereotype threat. A more careful qualitative and original analysis would be required to understand all the traits that are driving these differences.

Impact

In the previous section, we concluded that the instrument, despite showing DIF, is a fair measure of stereotype threat. Thus, we can use the instrument to draw conclusions regarding racial differences in outcomes. We present the results from the latent regression IRT model in Table 17. Again, we are using White as the reference category. We notice, on average, students who identify as Blacks have the highest stereotype belief and those who identify as Asians have the lowest stereotype belief. This difference is statistically significant. On the domain dimension, on average, Asians have significantly higher domain identification than Whites. On strength of fit, on average, Blacks have significantly higher estimates than Whites. On core strength dimension, on average, Latinos and Asians have significantly higher estimates than Whites.

Table 17

Differential Impact of Stereotype Threat Instrument based on Race

Regression Variable	Stereotype Belief	Domain	Strength of Fit	Core Strength
Asian	-0.403 (0.267)	-0.821 (0.381)	0.293 (0.254)	0.634 (0.307)*
Black	0.351 (0.267)	-0.483 (0.381)	-0.286 (0.254)	-1.778 (0.307)
Latino	0.045 (0.301)	-1.196 (0.429)	0.026 (0.286)	0.616 (0.346)
Others	-0.031 (0.197)	-0.238 (0.281)	-0.027 (0.187)	-1.562 (0.227)

**** p<0.01, ** p<0.05, * p<0.1*

In the previous section, we concluded that the instrument, despite showing DIF, is a fair measure of stereotype threat. Thus, we can use the instrument to draw conclusions regarding racial differences in outcomes. This has been described as differential impact, which captures the inherent differences in person estimates among groups. We use a latent Rasch regression model (Equation 6, Chapter 3) which moves beyond comparing mean scores and directly estimates

differences from item response data. We present the results in Table 17 where we are using White as the reference category. We notice, on average, students who identify as Latinos and Blacks have the highest stereotype belief. On average, Whites have the highest domain identification. On strength of fit, on average, Asians have the highest estimates, Blacks have the lowest. On core strength dimension, on average, Latinos and Asians have higher estimates than Whites, Blacks have the lowest.

Thus, we can see that each racial group has its own strengths and weaknesses on the four dimensions. On average, Black students have the highest stereotype belief, highest group identification, moderate levels of domain identification and the lowest core strength. This would make them the group that is most susceptible to stereotype threat. Latino students on the other hand with their high estimates on core strength, can shield themselves from the effects of stereotype threat. However, their domain identification is the lowest. It could be that these students have already disidentified with the data science domain and thus experiencing overall lower susceptibility to stereotype threat. Asians, on average, with their relatively higher core strength and lowest stereotype belief also experience lower chances of experiencing stereotype threat. Whites are somewhere in the middle of the spectrum for each of the dimensions. This would make them the group that is least susceptible to stereotype threat. Other racial groups, on average, have weaker core strength in comparison to the rest of the sample. They also have lower group identification. They are susceptible to stereotype threat.

CONCLUSION

DIF is an integral part of validity analysis. Yet, it is often ignored by assessment users. Interpreting group differences without incorporating DIF studies leads to flawed comparisons. In this regard, DIF is an important contribution in any study of assessment validity and fairness. Overall, in our 14-item Guttman instrument, we found significant DIF with a moderate to high effect size ranging between 0.55 to 1.05. All three items belonged to the stereotype belief dimension. Out of the three items, two of them showed DIF in favor of White students, one, which was also among the most difficult items of the instrument, showed DIF in favor of Latino and Black students. The first set of items favoring the Whites was self-focused, i.e. it captures the extent to which individuals believe the negative stereotype to be true about themselves. The third item favoring the Latinos and Blacks was group-focused, i.e. it captures the extent to which individuals believe the negative stereotypes to be true about the group.

An item depicting DIF does not necessarily mean it is unfair. It could be that it is measuring a secondary latent variable related to the central latent variable (stereotype belief). Thus, we recommend further qualitative analysis to understand these patterns of DIF. A DIF depicting item doesn't necessarily mean it is unfair. Hence, we use the estimates obtained from the instrument to analyze differences in stereotype threat experience among students across different racial groups. Overall, we found asymmetrical pattern of group differences in each of

our dimension. The ordering of racial groups is inconsistent, i.e., there appears to be no dominating group in the instrument. Thus, we conclude that each racial group has their strengths and weaknesses and their own ways of dealing with stereotype threat.

Conclusion and Future Research

Three decades of research have demonstrated the deleterious effects that stereotype threat can exert in the lives of everyone, especially those who are deemed as minority in various situations. Some groups have been experiencing the repercussions of being constantly stereotyped, in academic achievement tests, in non-academic activities like sports, even choice of peers and social circles can be altered because of stigma. Stereotypes, in a lot of situations, has become the root cause of violence, in the form of hate crimes, shootings and has even driven individuals to suicide. Thus, stereotype threat known to be a social psychological phenomenon, is not only an important area of research in the field of psychology, but needs a collective, cross functional attention from other individuals in other fields like public health, politics, economics, psychometrics and measurement.

There have been many solutions proposed, especially with regards to how surroundings, institutions and offices can be made more inclusive. Centers and support groups have been established at various levels to support students who are facing adverse consequences of stereotype threat. However, these efforts are less targeted and more focused on spreading awareness at a macro level. Stereotype threat is an individualistic experience and it is more complex than just merely expecting individuals to walk into these support centers and acknowledging that they need help.

Stereotype threat, as we described in this research gets trigged through very complex phenomena. Despite the rich theory on the moderators and mechanisms through which stereotype threat impacts individuals and the surroundings, there remains heterogeneity with regards to (a) the construct (stereotype threat), (b) the methods used for investigation and (c) the measures. This has resulted in an abundance of approaches and lack of consensus on the very fundamental concept, "how do we claim, measure, investigate or detect stereotype threat?" This research is an attempt to take one positive step towards using a uniform, empirical approach in congruence with well-established theory, to measure stereotype threat more uniformly.

