Reading between the Lines: Revealing and Resisting the 'Hidden' Gender Bias in Math Story Problems

Gabriella Silva Gorsky

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Min Li, Chair

Angela Ginorio

Katherine Lewis

College of Education

University of Washington

Abstract

Reading between the Lines: Revealing and
Resisting the 'Hidden' Gender Bias in Math Story Problems

Gabriella Silva Gorsky

Chair of the Supervisory Committee:
Associate Professor Dr. Min Li
Education

Educational testing has a long and complicated history in the United States. Currently, state- and district-level assessment systems shape the educational process in a number of ways, from curricular design to teacher evaluation to funding allocation (Au, 2007, 2011; Condie, Lefgren, & Sims, 2014; Harris, 2012; Madaus & Clarke, 2001; Medina & Neill, 1990; Sass, Semykina, & Harris, 2014). This work explores tests as *both* a measure of and a mechanism in the maintenance and spread of inequity. In the first study, 'gold-standard' math story problems are qualitatively coded by three reviewers for the presence of gender bias and stereotypes to reveal patterns in representation of gender. In the second study, a text classifier is developed to parse math story problems by gender and activity. The final piece presents math teacher reflections on making math and math assessment more equitable for students and offers a framework and resources for creating, modifying, and discussing problematic test questions with colleagues and students. The three papers together begin a discussion regarding the goals and impacts of culturally embedded

assessment and challenge many of the assumptions made in past and current research about gender differences in math, what makes assessment equitable, and how we can move forward towards academic justice.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I'd like to thank the teachers. My teachers. The spirits. The ancestors. The elders. The babies and the youth. The trees and moss and sunshine. My fear. My pain. Mother, father. Hot tea. Ice packs. Medicine. Procrastination and anxiety. Crystals, candles, and cards. Clarity and confusion. The pavement and the grass. Wheels and walking and waiting and pacing. Dark days. The best of times and worst. Friends like lovers, like siblings, like family. Partnership, presence, participation. Doubt turned certainty and back again. Reflection, dissociation. Pushing and pulling. Searching for answers and finding more questions. Mentors, masters. Guides of all forms. Music. Mindfulness. Messiness. Mysticism. Clutter and classrooms and dance studios and art supplies. Words of wisdom. Silence. Structure. Blue water, blue sky. Bitten nails. Trauma. Treasures. Invisible scars; invisible selves. The dead. The living. Tonight. Tomorrow. A quiet place. Laundry and dishes. Things left undone. Things yet to begin. Impatience, imperfection. Imposter syndrome. Insomnia. The 11th hour. Dawn. Sleeping in. Staying up. Regrets. No regrets. Divinity. Faith. All that is unknown, unseen. Tears. Rivers. Lakes. The mountains and the plains. Fresh snow. Hot wind. Rose water. Ritual. Rites of passage. Routine. Chaos. Control. Commitment. Comradery. Compassion. Intensity. Softness. The truth. Lies. Sunrise. Sunset. Disappointment. Delight. Tricks and treats. Sweet. Bitter. Sour. Savory. Snuggles. Struggles. Sweet KB. Teeth and claws. Fight or flight. Freedom. Release. Humility. Hunger. Here and now. History. Healing. Waves. Frequencies. Resonance. Dissonance. Conflict, contradiction. Consensus. Connection, collaboration. Community, immunity. The subtle. The obvious. Camping. Camp. Creation. Collapse. Rebirth. Restraint. Restoration. Respect. Models. Measures. Modesty. Math. Maps. The path. False starts. No end. Risks. Rewards. Serenity. Self-love. What has come before. Whatever comes next.

DEDICATION

Past, Present, Future

Chapter 0. ORIGINS

I have long held the belief that a radical social justice framework should provide the foundation for all meaningful educational research and that this epistemology *must* be rooted in a new and broader understanding of "multiculturalism" and the intersections of identity. As I continue down this path of academic and interpersonal discovery, I find that this orientation towards research and educational/social equity still guides the work I do, both on and off campus. The last few years have ushered in a wave of academic policies that have ravaged the landscape of public education. The move towards privatization, especially via charter schools and the voucher system, has meant that the existing chasm of access to high-quality educational experiences that exists between students with vast privilege and those experiencing all-encompassing social inequity has only widened. This makes our work as educational researchers and practitioners that much more important as we look toward the future.

When I began my graduate career, I didn't fully recognize that we were entering a global Data Age. The signs were certainly there, but few of us could have predicted all the ways that "Big Data" would become so fundamental to the ways we exist as citizens. From government and police surveillance to targeted advertising and geo-tracking, few in society can claim any sort of anonymity. Young people today have never lived in a time where this was not true and the effects of this particular acculturation can be seen in critical ways in the educational system. The rhetoric around educational policy has continued on a path of depersonalization and anti-accountability. Coupled with already rampant, but increasingly visible, racism, classism, misogyny, ableism, transphobia, and Islamophobia, our most vulnerable students are put even more at risk of getting "swept into the margins" through the reduction and erasure of their

complex identities to “mere facts and figures, grades, and test scores.” “Achievement” is centered, but at what cost?

Several years ago, writer Dr. Andrew Solomon asserted, “If we tolerate difference, what we’re doing is opening and expanding our idea of humanity.” At the time, Solomon’s words resonated with me deeply. At the time, I didn’t know how *personally important* mental health awareness and disability rights advocacy would become for me. Since then, I have again and again witnessed the vastness of inequity and find that his words fall short. We cannot merely “tolerate” difference! We must embrace and hold close *all* of the things that which makes us unique. And yet! We must also recognize the deep, transgenerational bond that is formed through shared struggle. What makes us unique, also unites us and the power of collective identity should not be undermined within educational systems that historically center individual achievement. It is as crucial as ever that the work I do as an educational measurement researcher and scholar is built on a foundation of authentic celebration for the individual path of discovery that students are on. Not only do education researchers have the “potential to transcend the pages of academic journals and publications to become true catalysts for far-reaching societal change,” it has become a moral imperative. *Good educators must also be activists*. Full stop.

For a long time, I didn’t know how atypical my public-school experiences had been, nor how grateful I would come to be for having grown up in Chicago with the parents I had or the opportunities I was gifted. My overwhelmingly positive early childhood and adolescent experiences both inside and outside the classroom planted a social justice seed in me. I have watched as my elementary and high school classmates and I have grown into compassionate, articulate, kind, critical, civic-minded artists, doctors, lawyers, social workers, builders/makers,

entrepreneurs, and educators. We are the product of public education done well. Was it perfect? Absolutely not. But we were being prepared for authentic engagement in a pluralist future.

My path to doctoral studies seemed much less cohesive when I started my program. The impact of various opportunities for self-expression, leadership, and critical thinking on my identity and skill development felt inconsistent and sporadic. In the time since, as I have been confronted with the many privileges that have brought me to this point, my early "gifted" placement no longer seems irrelevant to my long-term success. As a queer, mixed race non-binary femme raised by a single mother, the narrative of the academy says I *shouldn't be here*, but I am and it is no accident. With each passing year, the myriad ways in which even my very early academic experiences have set me on this path become clearer and clearer. As this crystalizes, so does my commitment to serving the needs of young people, especially those who are the most vulnerable, those whom the system says shouldn't succeed. I absolutely did not get here on my own and neither can they.