In this research, we develop the stereotype threat instrument using the BEAR Assessment System (BAS). BAS, built from three decades of research in assessment design, provides a strong foundation for the stereotype threat instrument. Using BAS principles of assessment development, we first developed a solid understanding of the theoretical framework of stereotype threat. We then interviewed experts from different fields (such as psychology, psychometrics and policy) and conducted multiple focus groups and interviews with a selected sample of students. Thus, the proposed instrument is an amalgamation of the vast literature on stereotype threat, the perspectives of the researchers and the views of experts from various fields

To understand the psychometric properties of the proposed instrument, we engaged in an extensive validation process. While the evidence we gathered lends credibility for usage of the instrument to measure stereotype threat across various situations, domains and for multiple groups, we feel that there is more that needs to be done. We acknowledge that instrument development is an iterative process and the empirical evidence obtained from this project should

be used to refine the instrument further. As next steps, using a larger sample would lend more credibility to the empirical findings. We chose a convenience sample of community college transfer students engaging in data science domain as a first case to test out the functioning of the instrument. In future, we would like to test the instrument using a randomized sample experiencing stereotype threat in different situations. We hypothesized stereotype threat to be a multidimensional construct, comprising of four dimensions - Stereotype Belief, Domain, Strength of Fit and Core Strength. The strength of a multidimensional model is that it allows us to look at an individual's outcomes at the sub-dimensional or micro level. While this information is useful, there are models like the composite model (Wilson & Gochyyev, 2020) that allow us to focus on outcomes both at the sub-dimensional level as well as the overall level. This perspective should be considered.

Lastly, as mentioned previously, individuals have multiple identities. Individuals are capable of shifting focus from one identity to a different identity to protect themselves from the stereotype effects of one. Thus, investigating the functionality of the instrument on multiple identities would give us a holistic perspective on the effects of stereotype threat.

REFERENCES

Adams, R., & Khoo, S. (1996). *Quest [computer program]*. Melbourne, Australia: Australian Council for Educational Research.

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1-23. https://doi.org/10.1177/0146621697211001

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel Item Response Models: An Approach to Errors in Variables Regression. *Journal of Educational and Behavioral Statistics*, 22(1), 47–76. https://doi.org/10.2307/1165238

Adams, R.J, Wu, M.L, Cloney, D., and Wilson, M.R. (2020). *ACER ConQuest: Generalised Item Response Modelling Software* [Computer software]. Version 5. Camberwell, Victoria: Australian Council for Educational Research.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

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

Aronson, J., Fried, C. B., & Good, C. (2002). Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. *Journal of Experimental Social Psychology, 38*(2), 113–125. https://doi.org/10.1006/jesp.2001.1491

Ashmore, R. D., Deaux, K., & McLaughlin-Volpe, T. (2004). An Organizing Framework for Collective Identity: Articulation and Significance of Multidimensionality. *Psychological Bulletin, 130*(1), 80–114. https://doi.org/10.1037/0033-2909.130.1.80

Balgiu, B. A. (2017). Self-esteem, personality and resilience. Study of a student's emerging adults emerging adults group. *Journal of Educational Sciences & Psychology, 7*(1), 93-99.

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology, 71*(2), 364–374. https://doi.org/10.1037/0022-3514.71.2.364

Barney, L. J., Griffiths, K. M., Christensen, H., & Jorm, A. F. (2010). The Self-Stigma of Depression Scale (SSDS): development and psychometric evaluation of a new instrument. *International Journal of Methods in Psychiatric Research, 19*(4), 243–254. https://doi.org/10.1002/mpr.325

Baumgardner, A. H. (1990). To know oneself is to like oneself: Self-certainty and self-affect. *Journal of Personality and Social Psychology, 58*(6), 1062-1072. https://doi.org/10.1037/0022-3514.58.6.1062

Betz, N. E., & Klein, K. L. (1996). Relationships Among Measures of Career Self-Efficacy, Generalized Self-Efficacy, and Global Self-Esteem. *Journal of Career Assessment, 4*(3), 285–298. https://doi.org/10.1177/106907279600400304

Birchwood, M., Trower, P., Brunet, K., Gilbert, P., Iqbal, Z., & Jackson, C. (2007). Social anxiety and the shame of psychosis: a study in first episode psychosis. *Behaviour Research and Therapy, 45*(5), 1025–1037. https://doi.org/10.1016/j.brat.2006.07.011

Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In: Lord, F.M. and Novick, M.R. (Eds.), *Statistical Theories of Mental Test Scores*. (pp 397-479). Addison-Wesley.

Bock, R. D., Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika 46*, 443–459. https://doi.org/10.1007/BF02293801

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35*(2), 179–197. https://doi.org/10.1007/BF02291262

Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review, 15*(1), 1–40. https://doi.org/10.1023/A:1021302408382

Brockner, J. (1979). The effects of self-esteem, success–failure, and self-consciousness on task performance. *Journal of Personality and Social Psychology, 37*(10), 1732–1741. https://doi.org/10.1037/0022-3514.37.10.1732

Brockner, J. (1988). *Self-esteem at Work: Research, Theory, and Practice*. Lexington, MA: Lexington Books.

Brown, J. D., & Dutton, K. A. (1995). The thrill of victory, the complexity of defeat: Self-esteem and people's emotional reactions to success and failure. *Journal of Personality and Social Psychology, 68*(4), 712–722. https://doi.org/10.1037/0022-3514.68.4.712

Brown, J. D. (1998). *The Self*. New York: McGraw-Hill.

Brown, R. P., & Pinel, E. C. (2003). Stigma on my mind: Individual differences in the experience of stereotype threat. *Journal of Experimental Social Psychology, 39*(6), 626–633. https://doi.org/10.1016/S0022-1031(03)00039-8

Cadaret, M. C., Hartung, P. J., Subich, L. M., & Weigold, I. K. (2017). Stereotype threat as a barrier to women entering engineering careers. *Journal of Vocational Behavior, 99,* 40–51. https://doi.org/10.1016/j.jvb.2016.12.002

Casad, B. J., & Bryant, W. J. (2016). Addressing stereotype threat is critical to diversity and inclusion in organizational psychology. *Frontiers in Psychology, 7*(8). https://doi.org/10.3389/fpsyg.2016.00008

Community College Research Center. (2018 & 2019). *Community College Report*. https://ccrc.tc.columbia.edu/Community-College-FAQs.html

Chen, G., Gully, S. M., Whiteman, J. A., & Kilcullen, R. N. (2000). Examination of relationships among trait-like individual differences, state-like individual differences, and learning performance. *The Journal of Applied Psychology, 85*(6), 835–847. https://doi.org/10.1037/0021-9010.85.6.835

Chen, G., Gully, S. M., & Eden, D. (2004). General self-efficacy and self-esteem: Toward theoretical and empirical distinction between correlated self-evaluations. *Journal of Organizational Behavior, 25*(3), 375–395. https://doi.org/10.1002/job.251

Cohen, G. L., & Garcia, J. (2005). "I Am Us": Negative Stereotypes as Collective Threats. *Journal of Personality and Social Psychology, 89*(4), 566–582. https://doi.org/10.1037/0022-3514.89.4.566

Cook, D.A., & Hatala, R. (2016). Validation of educational assessments: a primer for simulation and beyond. *Advances in Simulation, 31*, 1-12. https://doi.org/10.1186/s41077-016-0033-y

Cooley, C. H. (1902). *Human Nature and the Social Order*. New York: Scribner's.

Corrigan, P. W., & Watson, A. C. (2002). The paradox of self-stigma and mental illness. *Clinical Psychology: Science and Practice, 9*(1), 35–53. https://doi.org/10.1093/clipsy.9.1.35

Costa, P. T., Jr., & McCrae, R. R. (1994). Stability and change in personality from adolescence through adulthood. In C. F. Halverson, Jr., G. A. Kohnstamm, & R. P. Martin (Eds*.), The developing structure of temperament and personality from infancy to adulthood* (pp. 139–150). Lawrence Erlbaum Associates, Inc.

Crisp, G., & Nuñez, A. (2014). Understanding the Racial Transfer Gap: Modeling Underrepresented Minority and Nonminority Students' Pathways from Two-to Four-Year Institutions. *The Review of Higher Education 37*(3), 291-320.

Crocker, J., & Major, B. (1989). Social stigma and self-esteem: The self-protective properties of stigma. *Psychological Review, 96*(4), 608–630. https://doi.org/10.1037/0033-295X.96.4.608

Cross, W. E., Jr. (1991). *Shades of Black: Diversity in African-American Identity*. Temple University Press.

Crowe, A., Averett, P., & Glass, J. S. (2016). Mental illness stigma, psychological resilience, and help seeking: What are the relationships? *Mental Health and Prevention, 4*(2), 63–68 https://doi.org/10.1016/j.mhp.2015.12.001

Daley, S. G., & Rappolt-Schlichtmann, G. (2018). Stigma consciousness among adolescents with learning disabilities: Considering individual experiences of being stereotyped. *Learning Disability Quarterly, 41*(4), 200–212. https://doi.org/10.1177/0731948718785565

Davey, T., & Hirsch, T. M. (1991). Concurrent and consecutive estimates of examinee ability profiles. Paper presented at the Annual Meeting of the Psychometric Society, New Brunswick NJ.

Davis, C. III, Aronson, J., & Salinas, M. (2006). Shades of Threat: Racial Identity as a Moderator of Stereotype Threat. *Journal of Black Psychology, 32*(4), 399–417. https://doi.org/10.1177/0095798406292464

De Boeck P., Wilson M. (2004). A framework for item response models. In: De Boeck P., Wilson M. (Eds) *Explanatory Item Response Models. Statistics for Social Science and Public Policy*. Springer, New York, NY. https://doi.org/10.1007/978-1-4757-3990-9_1

Deaux, K. (1996). Social identification. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social Psychology: Handbook of Basic Principles* (pp. 777–798). The Guilford Press.

Devaux, M., & Sassi, F. (2016). Social disparities in hazardous alcohol use: self-report bias may lead to incorrect estimates. *European Journal of Public Health, 26*(1), 129–134. https://doi.org/10.1093/eurpub/ckv190

DeVellis, R. F. (1991). *Scale development: Theory and applications.* Sage Publications, Inc.

Devine, P. G., & Brodish, A. B. (2003). Modern Classics in Social Psychology. *Psychological Inquiry, 14*(3-4), 196–202. https://doi.org/10.1207/S15327965PLI1403&4_3

Dijksterhuis, A., & van Knippenberg, A. (1998). The relation between perception and behavior, or how to win a game of Trivial Pursuit. *Journal of Personality and Social Psychology, 74*(4), 865–877. https://doi.org/10.1037/0022-3514.74.4.865

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Lawrence Erlbaum Associates, Inc.

Eckberg, D. A. (2015). Race and Research Methods Anxiety in an Undergraduate Sample: The Potential Effects of Self-Perception. *The Journal of Classroom Interaction, 50*(2), 145–155. http://www.jstor.org/stable/44735495

Eden, D., & Aviram, A. (1993). Self-efficacy training to speed reemployment: Helping people to help themselves. *Journal of Applied Psychology, 78*(3), 352–360. https://doi.org/10.1037/0021-9010.78.3.352

Embretson, S. E., & Reise, S. P. (2013). *Item response theory.* Psychology Press.