I am only beginning to understand the critical ways that my early exposure to the arts and my initial pursuit of an art degree truly did inform the texture of my entire life. Going from the arts to psychology to education is not so disjointed a course to take, after all. Access to arts spaces provided me with alternative frameworks for knowing and being that fundamentally guide the way I interact with educational contexts, content, and people. Even my understanding of the teaching, learning, and application of mathematical principles has been informed by these creativity-centered experiences. Providing students a high quality education means encouraging not one, but many paths to self-actualization and fulfillment, and giving them the tools and networks they need to become whole people. This is especially true for our most vulnerable students.

Queer, disabled, black and brown youth continue to be inequitably served by the education system. Arts programs, among other sites for this sort of deep learning (such as the contemporary "maker" movement in education), provide a space for these young people to explore their intersectional identities, confront the institutional oppressions that perpetuate violence against marginalized communities, and develop deep and complex relationships that will set a life-long standard for their social selves. My challenge is to take these lessons and apply them to educational measurement--to envision assessment that also provides young people opportunities to critique systems, gain deeper self-awareness, and see the diversity of their community represented.

The idea that I could be doing research on gender identity in an educational measurement program was unfathomable to me when I first embarked on post-baccalaureate studies. The two paths seemed irreconcilable, but I was wrong. In reality, it is critical that we think deeply about the interplay of various facets of identity with how and why we measure student characteristics, especially high-stakes test performance. The aim of my current research is to expose the ways gendered assumptions and expectations are inscribed into our speech, into our systems of accountability, and into our academic institutions. Through integrating feminist and queer methodologies and looking to novel approaches to data analysis, I have been able to rethink the questions posed in doing assessment development work. I am driven by the connections we build when we work collaboratively, not only across disciplines, but also across social, cultural, and other structural divides, to solve complex problems like those of persistent gender inequity and pervasive cis-hetero-sexism.

0.1 Methodological Mission Statement

By leveraging my academic privilege and analytic skills, I hope to serve as a liaison between grassroots knowledge producers (i.e., teachers, parents, and students) and formal educational spaces. Top down approaches to equity and justice don't work. Further, while academic achievement along a normative trajectory provides invaluable social capital, I can no longer tolerate using this as *the* measure of success. Life satisfaction is multidimensional. Identity is multidimensional. My work as an educational measurement professional must also be multidimensional. It must recognize the inherently multifaceted identities of the students whose lived realities inform my research, and center them as the only true authorities on their experience.

Chapter 1. INTRODUCTION

From the early days of modern education as we know it, interest in gender differences in academic performance has steered a large portion of the educational research literature. Across subjects and contexts, researchers have been very invested in identifying and understanding gender differences in how people learn and display learning (Hyde & Linn, 2006; Voyer, Voyer, & Hinshaw, 2014). This is especially true when it comes to the science, technology, engineering, and math, frequently referred to as 'STEM'. This cluster of subjects and disciplines has long been plagued by a pervasive gender disparity in performance and professional attainment (Gayles, 2011; NSF, 2016; Shapiro & Williams, 2012; Smeding, 2012). Starting as early as kindergarten, boys tend to outperform girls in these subjects in school, as well as in skill-based tests (Cimpian et al., 2016). For decades, researchers across disciplines have studied this phenomenon.

While this shift towards a focus on gender equity in STEM has produced a number of impactful programs and interventions aimed at getting women, girls, and femmes[1] interested in math and related disciplines, the gender disparities in these fields endure (Gayles, 2011; NSF, 2016; Shapiro & Williams, 2012;). Further, while more women and femmes are gaining access to these spaces and fields, they are still largely in the minority when they get there. Despite inconclusive findings in research on gendered disparities in math performance and its underlying mechanisms, the long-term impacts are less ambiguous (Gayles, 2011; NSF 2016; Settles, O'Connor, & Yap, 2016; Smeding, 2012) In 2006, approximately 46% of PhDs in biology were women, 25% of PhDs in physical science, and 15% in engineering. Further, a mere 30% of

[1] *Femme* is a term used to describe people whose gender expression and/or internalized experience is defined in some way by a relationship with femininity and things associated with the feminine. People who fit within the framework of femme identity may claim a range of specific gender identities. While women and girls are targeted within a patriarchal system, so are those with proximity to or who are legible within contemporary or historical schemas of femininity

assistant professors in biology, 16% in physical science, and 17% in engineering were women (Hyde and Linn, 2006). In the decade since those statistics have been published, remain relatively stable. While women earn more than half of all bachelor's degrees and about half of all science and tech degrees, this is not reflected at the sub-program level with women earning only 17.9% of computer science degrees, 19.3% of engineering degrees, 39% of physical sciences degrees, and 43.1% of mathematics degrees (NSF, 2016). Assessment researchers are only beginning to address this dissonance as something more deeply rooted in our collective consciousness than "sex difference" research can ever describe and there is a great need for research on how such sexist and toxic ideals enter our psyches.

Almost contradictorily, there is already a lot of research on gender effects ("differences") in math assessment (Casad, Hale, & Wachs, 2017, Cimpian et al., 2016; Hyde & Linn, 2006; Voyer et al., 2014). Much of this includes research on stereotype threat, typically comparing different racial/ethnic groups across different settings, among other factors (Ambady, et al., 2001; Similarly, Amit, & Fried, 2005; Casad, Hale, & Wachs, 2017; Galdi, Cadinu, & Tomasetto, 2014; Huguet and Régner, 2008). Other research looks at factors predicting differential math achievement and STEM involvement by gender (Cimpian et al., 2016; Hyde & Linn, 2006; Shapiro & Williams, 2012). There is also a lot of research on differential item function (DIF) analysis and other ways of evaluating test item fairness (Gallagher, 1998; Gallagher et al., 1999; Ibarra, 2001; Kan & Bulut, 2014; Li, Cohen, & Ibarra, 2004; Mendes-Barnett & Ercikan, 2006; Ryan & Fan, 1996; Taylor & Lee, 2012). Additionally, there is a growing body of research exploring how the specific linguistic features of test questions impact student performance (Abedi, 2004; 2013).

However, it is rare to find research that explores the manifestations or origins of gender differences in STEM participation and math performance deeply (ecologically through an exploration of community- and society-level impacts on test performance, epistemically through a reframing of goas and assumptions about test performance, and critically through an interrogation of systems of oppression that operate to maintain and reproduce gender bias and achievement inequity), and few researchers, if any, attempt to consider their broader implications. For example, there is a lack of rigorous research on the item development process, especially when it comes to the bias and sensitivity review, despite the fact that these procedures determine later test exploration and procedural documentation publically available (LACCD, 2012; NCES, 2017; Washington State Office of the Superintendent of Public Instruction, 2011; SBAC, 2012). There is also a serious lack of teacher voice in the evaluation of test materials. What do teachers think makes tests problematic? How would teachers solve those problems? Researchers typically offer no clear guidelines for students/teachers to address stereotypes and bias they confront. Further, even less research exists about the implications and impacts of the methods we *do* use to address test bias or the (lack of) training we provide teachers when it comes to test literacy (Deluca & Bellara, 2013; Ellis & Smith, 2017; Sexton et al., 2018).

On the one hand, simply knowing that boys tend to outperform girls on a particular type of exam is valuable information. It tells us that there is some failure in the system, be it a failure of the instrument itself, a failure in the classroom, or a failure of society. What it *doesn't* tell us, is about a failure of girls. However, outcomes like this, that identify the differential performance of a particular group, are very often interpreted this way: as "proof" of a group *deficit*. Similarly, it is worthwhile to know that certain kinds of test items are more difficult for English Language Learners (ELLs), but what does that finding *mean*? More often than not, findings such as these

become calls for individual student remediation as opposed to calls for a critical examination of the infrastructure of the educational system.