Espenshade, T. J., & Radford, A. W. (2009). *No longer separate, not yet equal: Race and class in elite college admission and campus life*. Princeton University Press. https://doi.org/10.1515/9781400831531

Feuerstahler, L., & Wilson, M. (2019), Scale Alignment in Between-Item Multidimensional Rasch Models. *Journal of Educational Measurement, 56*, 280-301. https://doi.org/10.1111/jedm.12209

Finn, J. D. (1989). Withdrawing From School. *Review of Educational Research, 59*(2), 117–142. https://doi.org/10.3102/00346543059002117

Fischer, M. (2010). Review of the book *No Longer Separate, Not Yet Equal*: Race and Class in Elite College Admissions and Campus Life. *Social Forces 89*(2), 723-725.

Flore, P. (2018). *Stereotype threat and differential item functioning: A critical assessment*. Gildeprint Drukkerijen.

Garfield, L. Y. (2006). The Cost of Good Intentions: Why the Supreme Court's Decision Upholding Affirmative Action Admission Programs Is Detrimental to the Cause. *Pace Law Review, 27*(15).

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*(5), 692–731. https://doi.org/10.1037/0033-2909.132.5.692

Gist, M. E., & Mitchell, T. R. (1992). Self-efficacy: A theoretical analysis of its determinants and malleability. *The Academy of Management Review, 17*(2), 183–211. https://doi.org/10.2307/258770

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology, 59*(6), 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216

Graham, S. (1994). Motivation in African Americans. *Review of Educational Research, 64*(1), 55–117. https://doi.org/10.3102/00346543064001055

Griffin, B.W. (2002). Academic Disidentification, Race, and High School Dropouts. *The High School Journal 85*(4), 71-81. doi:10.1353/hsj.2002.0008.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications, Inc.

Harter, S. (1990). Self and identity development. In S. S. Feldman & G. R. Elliott (Eds.), *At the threshold: The developing adolescent* (pp. 352–387). Harvard University Press.

Hedges, L. V., & Nowell, A. (1995). Sex Differences in Mental Test Scores, Variability, and Numbers of High-Scoring Individuals. *Science, 269*(5220), 41–45. http://www.jstor.org/stable/2889145

Heider, F. (1958). *The psychology of interpersonal relations*. John Wiley & Sons Inc. https://doi.org/10.1037/10628-000

Helms, J. E. (1995). An update of Helm's White and people of color racial identity models. In J. G. Ponterotto, J. M. Casas, L. A. Suzuki, & C. M. Alexander (Eds.), *Handbook of multicultural counseling* (pp. 181–198). Sage Publications, Inc.

Herrnstein, R. J., & Murray, C. A. (1994). *The bell curve: Intelligence and class structure in American life*. Free Press.

Hess, T.M., Hinson, J.T., & Hodges, E.A. (2009). Moderators of and Mechanisms Underlying Stereotype Threat Effects on Older Adults' Memory Performance. *Experimental Aging Research, 35*, 153-177. https://doi.org/10.1080/03610730802716413

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Lawrence Erlbaum Associates, Inc.

Hoyt, C. L., & Blascovich, J. (2007). Leadership efficacy and women leaders' responses to stereotype activation. *Group Processes & Intergroup Relations, 10*(4), 595–616. https://doi.org/10.1177/1368430207084718

Irribarra, D. T., & Freund, R. (2014). Wright map: IRT item-person map with ConQuest integration.

Jackson, P.A., & Seiler, G. (2013). Science identity trajectories of latecomers to science in college. *Journal of Research in Science Teaching, 50*, 826-857. https://doi.org/10.1002/tea.21088

James, W. (1981). *The principles of psychology*. Cambridge: Harvard University Press

Judge, T. A., Erez, A., & Bono, J. E. (1998). The power of being positive: The relation between positive self-concept and job performance. *Human Performance, 11*(2-3), 167–187. https://doi.org/10.1207/s15327043hup1102&3_4

Judge, T. A., Locke, E. A., Durham, C. C., & Kluger, A. N. (1998). Dispositional effects on job and life satisfaction: The role of core evaluations. *Journal of Applied Psychology, 83*(1), 17–34. https://doi.org/10.1037/0021-9010.83.1.17

Judge, T. A., Jackson, C. L., Shaw, J. C., Scott, B. A., & Rich, B. L. (2007). Self-efficacy and work-related performance: The integral role of individual differences. *Journal of Applied Psychology, 92*(1), 107–127. https://doi.org/10.1037/0021-9010.92.1.107

Kane, M. (2006). Content-Related Validity Evidence in Test Development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 131–153). Lawrence Erlbaum Associates Publishers.

Keller, J. (2007), Stereotype threat in classroom settings: The interactive effect of domain identification, task difficulty and stereotype threat on female students' maths performance. *British Journal of Educational Psychology, 77*, 323-338. https://doi.org/10.1348/000709906X113662

Kobrynowicz, D., & Branscombe, N. R. (1997). Who Considers Themselves Victims of Discrimination? Individual Difference Predictors of Perceived Gender Discrimination in Women and Men. *Psychology of Women Quarterly, 21*(3), 347–363. https://doi.org/10.1111/j.1471-6402.1997.tb00118.x

Laanan, F.S. (2001). Transfer student adjustment. *New Directions for Community Colleges, 2001*(114), 5-13. https://doi.org/10.1002/cc.16

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22* 140, 55.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171–196). Lawrence Erlbaum Associates, Inc.

Lopez, C., & Jones, J. S. (2017). Examination of Factors that Predict Academic Adjustment and Success of Community College Transfer Students in STEM at 4-Year Institutions, *Community College Journal of Research and Practice, 41*(3), 168-182. https://doi.org/10.1080/10668926.2016.1168328

Lord, F.M. (1980). *Applications of Item Response Theory To Practical Testing Problems* (1st ed.). Routledge. https://doi.org/10.4324/9780203056615

Major, B., & Crocker, J. (1993). Social stigma: The consequences of attributional ambiguity. In D.M. Mackie & D.L. Hamilton (Eds.), *Affect, cognition, and stereotyping: Interactive processes in group perception* (pp. 345-370). San Diego, CA: Academic Press

Major, B., Kaiser, C. R., & McCoy, S. K. (2003). It's Not My Fault: When and Why Attributions to Prejudice Protect Self-Esteem. *Personality and Social Psychology Bulletin, 29*(6), 772–781. https://doi.org/10.1177/0146167203029006009

Mapuranga, R., Dorans, N. J. & Middleton, K. (2008). A Review of Recent Developments in Differential Item Functioning. *ETS Research Report Series, 1*-32. https://doi.org/10.1002/j.2333-8504.2008.tb02129.x

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

Marx, D. M., Monroe, A. H., Cole, C. E., & Gilbert, P. N. (2013). No Doubt About It: When Doubtful Role Models Undermine Men's and Women's Math Performance Under Threat. *The Journal of Social Psychology, 153*(5), 542-559. https://doi.org/10.1080/00224545.2013.778811

Martinková, P., Drabinová, A., Liaw, Y. L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking Equity: Why Differential Item Functioning Analysis Should Be a Routine Part of Developing Conceptual Assessments. *CBE Life Sciences Education, 16*(2), rm2. https://doi.org/10.1187/cbe.16-10-0307

Masters, G. N. (1988). Measurement models for ordered response categories. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 11–29). Plenum Press. https://doi.org/10.1007/978-1-4757-5644-9_2

McGuire, W. J., & McGuire, C. V. (1981). The spontaneous self-concept as affected by personal distinctiveness. In M. D. Lynch, A. A Norem-Hebeisen, & K. J. Gergen (Eds.), *Self-concept: Advances in theory and research* (pp. 147-171). Cambridge, MA: Ballinger

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57*(2), 289–311. https://doi.org/10.1007/BF02294510

Miller, C. T., & Kaiser, C. R. (2001). A theoretical perspective on coping with stigma. *Journal of Social Issues, 57*(1), 73–92. https://doi.org/10.1111/0022-4537.00202

O'Brien, L. T., & Crandall, C. S. (2003). 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. (1997). Race and academic disidentification. *Journal of Educational Psychology, 89*(4), 728–735. https://doi.org/10.1037/0022-0663.89.4.728

Osborne, J. W. (2001). Testing Stereotype Threat: Does Anxiety Explain Race and Sex Differences in Achievement? Contemporary *Educational Psychology, 26*, 291-310. https://doi.org/10.1006/ceps.2000.1052

Osborne, J. W., & Jones, B.D. (2011). Identification with Academics and Motivation to Achieve in School: How the Structure of the Self Influences Academic Outcomes. *Educational Psychology Review, 23*, 131–158. https://doi.org/10.1007/s10648-011-9151-1

Oyserman, D., Harrison, K., & Bybee, D. (2001). Can racial identity be promotive of academic efficacy? *International Journal of Behavioral Development, 25*(4), 379–385. https://doi.org/10.1080/01650250042000401

Packard, B. W. L., Gagnon, J. L., LaBelle, O., Jeffers, K., & Lynn, E. (2011). Women's experiences in the STEM community college transfer pathway. *Journal of Women and Minorities in Science and Engineering, 17*(2), 129. 10.1615/JWomenMinorScienEng.2011002470

Paek, I., & Wilson, M. (2011). Formulating the Rasch Differential Item Functioning Model Under the Marginal Maximum Likelihood Estimation Context and Its Comparison With Mantel–Haenszel Procedure in Short Test and Small Sample Conditions. *Educational and Psychological Measurement*, *71*(6), 1023–1046. https://doi.org/10.1177/0013164411400734

Pavlova M. A., Weber, S., Simoes, E., & Sokolov, A. N. (2014) Gender Stereotype Susceptibility. *PLoS One, 9*(12). https://doi.org/10.1371/journal.pone.0114802

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

Picho, K., & Brown, S. W. (2011). Can stereotype threat be measured? A validation of the Social Identities and Attitudes Scale (SIAS). *Journal of Advanced Academics, 22*(3), 374–411. https://doi.org/10.1177/1932202X1102200302

Pietri, E. S., Moss-Racusin, C. A., Dovidio, J. F., Guha, D., Roussos, G., Brescoll, V. L., & Handelsman, J. (2017). Using video to increase gender bias literacy toward women in science. *Psychology of Women Quarterly, 41*(2), 175–196. https://doi.org/10.1177/0361684316674721

Pinel, E. C. (1999). Stigma consciousness: The psychological legacy of social stereotypes. *Journal of Personality and Social Psychology, 76*(1), 114-128.

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and Longitudinal Modeling Using Stata (3rd ed., Vol. 1)*. College Station, TX: Stata Press.

Rasch, G. (1960). Studies in mathematical psychology: I. *Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Rasch, G. (1980). *Probabilistic model for some intelligence and achievement tests*. Chicago, IL: University of Chicago Press.

Rattani, S. (2016). SAT: Does Racial Bias Exist? *Creative Education, 7*, 2151-2162.

Rosenberg, M. (1979). *Conceiving the Self*. New York: Basic Books.

RStudio Team. (2015). RStudio: *Integrated Development Environment for R*. Boston, MA.

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

Santelices, M. V., & Wilson, M. (2010). Unfair treatment? The case of Freedle, the SAT, and the standardized approach to differential item functioning. *Harvard Educational Review, 80*(1), 106–133. https://doi.org/10.17763/haer.80.1.j94675w001329270

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., Johns, M. & Barquissau, M. (2004). The Costs of Accepting Gender Differences: The Role of Stereotype Endorsement in Women's Experience in the Math Domain. *Sex Roles: A Journal of Research, 50*, 835–850. https://doi.org/10.1023/B:SERS.0000029101.74557.a0

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review, 115*(2), 336–356. https://doi.org/10.1037/0033-295X.115.2.336

Schmitt, M. T., Branscombe, N. R., Kobrynowicz, D., & Owen, S. (2002). Perceiving Discrimination Against One's Gender Group has Different Implications for Well-Being in Women and Men. *Personality and Social Psychology Bulletin, 28*(2), 197–210. https://doi.org/10.1177/0146167202282006

Schwartz, R., & Ayers, E. (2011). Delta dimensional alignment: Comparing performances across dimensions of the learning progression for assessing modeling and statistical reasoning. Unpublished manuscript, University of California, Berkeley.