> *Why are boys outperforming girls? Is it due to differential in-class socialization? Is the teacher providing more support to boys? Why? Where does this tendency originate? Is the course text written in a way that is more relevant to boys? Is the material not being presented in a way that is equally accessible to both groups? Are ELLs being supported as multilingual or are they being funneled forcibly through an English-only program? What is reflected in the fact that they are considered "English Language Learners" in the first place? Why aren't there means of assessing students in many different ways such that linguistic proficiencies play less of a role?*

These are just a few of the many questions that can be asked in response to particular (real-world) statistical findings. While concerns about representation, the nuances of written language, or the political and social structures that underscore inequity may be acknowledged within scholarly circles, rarely are they addressed as part of the interpretation of specifically quantitative data in published research reports. There doesn't seem to be room for an interrogation of the educational system in these studies, no space to challenge the workings of an educational system that reproduces the marginalization of girls, students of color, disabled students, LGBT students, mentally ill students, poor students, and even the teachers who embody these identities. This leads me to the overarching question guiding my present research agenda: How can educational assessment be used, not to reproduce the status quo, but as a part of a

project to transform the way measurement tools are used and interpreted in the larger knowledge building process?

Taken together, my three explorations begin to answer questions about the function of the testing process itself in the perpetuation of stereotypes and social bias. Rather than focusing on test performance alone, as much of the existing literature tends to do, even in the service of test equity, I interrogate our tools and our methods. Through this lens, the math test questions I examine in Study 1 shed their assumed neutrality and become sites for understanding the almost innumerable ways in which harmful gender stereotypes are reproduced, even at the item level. Study 2 provides an opportunity to establish the validity and versatility of my approach to item review and position it within current trends in research and assessment, specifically through applying Machine Learning (ML) and Natural Language Processing (NLP) techniques to the process of item review. Lastly, Study 3 provides a reference for math educators, and perhaps even some test developers, to better understand the relative importance of considering the impact of repeated exposure to academic materials inscribed with structural gender inequity on students' academic experience and identity development. The goal is to offer an accessible resource for doing the work of identifying, unpacking, and rescripting the normative gender messages students confront on a daily basis, both in and out of the classroom.

1.1 WHY DO REPRESENTATIONS OF IDENTITY IN TESTS MATTER?

To set the context for this work, it is important to outline certain concepts that form the basis of the interest in gender representation in math and for the concern regarding the impacts of outdated assumptions about who testing serves and about how to make assessment equitable. By

bringing together divergent philosophies and epistemologies, the hope is that the work will have greater reach than if only a single theoretical or methodological track had been chosen.

1.1.1 *Stereotype Threat*

Stereotype threat is a foundational concept in the study of group-based differential test performance. First coined by Claude Steele and Josh Aronson in their foundational paper "Stereotype Threat and the Intellectual Test Performance of African Americans" (1995). Stereotype threat describes a "predicament" whereby our fear of conforming with negative stereotypes about a group to which we belong becomes self-evaluative and self-fulfilling. That is, we begin to judge ourselves against that stereotype and as a result, begin to conform to it. Steele and Aronson's critical finding was that the experience of stereotype threat has real impacts on academic performance. Black students who were primed to think about racial stereotypes underperformed relative both to white students and black students who had not been primed.

Steele and Aronson's foundational research has been replicated by the original authors and others with more diverse samples, looking at effects of not only race, but gender, contextual and environmental factors, and interventions aimed at masking the effects (Ambady, Shih, Kim, & Pittinski, 2001; Similarly, Amit, & Fried, 2005; Casad, Hale, & Wachs, 2017; Galdi, Cadinu, & Tomasetto, 2014; Huguet and Régner, 2008). While the results are typically mixed and the exact mechanism remains unarticulated (Schmader and Johns (2003) explore "minority stress" as one possible explanatory factor), what is clear is that test-takers' feelings and beliefs about their own identity, as well as the relationship of that identity to the task-at-hand (Brown, 2011, Emde, 2006; Gunderson et al., 2012; Hendy, Schorschinsky, & Wade, 2014; McKeller et al., 2019; Tomasetto, Alparone, & Cadinu, 2011; Williams, 1991), *matter* and this can be leveraged (for better or for worse) by test developers, test administrators, and test users. Further, many

stereotype threat studies take for granted the existence and internalization of well-known social stereotypes missing an opportunity to examine and understand where those ideas come from and how they are reproduced, not in a lab, but in society.

1.1.2 *Implicit Attitudes*

Around the same time that Steele and Aronson were researching the impact of stereotype threat, Greenwald and Banaji were investigating *implicit* attitudes. According to the researchers, this describes the way people unconsciously attribute particular qualities to individuals based on group membership due to learned associations which they may receive from family, community, government, media, etc. (Greenwald & Banaji, 1995). This early work led to the subsequent development of the Implicit Association Test, a –computer-based measure that requires individuals to quickly categorize images and words based on a set of predefined rules (Greenwald, McGhee, & Schwartz, 1998). The goal of the test is to identify cognitive social bias on the basis of association speed. The test is currently hosted as a public online social experiment by Harvard University (Project Implicit).

While the validity and reliability of these tests have been widely critiqued, the popularity of the assessment within the lay community demonstrates the salience of "snap judgements" to the greater public understanding of how bias is reproduced from the micro to the macro level (Bronfenbrenner, 1994; Cooley, Payne, & Phillips, 2014). These implicit attitudes are deeply ingrained and quite often enacted unconsciously through explicit stereotyping or via a more insidious type of interpersonal interactions called microaggressions. These same snap judgements and decisions are made by students when they are consuming test materials, rendering the specific language used in those tests important not only from an academic achievement perspective, but from an identity development one, as well.

1.1.3 *Microaggressions*

First conceptualized by Harvard University psychiatry professor Chester R. Pierce in 1970, the term *microaggressions* was a term coined to describe the small, but significant ways African Americans were differentially treated on the campus by peers, instructors, and the very institution, itself. Since then, and increasingly in recent years, the concept of *microaggressions* has begun to take on mainstream linguistic and sociological significance, especially in educational contexts where repeated exposure functions as a form of subliminal bullying.

Sue's (2010) book *Microaggressions in Everyday Life: Race, Gender, and Sexual Orientation* (2010), considers the short- and long-term impact and significance of these "brief and commonplace daily verbal, behavioral, or environmental indignities, whether intentional or unintentional, that communicate hostile, derogatory, or negative… slights and insults toward people" on the basis of perceived race, gender, and/or sexual identity on both individuals and groups. Moreover, there is increasing acknowledgement of the negative impact *microaggressions* can have in terms of reproducing existing biased social structures, as well as contributing to minority stress and its subsequent physical and mental health consequences (Bolte, Brown, Segrist, & Dudley, 2015; Lester, Yamanaka, & Struthers, 2016; Meyer, 2003; Nadal, 2013; Prather & Ruiz, 2015).

1.1.4 *Intersectionality*

It is essential to draw on the foundational work of Black feminist scholar Kimberlé Crenshaw (1991), who highlights the embodiment of intersectionality through an analysis of how violence perpetrated against women (in a myriad of forms) occurs at the intersection of race, class, and gender. Each of these identity markers does not operate in isolation from the others. Rather, they have a symbiotic relationship that not only can't be reduced to one or another individual

characteristic, but which also shifts in meaning across spaces and time. What elevates her argument over those of other scholars of intersectionality is her articulation of three distinct manifestations of intersectionality: structural, political, and representational.