Serpe, R. T., & Stryker, S. (2011). The symbolic interactionist perspective and identity theory. In S. J. Schwartz, K. Luyckx, & V. L. Vignoles (Eds.), *Handbook of identity theory and research* (pp. 225–248). Springer Science + Business Media. https://doi.org/10.1007/978-1-4419-7988-9_10

Shapiro, J. R., & Neuberg, S. L. (2007). From Stereotype Threat to Stereotype Threats: Implications of a Multi-Threat Framework for Causes, Moderators, Mediators, Consequences, and Interventions. *Personality and Social Psychology Review, 11*(2), 107–130. https://doi.org/10.1177/1088868306294790

Shepard, L. A. (1982). Defintions of bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias.* (pp 9-30). Baltimore: John Hopkins University Press.

Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science, 10*(1), 80–83. https://doi.org/10.1111/1467-9280.00111

Shih, M., Ambady, N., Richeson, J. A., Fujita, K., & Gray, H. M. (2002). Stereotype performance boosts: The impact of self-relevance and the manner of stereotype activation. *Journal of Personality and Social Psychology, 83*(3), 638–647. https://doi.org/10.1037/0022-3514.83.3.638

Shih, M. (2004). Positive Stigma: Examining Resilience and Empowerment in Overcoming Stigma. *The ANNALS of the American Academy of Political and Social Science, 591*(1), 175–185. https://doi.org/10.1177/0002716203260099

Shih, M., Pittinsky, T. L., & Trahan, A. (2006) Domain-specific Effects of Stereotypes on Performance. *Self and Identity, 5*(1), 1- 14. https://doi.org/10.1080/15298860500338534

Shih, M. J., Pittinsky, T. L., & Ho, G. C. (2012). Stereotype boost: Positive outcomes from the activation of positive stereotypes. In M. Inzlicht & T. Schmader (Eds.), *Stereotype threat: Theory, Process, and Application* (pp. 141–156). Oxford University Press.

Shulock, N., & Moore, C. (2007). Rules of the Game: How State Policy Creates Barriers to Degree Completion and Impedes Student Success in the California Community Colleges. Sacramento: California State University-Sacramento, Institute for Higher Education Leadership and Policy.

Simon, B., & Klandermans, B. (2001). Politicized collective identity: A social psychological analysis. *American Psychologist, 56*(4), 319–331. https://doi.org/10.1037/0003-066X.56.4.319

Smith, J. L., & White, P. H. (2001). Development of the Domain Identification Measure: A Tool for Investigating Stereotype Threat Effects. *Educational and Psychological Measurement, 61*(6), 1040–1057. https://doi.org/10.1177/00131640121971635

Smith, J. L. (2004). Understanding the Process of Stereotype Threat: A Review of Mediational Variables and New Performance Goal Directions. *Educational Psychology Review 16*, 177–206.

Smith, W. A., Hung, M., & Franklin, J.D. (2012). Between hope and racial battle fatigue: African American men and race-related stress. *The Journal of Black Masculinity, 2*(1), 35-58.

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

Spencer, S. J., Steele, C. M., & Quinn, D. M. (2002). Stereotype threat and women's math performance. In A. E. Hunter & C. Forden (Eds.), *Readings in the psychology of gender: Exploring our differences and commonalities* (pp. 54–68). Allyn & Bacon.

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

Steele, C. M. (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., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. *Advances in Experimental Social Psychology, 34*, 379-440.

Steen, L. A. (1987). Mathematics education: A predictor of scientific competitiveness. *Science, 237*, 251-253.

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

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Tajfel, H. (1981). Human Groups and Social Categories: *Studies in Social Psychology*. Cambridge: Cambridge University Press.

Tesser, A., & Campbell, J. (1980). Self-definition: The impact of the relative performance and similarity of others. *Social Psychology Quarterly, 43*(3), 341–346. https://doi.org/10.2307/3033737

Tesser, A. (1988). Toward a self-evaluation maintenance model of social behavior. In L. Berkowitz (Ed.), *Advances in experimental social psychology, Vol. 21*. Social psychological studies of the self: Perspectives and programs (pp. 181–227). Academic Press.

Thayer, S. W., & Olivo, C. (2021, July 8). To close racial equity gaps, make it simpler for community college students to transfer. *EdSource*.

Thoits, P. A. (1986). Social support as coping assistance. *Journal of Consulting and Clinical Psychology, 54*(4), 416–423. https://doi.org/10.1037/0022-006X.54.4.416

Torres, K., & Charles, C. (2004). METASTEREOTYPES AND THE BLACK-WHITE DIVIDE: A Qualitative View of Race on an Elite College Campus. Du Bois Review: *Social Science Research on Race, 1*(1), 115-149.

Trytten, D. A., Lowe, A.W., & Walden, S.E. (2012). "Asians are Good at Math. What an Awful Stereotype" The Model Minority Stereotype's Impact on Asian American Engineering Students. *Journal of Engineering Education, 101*, 439-468.