This intersectional lens can easily be applied to item review, and should be. Testing, while experienced by the majority of children and adolescents in the United States, still reproduces vast social inequity. Access to educational resources, academic and social supports, and high quality learning environments and materials is not equitably distributed, and this is undeniably linked with racialized, gendered, and classed identities (Gorski, 2012; Howard, 2010). Within a single classroom, students do not receive the same educational experience. Stereotypes about feminine and masculine people, students of color, non-English speaking students, disabled students, poor and rich students, single-parent students, and all the infinite intersections of these and other [marginalized] identity markers are not challenged, but perpetuated through implicit and explicit differential treatment.

By using intersectionality as a methodology for framing test development and evaluation, the way these particular oppressions are played out can be interrogated in a more nuanced and less reductive way. In regards to Crenshaw's three modes of intersectionality, the educational system is undoubtedly mired in structural intersections of oppression (stemming from the larger systemic oppression that guides the operations of formal schooling), is relatedly subject to the sway of political intersectional oppression, which is manifested through lobbyist groups, state boards of education, the federal Department of Education, and of course presidential mandates, and is further implicated in representational intersectional oppression when we consider stereotypes about the "brainy Asian student," the struggle of femme-identified people to achieve in STEM fields, or the "dumb" jock (who is almost always portrayed as masculine).

1.1.5 *Liminality and Third Space*

Third Space Theory was first articulated by Harvard English and American Literature professor Homi K. Bhabha (2004) to describe the discursive location where two cultures meet. This theory, which derives from Vygotsky's (1962) sociocultural tradition, views this meeting point as a literal and theoretical 'space' infused with kinetic, generative cultural tension. Within this space, the nuanced hybrid identities of people who are not "either/or," but rather "both" become legible. This betweeness can also be described as *liminality*, a concept taken up from Anthropology and increasingly present within Queer and Feminist literature, especially as it pertains to the experiences of people of color and other marginalized groups.

While Third Space Theory has been taken up formally within educational research, it has mostly focused on the role of cultural capital on the development of academic skills within specific content domains (Maniotes, 2005; Levy, 2008; Skerrett, 2010). The way Queer, Indigenous, and Feminist authors have taken up the idea of the Third Space, especially as it relates to the "potentiality" aspect of *liminality*, provides a novel perspective within educational measurement regarding how the location of certain student identities within this tense space might impact how they receive and internalize cultural, in this case, gendered, messages (Pérez, 1999; Boyd & Ramírez, 2013; Driskill, 2016). We know with certainty that students' experiences of gender are not all the same, and yet we apply broad guidelines to the administration of gendered signifiers within testing. What does it mean for students both inside and outside of this Third Space when it comes to gendered identity, when the standards simply reproduce a more or less static status quo?

1.1.6 *The 'Politics of Absence'*

In the introduction to the special issue of *Trans Studies Quarterly,* Rawson (2015) engages what is described as a "politics of absence" (p. 547). Rawson is borrowing this construction from Harrison Apple, one of the contributors to the special issue. In his article, Apple frames his archival analysis on the absence of certain identities from the archives of a particular nightclub. The absence of African American patrons from the collection raised a flag regarding the erasure of particular identities. It was through this exploration that Apple was able to uncover a rich history of transgender[2] gathering and social organizing that happened in that place and time. This analysis of absence helps highlight the challenges facing scholars across disciplines whose work involves examining marginalized identities.

What Apple demonstrates in their research is that it was the lack of documentation that brought to light the erasure of transgender identity. Within the context of academic testing, the fact that various procedures already exist for evaluating items means that uncovering implicit bias requires a reading between the lines. Along this vein, this research was entered into with an awareness that what was *not* present within items would be just as important as what was.

1.2 RETHINKING METHODOLOGIES

Most studies do things "as usual," rather than considering novel approaches. The majority of existing bias/fairness research defaults to binary comparisons (boy/girl, white/non-white, non-ELL/ELL), in order to compare the "other" to the hegemonic "norm" (as in the case of using "reference" and "comparison" groups). The focus tends to be on what is present and measurable,

[2] *Transgender* is a word used to describe the gender identity of someone who does not identify with the gender/sex they were assigned at birth. In contrast, *cisgender* describes someone who does identify with the gender they were assigned at birth.

understandably, but we can't also ignore what is not there. Through this work, I hope to infuse into the existing discourse on the intersection of gender identity and math assessment a more explicit acknowledgement of the impact of repeated exposure to (negative) stereotypes about one's group(s), repeated invisibility of one's group identity, and the fact that negative effects may not be captured in the test score. While differential achievement serves as one potential indicator of identity-group-based social inequity, certain long-term outcomes and social-emotional impacts may not be evident in the scores. One clear example of this is the phenomenon of increased college enrollment of black and brown students, but significantly more barriers to completion compared to white students with comparable academic profiles (Hollands et al., 2012; Wei et al., 2011; Shapiro et al., 2017). Similarly, mental health disparities exist between members of different groups, especially when it comes to access to appropriate services and resources (Daniel, Golberstein, & Hunt, 2009; McGuire & Miranda, 2008; Rosenthal & Wilson, 2012). High achievement and access to opportunity doesn't protect marginalized people from negative long-term outcomes (Chetty et al., 2018).

Through an equity-driven intentional process of revealing implicit gender bias in math story problems, identifying its impact on students, and rethinking item development, the three studies in my dissertation fill several distinct gaps in the measurement, and specifically, test bias, literature:

a. Affirm the pervasiveness of implicit gender bias within math story problem contexts, even among 'gold-standard' item sets.

b. Examine the relationship between the presence of implicit gender bias in math story problems and patterns of test performance.

c. Explore the ways educators and youth understand and interpret the messages they are sent via math story problem contexts.

d. Provide rigorous guidelines for item development, bias identification/evaluation, and instrument improvement that are accessible to a spectrum of stakeholders in student academic assessment.

e. Normalize a collaborative approach to item bias reduction that centers the needs, perspectives, and voices of historically marginalized communities.

The intention of this work is to present a snapshot of the ways gendered (and other socially biased) messages are deeply encoded into the educational materials that students interact with on a regular basis, in some cases daily. While the validated items explored in Study 1 represent only a small subset of all the potentially problematic resources students will encounter, the fact that they *are* stringently evaluated by education, measurement, and psychometric experts made them the ideal location to launch this ongoing inquiry. Even a person who is knowledgeable about educational and assessment systems might be shocked to find out how much is *still* embedded in test questions that "should have been" *objective, unbiased*, and *neutral.* My argument is not that the tests themselves are bad, nor do I claim to know what the best course of action is for addressing these problems at the institutional level. Rather, my hope is to legitimize the questioning and critique of so-called objective approaches to minimizing test unfairness, thus empowering students, teachers, and families to gain perspective about what the tests are really evaluating, as well as the impact they have on the holistic development of young identities.

Chapter 2. RETHINKING CONSEQUENTIAL VALIDITY: LINGUISTIC DECONSTRUCTION OF 'GOLD-STANDARD' MATH STORY PROBLEMS TO UNCOVER IMPLICIT GENDER BIAS

GABRIELLA SILVA GORSKY, NIXI WANG, & DAVID PHELPS

2.1 ABSTRACT

This work represents a novel exploration of the critical measurement unit —the test item—as a means to expose social inequity. To explore the ways implicit gender bias and stereotypes continue to permeate the existing body of standardized test questions, a rigorous procedure for the deconstruction of math story problems was developed. Collaborative coding of the data revealed consistent trends of implicit gender bias manifested in the item set. Story problem contexts were not independent of the implied gender of the characters represented. These findings radically challenge fundamental assumptions made about how the measurement community evaluates test questions for bias and sensitivity. They extend the conversation about testing and measurement to the item development process, especially when it comes to the usual research focus on the use of scores in student promotion/placement, teacher retention, and allocation of funding.

2.2 INTRODUCTION

Since the inception of formalized testing and measurement, especially within psychology and education, researchers have been interested in group differences. In the United States, social

justice movements around issues of gender, race, class, and disability built momentum over the course of the 20th century (Au, 2007, 2011; Condie, Lefgren, &Sims, 2014; Harris, 2012; Madaus & Clarke, 2001; Medina & Neill, 1990; Sass, Semykina, & Harris, 2014). These shifts introduced new questions for the measurement community: Were psychological and educational tests 'fair'? Were they measuring the same traits the same way for different individuals? For different groups? If they weren't, was it due to a flaw in the measurement tool? In the statistical model used? Or did score deviations reflect *real* between-groups differences? The measurement community still struggles to answer some of these questions (Anastasi, 1986; Haertel, 2013; Kane, 2013), but over the last several decades, a number of methodological approaches have been developed to identify issues of fairness and statistical bias at the test and item level (Camilli, & Shepard, 1994; Zumbo, 2007; Zwick, 2012).

Our current models for assessment frequently take as a given that academic success and learning are well captured by test scores. As a result, a lot of effort is put towards improving student performance on assessments, especially those that are considered "high-stakes" (used for important student-, teacher-, school-, and district-level decision-making) (). In order for the performance of a student to be meaningful, we have to have some confidence that their score is comparable to their previous assessment and to that of other students in other groups. When an individual item or whole test do this satisfactorily, it will sometimes be described as being *valid* (to use as intended). "Fairness" is one way that validity is articulated within assessment.

In the math story problem development and evaluation process, an item is typically considered fair if it is free from contexts that are not considered broadly accessible to students with a range of backgrounds and if it measures the same underlying skill(s) regardless of group membership (gender, race/ethnicity, language proficiency, and socioeconomic status are

frequently used) (Camilli & Shephard, 1994; ETS, 2014; Kane, 2010). Within this framework, "accessibility" is primarily evaluated through test performance. The assumption being made is that if a student can correctly answer a question, then the item worked well and the contextualization of it is unproblematic. This represents major assumptions, however, both about how inequity operates and about how success is framed within our current assessment systems, this research seeks to explore.

Through iterative interrogation of assumptions, as a research community we are tasked with rethinking what "success" really looks like for a diverse population of children, teens, and adults, as well as how the research and findings we generate contributes to and reproduces some of the same destructive mechanisms to which we claim to stand in fierce opposition. It is not enough to simply study these gaps in "achievement" and "opportunity" as mere facts and figures that describe inequity; it is essential that the direction be shifted to a more nuanced discourse about how our validation and bias-reduction methods interact with our ideals and how that in turn shapes and reshapes our collective understanding of why we collect data in the first place, as well as how they are used. Central to this discussion are the principles of validity and bias, which can carry multiple contextual meanings. In order to engage in this deeper critique, it is crucial that we first clarify the concepts and methods regularly employed by the measurement community to address these systemic issues.

2.2.1 *Conceptualizing Gender*

There are some important concepts and ideas that are necessary to define before diving into the purpose and nature of the research conducted. While some of these ideas may be familiar, others take on a particular meaning or nuance in the context of this research. These ideas provide a basis for deconstructing messages in math story problems and considering the impact they may

have on achievement and identity (Mattarella-Micke & Beilrock, 2010; Roth, 1996; Schley & Fujita, 2014).

- **Gender**- In this work, *gender* is defined as the internal and social experience of identifying with femininity and masculine along a spectrum, though gender is often presented as a binary (masculine/feminine). The experience of gender is related to, but not defined by biological sex.
- **Romantic Orientation-** Sometimes referred to as "sexual orientation," *romantic orientation* here refers to the ways a person experiences romantic attraction to people of the same or different genders. The choice to use "romantic" versus "sexual" was made to highlight romantic attraction as the key feature of partnership as opposed to sex. It is also more inclusive of people who identify as asexual and do not experience sexual desire.
- **Family Structure-** In this work, *family structure* refers to the ways a family unit might be organized. The normative model depicts a hetero-romantic-headed household ("mom" and "dad") with one or more biological children. Other family structures might include single-parent households, extended relative households (with grandparents or others living in one home), adopted children, childless families, gay/lesbian-headed households, or non-romantic cooperative living.
- **Patriarchy-** For this study, *patriarchy* was defined as a belief system that upholds the viewpoint that men and masculinity are superior to women and femininity. This can be expressed through gender role expectations, unequal distribution of power between people of different genders in society, and the images and symbols associated with masculinity compared to femininity (i.e., "strong," "rational," "brave," "protective"; "soft," "sensitive," "small," "fragile"), among other things.

2.2.2 *Validity*

At the most elemental level, validity describes the degree to which an assessment measures what it is *supposed* to measure (Borsboom, Mellenbergh, & Van Heerden, 2004). Therefore, a test itself is not described as being valid or invalid, but rather what is evaluated is whether or not the proposed interpretation and use of that test for a given purpose is well-grounded. In particular, Kane (2013) offers an argument-based approach to validation which matches clearly articulated proposed interpretations and uses (premises) with appropriate sources of evidence (rationales). Under this framework, determining the validity of intended interpretation and/or use involves evaluating how well the rationales support the premises being put forth. These are the 'arguments' to which Kane refers. While procedural premises pertaining to the design of an assessment tool, analysis of the data, and how it will be used are typically given meticulous attention, consequential premises pertaining to the potential positive and negative outcomes of the use of the assessment are often left with underdeveloped rationales. It is this latter set of premises that is of central importance to the present discussion. Simply put, the range of potential impacts and consequences deserves greater attention.

It has become utterly apparent that the greatest threats to appropriate test use and interpretation do not reside within the domain of statistical assumptions, but rather with those of a theoretical nature. Kane (2013) explains that when we validate an interpretation or intended use of a test score, we are interrogating the *plausibility* of that claim in light of the evidence we have. However, throughout their writing, Kane also points to the fact that interpretation is the least developed facet when it comes to validation work (a minor critique of Messick (1987), whose earlier work understated this component of validation work]. Despite arguably being the most important component of how we understand what makes an assessment tool "valid" in a

conventional sense, a deep exploration of the consequences of test interpretation is done from outside of our discipline (Au, 2011; Green & Griffore, Lomax et al., 1995; Madaus & Neill, 1990).

Anastasi provides a strong foundation for understanding and evaluating validity (especially construct validity). In context, Anastasi's (1986) views seem progressive and almost "radical" (i.e., when she describes the traditional tripartite approach to validation as "crude and oversimplified"). Anastasi avows: "Test validity is a living thing; it is not dead and embalmed when the test is released" (p. 4). This is a critical ideology, because it reiterates the importance of adaptability as a part of the scientific method and stands as a reminder that constructs are not static. According to Anastasi, "empiricism need not be blind" (p. 6), a perspective which centers context in our understanding of validity. Anastasi (1986) proposes that valid measures can nonetheless be used to define and evaluate broad traits by assessing individuals across scenarios and aggregating outcomes. However, Anastasi leaves a conceptual gap: what is the responsibility is of the researcher(s) in terms of highlighting the importance of those particular contexts that produce/contribute to difference?