Van der Linden, W. & Hambleton, R. K. (1997). *Handbook of item response theory*. New York: Springer-Verlag

Vass, V., Morrison, A. P., Law, H., Dudley, J., Taylor, P., Bennett, K. M., & Bentall, R. P. (2015). How stigma impacts on people with psychosis: The mediating effect of self-esteem and hopelessness on subjective recovery and psychotic experiences. *Psychiatry research*, *230*(2), 487–495. https://doi.org/10.1016/j.psychres.2015.09.042

Volodin, N., & Adams, R. J. (2002). The estimation of polytomous item response models with many dimensions. https://research.acer.edu.au/ar_misc/14

Walton, G. M., & Spencer, S. J. (2009). Latent Ability: Grades and Test Scores Systematically Underestimate the Intellectual Ability of Negatively Stereotyped Students. *Psychological Science*, *20*(9), 1132–1139. https://doi.org/10.1111/j.1467-9280.2009.02417.

Wang, W. C. (2008). Assessment of differential item functioning. *Journal of Applied Measurement, 9*(4), 387–408.

Watson, A. C., Corrigan, P., Larson, J. E., & Sells, M. (2007). Self-stigma in people with mental illness. *Schizophrenia bulletin, 33*(6), 1312–1318. https://doi.org/10.1093/schbul/sbl076

Willits, F. K., Theodori, G. L., Luloff, A. (2016). Another Look at Likert Scales. *Journal of Rural Social Sciences, 31*(3).

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Erlbaum.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181-208. https://doi.org/10.1207/S15324818AME1302_4

Wilson, M., Gochyyev, P. (2020). Having your cake and eating it too: Multiple dimensions and a composite. *Measurement*, *151*(1). https://doi.org/10.1016/j.measurement.2019.107247

Wilson, M., Bathia, S., Morell, L., Gochyyev, P., Koo, B. W., & Smith, R. (in press). Seeking a Better Balance Between Efficiency and Interpretability: Comparing the Likert Response Format with the Guttman Response Format. *Psychological Methods.*

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis: Rasch Measurement*. Chicago: MESA Press.

Xavier, L. F., Fritzsche, B. A., Sanz, E. J., & Smith, N. A. (2014). Stereotype threat: How does it measure up? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 7*(3), 438–447. https://doi.org/10.1111/iops.12176

Zwick, R. (1990)/ When Do Item Response Function and Mantel-Haenszel Definitions of Differential Item Functioning Coincide? *Journal of Educational Statistics*, 15(3), 185-197.

APPENDIX

Transfer Students Stereotype Threat Instrument

This study aims to understand the experiences of transfer students studying in a four-year university.

Section 1
This section aims to understand you as a person. Please answer the following questions to the best of your knowledge. Remember there is no right/wrong answer.

CS1. Which is the one statement that best describes you -

a) I have nothing to be proud of.
b) I do not have much to be proud of.
c) I have several good qualities.
d) I am confident about my abilities.
e) I am proud of what I have achieved so far.

CS2. Which is the one statement that best describes you –

a) I never take on a difficult task.
b) I start a difficult task but give up too quickly.
c) I make a positive attempt to complete a difficult task.
d) I can complete a difficult task most of the times.
e) I can excel in any difficult task given to me.

CS3. Which is the one statement that best describes you -

a) I am unable to snap back when something bad happens.
b) I have a hard time making it through stressful events.
c) It takes time, but eventually I do get over set-backs.
d) It does not take me long to recover from failures.

Section 2
This section aims to understand your interest in data science or data science related courses (e.g - statistics, computer science, math, econometrics, etc.)

DN1. Which is the one statement that best describes you -

a) I would rather spend time doing something else than doing data science.
b) I do not care much about being good in data science.
c) I am willing to put in expected hours to excel in data science.
d) I am motivated to do well in data science.
e) I am passionate about doing well in data science.

DN2. Which is the one statement that best describes you -

a) My skills in data science are poor.
b) My skills in data science are average.
c) My skills in data science are above average.
d) My skills in data science are excellent.

DN3. Which statement best describes you -

a) I always do badly in data science.
b) I sometimes do badly in data science.
c) I do not care whether I do well or not in data science.
d) I sometimes do well in data science.
e) I always do well in data science.

DN4. Which is the one statement that best describes you -

a) I think data science is boring.
b) I think data science is a little interesting.
c) I think data science is interesting.
d) I think data science is very interesting.
e) I think data science is a totally engaging field of study.

DN5. Which is the one statement that best describes you -

a) A career in data science would not be a good fit for me.
b) I am not sure if I am interested in data science as a career.
c) I might have an interest in data science as a career.
d) A career in data science could be a good fit for me.
e) A career in data science would be a great fit for me.

Section 3
This section aims to understand your experiences as a transfer student. Please answer the following questions to the best of your knowledge. Remember there is no right/wrong answer.

SF1. Which is the one statement that best describes you -

a) I wish I wasn't a transfer student.
b) I am not proud about being a transfer student.
c) I consider myself a transfer student.
d) I am a proud transfer student.
e) I feel that everyone should want to be a transfer student.

SF2. Which is the one statement that best describes you –

a) I sometimes have negative feelings about being a transfer student.

b) I am not comfortable with anybody knowing that I am a transfer student.
c) I am comfortable with a few of my best friends knowing I am a transfer student.
d) I do not care who knows I am a transfer student.
e) Being a transfer student is an important part of my self-image.

SF3. Which is the one statement that best describes you -

a) Being a transfer student is not a part of who I am.
b) I am not sure if being a transfer student is a part of my identity.
c) Being a transfer student might be a small part of my identity.
d) Being a transfer student is a part of my identity.
e) Being a transfer student is a big part of my identity.

SB1. Which is the one statement that best describes you –

a) My being a transfer student does not influence what people think of me.
b) Some people judge me based on my transfer student status.
c) People from other groups almost always interpret my behavior based on me being a transfer student.

SB2. Which is the one statement that best describes you -

a) I almost always feel myself to be a victim of the stereotypes that are associated to the transfer student community.
b) I often feel that I am a victim of the stereotypes that are associated to the transfer student community.
c) I do not notice whether people treat me as a victim of the stereotypes that are associated with the transfer student community
d) I never feel the stereotypes that are associated to the transfer student community to be also true about myself.