Haertel (2013) implores measurement professionals to expand their scope when it comes to giving appropriate attention to "unintended testing consequences." It is, as they put it, in large part due to "an entirely understandable confirmationist bias" that results in an incomplete evaluation of all plausible outcomes (p. 87). They go so far as to suggest that perhaps the measurement community may not the best-suited for this task at all, and suggests that consequential validity[3] should, as a necessity, be a more interdisciplinary domain. Haertel asserts

[3] The potential and actual positive or negative individual or social consequences of the use of a test or its scores/data.

that measurement professionals are certainly capable of bringing an appropriately wide lens to the validity inquiry given the opportunity to engage in scholarly collaboration with interdisciplinary colleagues who can draw from a more diverse body of knowledge and experience. When it comes to evaluating the construction of social identity within and through test content, the opportunity to engage in scholarly collaboration with interdisciplinary colleagues who can draw from a more diverse body of knowledge and experience is critical for bringing an appropriately wide lens to the validity inquiry.

While the study of group performance differences has more than a century-long history, the high-stakes testing movement catalyzed more than three decades of educational research largely focused on understanding and predicting assessment scores (Green & Griffore, 1980; Howard, 2010; Lomax et al., 1995; Madaus & Clark, 2001; Medina & Neill, 1990). A broad literature now exists exploring the ways that identity manifests in academic performance, but though this has yielded a surplus of scholarship exploring, exposing, and explaining, among other things, gendered performance deficits in math and other subjects, the studies are critically lacking when it comes to naming and challenging the oppressive ideological systems that produce unilateral inequity in education.

2.2.3 *Bias vs. Fairness*

Before moving into a discussion of methods of identifying bias, it is important to first articulate what is actually meant by *bias.* Statisticians and society both use the term bias, but often in very distinct ways. In statistics, we are typically addressing ‘measurement bias,’ a systematic error in how an instrument measures a trait in specific group (Camilli & Shepherd, 1994). The key here is that the error is systematic; it is not random or by chance: the test *consistently* does not produce valid results for one or more groups. For example, a math test may advantage people

who learned English as their first language such that English Language Learners (ELL) routinely score below native speakers of the *same underlying ability.*

Often, however, even within measurement communities, when we use 'bias' more colloquially, we are describing *fairness,* which is more of a qualitative assessment of the social impact of test administration, use, or content. To understand the subtle, but important difference between bias and fairness, take the same hypothetical test I just described, but now consider a case where no systematic score differences can be found, but where some questions contain content that is unfamiliar or offensive to a certain subgroup, or where the test results are interpreted or used in different ways for different groups. Historically, both systematic, statistical bias and fairness-type bias have been found in a wide variety of instruments on the basis of gender, race, social class, and disability, among other identity markers. Both indicate a fundamental flaw in the design, administration, and use of some measurement tool, but the ways in which different sorts of bias/fairness issues are addressed varies widely across educational and professional contexts.

What happens when our assessment systems don't equitably capture the full range of student learning and experience? This research explores the intersection of measurement bias, how our systems differentially assess students from different groups, social bias, the beliefs and ideas about those groups that we bring to our work, and fairness, where measurement and social bias meet. So, why do bias and fairness matter in assessment? In short, the identification of statistical and/or social bias in tests represents a validity problem (Kane, 2010; Smith & Fey, 2000). If whole tests, or even individual items, advantage one group over another, then our ability to make claims about test scores or group trends is compromised. Bias at the test or item level can originate from a number of sources. While group-level score discrepancies may be a

true or partially true representation of the bias and inequity in society, it is important to remember that tests have also historically been used in ways that compound their effects (Medina & Neill, 1988). In the absence of measurement bias, the presence of social bias nonetheless represents a failure to "sanitize" test questions of existing norms, values, and stereotypes about individuals and identities.

Many researchers and test developers have articulated different ways the content of whole tests or individual items can be biased, unfair, or insensitive (Abedi & Levine, 2013; Camilli & Shephard, 1994; ETS, 2014; SBAC, 2012, Warne et al., 2014). For example, a test might be considered biased if it contains more references to one gender than another or if gender roles are presented stereotypically (American Association of University Women, 1992). Given the negative effect stereotype activation can have on the test performance of women and girls, these patterns can have real impacts and are too harmful to be ignored (Ambady et al., 2001; Good, Woodzicka, & Wingfield; 2010; Picho & Schmader, 2017). Though AAUW (1992) reports mixed evidence that these features impair test performance, other research has demonstrated that these things make a difference (Au, 2011; Green & Griffore, 1980; Lomax et al., 1995; Madaus & Neill, 1990). Regardless, the presence of insensitive, biased, or insensitive content undermines the claim that tests are objective and value neutral.

2.3 How We Assess for Bias in Tests

2.3.1 *Bias and Sensitivity Review*

Most large-scale assessments already have procedures in place for initial bias and sensitivity review. In 1988, the National Assessment Governing Board (NAGB) was created by the U.S. Congress to oversee the formation of policy guidelines for the National Assessment of Educational Progress (NAEP), the only nationally representative, long-term evaluation of

American student achievement and ability (NCES, 2000). These guidelines included standards for "determining the appropriateness of test items and ensuring they are free from bias" as it relates to the potential negative psychological impact of cognitive bias, as well as the more immediate practical consequences of statistical bias. The development process for NAEP items is consistent with other broadly administered measures such as that of the Common Core State Standards (CCSS), Smarter Balanced Assessment Consortium (SBAC), and even state-level curricular standards such as the Washington State Next Generation Science Standards (NGSS).

However, despite extensive process documentation which often includes exhaustive lists of what themes and topics should and shouldn't be present in assessment items, explicit language on content review procedures remains generalized and short-winded (LACCD, 2012; NCES, 2017; WA OSPI, 2011; SBAC, 2012). Test documentation for NAEP and other widely used scales typically include only a few sentences regarding a concrete qualitative review of items for potential sources of bias and/or contexts which might cause students emotional distress on the basis of identity or culture (NCES, 1999, 2014, 2017; SBAC, 2012; CCSS, 2013; NGSS, 2013; PARCC, 2016). While it is likely that this process does, in fact, occur, the lack of transparency and specificity when it comes to this process, especially when compared to the extensive statistical reporting for item/examinee sampling procedures, coupled with a general lack of diversity within education and the educational measurement community (Dee, 2005; Gorski, 2012; Howard, 2010; Taie & Goldring, 2017), introduces some amount of skepticism regarding not only the rigor, but also the purpose and efficacy, of these early-stage procedures for identifying bias.

2.3.2 *Differential Item Function Analysis (DIF)*

The results of a DIF analysis indicates if students in different groups require different latent ability (θ) in the subject area in order to have the same probability of answering a given item correctly, which may signify that it is written in a biased way. There are a number of methods for computing DIF, including, Item Response Theory (IRT-DIF), Mantel-Haenzel (MH-DIF), and regression (R-DIF). The most common way the DIF procedure is applied is via 2-Parameter IRT models which take into account the *difficulty* of an item (*b*)—proportion of students in a sample who got the item correct—and the *discrimination* of that item (α)—a measure of how well an item distinguishes between high and low performers.

The existing body of literature on Gender DIF in math is extensive. In general, this research compares boys' and girls' relative performance on some type of assessment (Kan & Bulut, 2014; Lan & Li, 2014; Li, Cohen, & Ibarra, 2004; Mendes-Barnett & Ercikan, 2006; Ryan & Fan, 2996; Taylor & Lee, 2012). This work was borne of an effort to close the gendered achievement gap that persists in mathematics, specifically, and other STEM disciplines more generally. Research by Kan and Bulut (2014) revealed that some contextualized math problems favor girls, while non-contextualized problems did not favor boys or girls. Other research by Ibarra (2001) attributes the occurrence of DIF to test-takers' culturally-induced cognitive dissonance between academic and social expectations, described by the Multi-context model. Li, Cohen, and Ibarra (2004) later integrated this socio-cultural model and Gallagher and colleagues' process-focused model (Gallagher, 1998; Gallagher et al., 1999) into a coding scheme which identified factors predicting gender DIF. These, and other gender-related studies, help us understand when and why gender DIF occurs (Mendes-Barnett & Ercikan, 2006; Ryan & Fan, 1996; Taylor & Lee, 2012). It is additionally important to acknowledge that the existence of DIF

for items within a scale only demonstrates statistical bias, but does not necessarily indicate *contextual* bias, be it statistical or socio-cultural.

2.3.3 *Problematizing DIF*

While IRT methods for detecting DIF are commonplace and generally accepted as a robust method for identifying bias and/or differentially functioning items, it nonetheless has some drawbacks. The major critique of IRT DIF is that it relies on total score for matching the comparison groups. If items are problematic (DIF or otherwise), the comparison becomes circular and loses some validity (Hidalgo-Montesinos & Gómez-Benito, 2003). For example, if a test contains items that privilege native English speakers, the ability parameter of the IRT DIF model that will be estimated based on the total score of the test will be slightly inaccurate for students in both that group and the group of English learners, whose scores will be higher and lower (respectively) than they would be if those items were omitted or didn't privilege a particular group. Additionally, IRT DIF fails to account for situations where the two groups have different ability distributions. Essentially, IRT assumes that the scores from the two groups are fundamentally comparable, when, in fact, they may not be (Zumbo, 2007).

The DIF approach, overall, has several additional key limitations that call into question its use as 'the' method of identifying "unfair" items. For one, DIF is often limited to two-group comparisons. Though procedures for multiple group comparisons exist and are being refined, much of the literature on DIF and gender is limited to binary comparisons. Further, using DIF as the standard for item, and by default, test, fairness preserves assumptions about standardized tests and reproduces existing test culture. It keeps the focus on final score as a measure of achievement and appropriateness for inclusion in a test booklet, further cementing final test score as "the" measure of achievement and learning. This forecloses on a more holistic "whole

student" approach to teaching, learning, and assessment, undermining many skills and challenges which cannot be adequately captured by an academic test score (such as empathy, perseverance, burnout, or boredom), but are no less fundamental to students' identity development , as well as their academic and cultural self-concept (Mayes, 2007).

What this approach fails to account for are the social mechanisms that produce the potential for bias in the first place and necessitates a troublingly narrow definition of how a given student may or may not be harmed by exposure to certain types of items. Simply because two students are equally likely to get a particular item correct does not exempt that item from being embedded with a range of socio-cultural ideals, stereotypes, and norms. For example, an item might portray a stereotypical scene such as "Mr. Jones" watching a basketball game for a certain amount of time, while "Mrs. Jones" cooks dinner, for a different amount of time, and may ask students to solve a math problem relating to the time. The context does reproduce messages about traditional gender roles, but this alone does not predict student performance. Students of different genders may *not* be impacted when it comes to performance on the item, but it does not change the fact that the context was flawed from the start, nor that the representation may impact students psychologically. Furthermore, cultural consciousness is less shaped by individual major events, than it is by the mundane repetition and reiteration of seemingly innocuous ideas, values, and perspectives. More concerning than the overt biases in a test question (that will likely be deemed unsuitable for inclusion) are the more insidious and subliminal patterns that serve to reinforce existing social constructs (Arendasy & Sommer, 2012; Hernandez et al., 2014; Keith, 2016, Sterzing et al., 2017; Sue, 2010, 2010).The sum of these occurrences are nothing short of measurement microaggressions[4].

[4] See definition in Chapter 1, section 1.2.3.

2.4 The Purpose of This Work

The broad aim of this research was to disrupt some of our preconceived notions about testing and 'fairness.' In particular it raises several important and interrelated questions: *Who does academic testing serve? (To what end? What is the impact?) In what ways are our protocols flawed and/or harmful? How can we do a better job of acknowledging these flaws and limitations while still maintaining systems of accountability?*

More specifically, this research aims to challenge traditional understandings of what the measurement community calls *consequential validity*. Messick (1988) argued ". . . it is not that adverse social consequences of test use render the use invalid but, rather, that adverse social consequences should not be attributable to any source of test invalidity such as construct–irrelevant variance." In other words, Messick is defining adverse consequences in a specific way that limits the culpability of the test to only those situations where negative outcomes can be directly tied to problems with the test itself (ex., if a certification test systematically underscored a particular group ("test invalidity") and then that group was systematically barred from gaining employment ("adverse social consequences"). This perspective introduces a particular set of rather narrow assumptions about the social consequences of tests and testing that require consideration. A major catalyst for conducting this research was recognizing a need to push back against those assumptions. How do we define "adverse social consequences"? Risk is almost entirely context-dependent, rendering it nearly impossible to encapsulate for entire student populations. Furthermore, how often can we really claim that we have thoroughly examined all potential sources of test invalidity? It remains uncertain whether or not this is possible, either.

What would it mean to rethink how we define and assess bias and risk? How would that impact our strategies for addressing it? The work outlined here attempted to answer three

specific questions. First: Do test items which have already been screened for bias using both qualitative and statistical methods still contain contextual patterns and themes which reproduce gendered stereotypes? If they do: Which stereotypes are being reproduced? To what extent do a diverse group of reviewers agree about the presence of stereotyping and types of bias identified?

2.5 METHODOLOGY

In order to test the hypothesis that young people are exposed to test items that, while allegedly absolved of explicit bias, may still reproduce gendered social stereotypes, a sample of items from a nationally administered standardized exam was selected and examined for the existence of consistent patterns of biased themes.

2.5.1 *Source of Items*

The final question sample was drawn from the pool of released items from the National Assessment of Educational Progress (NAEP) spanning 1990 to 2013. Math story problems were chosen as a domain due to the fact that a broad body of research exists detailing the ways that linguistic characteristics mediate math performance for marginalized student populations (Green & Griffore, 1980; Hendrix, et al., 2009; Levine & Levine, 2013; Lomax et al., 1995). Math problems offer an opportunity to explore both content knowledge and the effect of non-content related stimuli simultaneously.

The pool was further restricted to only Grade 8 items. NAEP administers tests to 4th and 12th graders, as well. Grade 8 was chosen because this period of adolescence is important to the way individuals begin to solidify a sense of self, as well as being an important academic transition year, especially for marginalized student populations, that renders it of particular relevance for this type of exploration (Cognato, 1999, as cited in Mizelle & Irvin, 2000; Eccles, Midgley, & Adler, 1984; Gillock & Reyes, 1996). Out of 440 released Grade 8 math items, 120

were identified as “contextualized.” These problems, in contrast to “naked number problems,” are narrative in structure, not always, but most often, describing characters and actions by those characters to situate the math content within a real-world scenario. All 120 items were evaluated using a novel evaluation rubric, which is described in the next section.