SB3. Which statement best describes your views -

a) Most of the stereotypes about transfer students are true.
b) Some of the stereotypes about transfer students are true.
c) None of the stereotypes about transfer students are true.

Section 4
Additional questions

1. Some people say that a lot can be overcome with the right mindset and will power. Do you think so?
   What helps you navigate the challenges in life?

2. What type of influence does being a transfer student have in your life?
   a) Positive
   b) Neutral
   c) Negative

3. Could you explain why you chose positive/negative/neutral in the question above?

4. How much do you value data science?

5. How many data science related courses/activities are you currently involved in?

6. Are you aware of any stereotypes about transfer students? How does it make you feel?

Section 5
Demographics

1. What is your gender?
   a) Male
   b) Female
   c) Other
   d) Prefer not to answer

2. What is your age?

3. How do you identify racially/ethnically?
   a) Latino/a or Hispanic
   b) Black or African-American
   c) Asian
   d) Pacific Islander
   e) American Indian or Native American
   f) Bi-racial
   g) White
   h) Other

4. Are you a first-generation student?
   a) Yes
   b) No

5. What year did you transfer?

6. What is your expected year of graduation?

7. What are you majoring in?

8. What is your current GPA?

9. Are you an international student?
   a) Yes
   b) No

Generic Stereotype Threat Instrument

This study aims to understand the experiences of [Group], engaging in [Domain], in [Context]

Section 1
This section aims to understand you as a person. Please answer the following questions to the best of your knowledge. Remember there is no right/wrong answer.

CS1. Which is the one statement that best describes you -

a) I have nothing to be proud of.
b) I do not have much to be proud of.
c) I have several good qualities.
d) I am confident about my abilities.
e) I am proud of what I have achieved so far.

CS2. Which is the one statement that best describes you –

a) I never take on a difficult task.
b) I start a difficult task but give up too quickly.
c) I make a positive attempt to complete a difficult task.
d) I can complete a difficult task most of the times.
e) I can excel in any difficult task given to me.

CS3. Which is the one statement that best describes you -

a) I am unable to snap back when something bad happens.
b) I have a hard time making it through stressful events.
c) It takes time, but eventually I do get over set-backs.
d) It does not take me long to recover from failures.

Section 2
This section aims to understand your interest in [Domain]

DN1. Which is the one statement that best describes you -

a) I would rather spend time doing something else than doing [Domain].
b) I do not care much about being good in [Domain].
c) I am willing to put in expected hours to excel in [Domain].
d) I am motivated to do well in [Domain].
e) I am passionate about doing well in [Domain].

DN2. Which is the one statement that best describes you -

a) My skills in [Domain] are poor.
b) My skills in [Domain] are average.

c) My skills in [Domain] are above average.
d) My skills in [Domain] are excellent.

DN3. Which statement best describes you -

a) I always do badly in [Domain].
b) I sometimes do badly in [Domain].
c) I do not care whether I do well or not in [Domain].
d) I sometimes do well in [Domain].
e) I always do well in [Domain].

DN4. Which is the one statement that best describes you -

a) I think [Domain] is boring.
b) I think [Domain] is a little interesting.
c) I think [Domain] is interesting.
d) I think [Domain] is very interesting.
e) I think [Domain] is a totally engaging field of study.

DN5. Which is the one statement that best describes you -

a) A career in [Domain] would not be a good fit for me.
b) I am not sure if I am interested in [Domain] as a career.
c) I might have an interest in [Domain] as a career.
d) A career in [Domain] could be a good fit for me.
e) A career in [Domain] would be a great fit for me.

Section 3
This section aims to understand your experiences as a member of [Group]. Please answer the following questions to the best of your knowledge. Remember there is no right/wrong answer.

SF1. Which is the one statement that best describes you -

a) I wish I wasn't a member of [Group]
b) I am not proud about being a member of [Group].
c) I consider myself a member of [Group].
d) I am a proud member of [Group].
e) I feel that everyone should want to be a member of [Group].

SF2. Which is the one statement that best describes you –

a) I sometimes have negative feelings about being a member of [Group].
b) I am not comfortable with anybody knowing that I am a member of [Group].
c) I am comfortable with a few of my best friends knowing I am a member of [Group].
d) I do not care who knows I am a member of [Group].
e) Being a member of [Group] is an important part of my self-image.

SF3. Which is the one statement that best describes you -

a) Being a member of [Group] is not a part of who I am.
b) I am not sure if being a member of [Group] is a part of my identity.
c) Being a member of [Group] might be a small part of my identity.
d) Being a member of [Group] is a part of my identity.
e) Being a member of [Group] is a big part of my identity.

SB1. Which is the one statement that best describes you –

a) My being a member of [Group] does not influence what people think of me.
b) Some people judge me based on my [Group] membership.
c) People from other groups almost always interpret my behavior based on me being a member of [Group].

SB2. Which is the one statement that best describes you -

a) I almost always feel myself to be a victim of the stereotypes that are associated to the [Group]
b) I often feel that I am a victim of the stereotypes that are associated to the [Group].
c) I do not notice whether people treat me as a victim of the stereotypes that are associated with the [Group].
d) I never feel the stereotypes that are associated to the [Group] to be also true about myself.

SB3. Which statement best describes your views -

a) Most of the stereotypes about the [Group] are true.
b) Some of the stereotypes about the [Group] are true.
c) None of the stereotypes about the [Group] are true.

Section 4
Additional questions

1. Some people say that a lot can be overcome with the right mindset and will power. Do you think so?
    a. What helps you navigate the challenges in life?

2. What type of influence does being a member of [Group] have in your life?

a) Positive
b) Neutral
c) Negative

3. Could you explain why you chose positive/negative/neutral in the question above?

4. How much do you value [Domain]?

5. How many [Domain] related activities are you currently involved in?

6. Are you aware of any stereotypes about the [Group]? How does it make you feel?