Table 2.1

Item Characteristics for the NAEP Grade 8 Math Items

Item Characteristics					
Year	*n*	*%*	Content Area	*n*	*%*
1990	3	2.5	Algebra	26	21.7
1992	22	18.3	Data Analysis, Statistics, & Probability	29	24.2
1996	8	6.7	Geometry	10	8.3
2003	16	13.3	Measurement	12	10.0
2005	10	8.3	Number Properties & Operations	43	35.8
2007	18	15.0			
2009	8	6.7			
2011	14	11.7	Average Pass Rate	0.58	
2013	21	17.5			
Graphic	*n*	*%*	Type	*n*	*%*
No	73	60.8	Multiple Choice	74	61.7
Yes	47	39.2	Short Constructed Response	36	30.0
			Extended Constructed Response	10	8.3

Note. Total *n* = 120.

2.5.2 *Instrument*

The novel coding rubric was composed of two parts: Intrinsic Bias (IB) and Item Diagramming (ID) (See Appendix 2 A). The IB section consisted of 6 Yes-or-No questions regarding the presence of problematic elements (i.e., “Does the context of the item favor one family structure over others?” and “Does the context of the item assume that gender identity predicts interest or behavior?”). If a coder chose ‘yes’ for an IB question, they wrote a description of the presence of that element in a particular item. The language used in the protocol was consistent with that used in the bias and sensitivity guidelines provided in standardized test documentation (NCES, 1999,

2014; SBAC, 2012; CCSS, 2013; WA OSPI, 2011; PARCC, 2016), but taken a step further to allow for the deeper and more radical item-by-item interrogation requisite of queered methodology. For example, when presented with an item about re-carpeting a room, reviewers may indicate that the item did assume that gender identity predicts interest or behavior and then provide a concise, but detailed rationale (example Appendix 2 B).

The ID section consisted of dissecting each item into component relational parts: individual characters or 'agents' described in the item, as well as each person's implied or explicitly labeled gender (via name and pronoun use), and a rich description of what they are doing in the scenario. The principal goal of this procedure was to link gendered identities with gendered tasks in order to map patterns of stereotyped behaviors based on implicit gender bias within the contextualized items. An embedded goal of this process was to identify overarching task types, such as "home repair," "sports," "science," or "caretaking." This allowed for concise pairing of gendering with behavior.

2.5.3 *Procedure*

Panel Review. In order to triangulate the findings, a panel review was organized. Three reviewers, including the primary investigator, all well-versed in critical theory, specifically gender bias, math learning, or both, were recruited to participate in 3 6-hour collaborative review sessions. The group went through the first item together to make sure all were comfortable with the process. Each reviewer then independently reviewed all items using the rubric, so that every item had been evaluated 3 times. Reviewers made note of questions or comments that arose during their review process. These were discussed halfway through the session and again at the end. This procedure ensured that the item review remained an independently-guided task, but

still left room for the process to be a creative joint effort as opposed to one that is static and unamenable.

Collaborative discussion focused on highlighting particularly problematic or confusing contexts. It also gave each reviewer an opportunity to reflect on the ways their own understanding informed their review process. One reviewer was a white man born in the U.S. who offered his perspectives on the representation of masculinity and masculine identity from his own position as both a male-identified person and as an educator and researcher on equity in learning. Another reviewer was a Chinese woman in the U.S. to study internationally and she raised many interesting questions about the connection between names and gender in a U.S. contexts. The third reviewer identified as a queer mixed ethnic person and used their unique perspective on gender roles and liminal identity to inform their review. While there was rarely overt disagreement about concepts raised during discussion, each of the reviewers' identities played a clear part in how they were understanding and engaging the items for review.

Data Analysis. Scale reliability was calculated for the overall instrument using Cronbach's alpha for internal-consistency reliability. Inter-rater reliability analysis was conducted for each of the 3 review session time points, as well as for all time points together. Inter-rater reliability looks at the degree to which raters on agreed on a common task. In this research, inter-rater reliability, evaluated by the intra-class correlation, refers to the agreement of the 3 item reviewers when it came to how they answered the first rubric question: *Does the context of the item reinforce gender stereotypes?* A multiple logistic regression analysis was also used to identify which, if any, item characteristics such as content area, question type, pass rate, or year of administration significantly predicted being rated as reinforcing gender stereotypes.

Content Analysis. Two evaluative procedures were utilized to analyze the panel review data. First, the IB results were cross-validated for common 'yes' responses and descriptions of these occurrences were flagged for themes using an emergent thematic coding approach to discourse analysis (Schreier, 2014; Willig, 2014). Next, the ID results were evaluated for consistency across reviewers. The goal of coding was to identify systematic patterns in the combinations of different story problem character themes, roles, and behaviors. This process facilitated linking gendered identities with stereotypically gendered behaviors.

We used a relatively simple coding procedure to look for themes within and across items. First, for each question, we noted the number of "characters" featured in the item. Next, we identified if gender was made explicit through the use of gender pronouns (she/her/hers, he/him/his) or implied by name (or perhaps by activity). We then briefly summarized the action(s) of the character (Ex. Kara (girl) is washing cars for a fundraiser), which allowed us to sort items into activity categories (Ex. food preparation, sports, science). At this stage it was possible to start noting patterns in association between gender and activity. In order to look more closely at *how* characters of different genders were, we also began to note the sorts of words used to describe the actions of the characters such as those relating to agency and motivation (the "why" of what they were doing) and those having to do with the roles of the characters (competition, relevant/irrelevant details). We next considered whether or not the context felt believable. That is did it seem like something a "real person" would do (from our individual perspective) or did it feel artificial? Finally, we thought through what messages about school, gender roles, behavior, or other aspects of society or identity were being sent by the item.

2.6 RESULTS

2.6.1 *Scale and Item Characteristic Analysis*

Scale Reliability. Scale reliability was assessed using Cronbach's alpha. Item Bias questions were treated as a single dimension. The analysis was run for the overall response sample across all time points and raters, split by time point, split by rater, and split by time point and rater. Across all time points and raters, the scale reliability was generally weak (α_T= .521), with much room for scale improvement. However, interesting patterns emerged through more nuanced analysis. Despite being a collaborative task, over time, the reliability of the scale actually decreased (α_1= .628; α_2= .593; α_3= .409), suggesting that rather than moving item reviewers towards unanimity, collaborative review actually introduces greater diversity of perspective. Further, differences between raters reveal disparities in scale salience for people of divergent backgrounds (α_a= .485; α_b= .589; α_c= .516) that raise a question about the true uniformity of review practices in the testing industry. These patterns are only strengthened when rater and time are intersected (α_{a1}= .618, α_{a2}= .521, α_{a3}= .376; α_{b1}= .639, α_{b2}= .658, α_{b3}= .491; α_{c1}= .644, α_{c2}= .585, α_{c3}= .493), implicating that rather than converging in understanding, over time, the task became more complex and nuanced.

Inter-rater Reliability. Given, low scale reliability, much higher rates of affirmative response on the first two scale questions (>55%) compared to the following four (<5%), and a strong correlation between them (r_b= .748, p<.001), the discussion of inter-rater reliability will focus on agreement on the first scale question only: *Does the context of the item reinforce gender stereotypes?* Overall, the intra-class correlation for average measures was an underwhelming 0.60. When examined at each time point, however, a similar pattern emerges in the analysis of inter-rater reliability as did in the scale reliability analysis. Across time, it is clear that rater

agreement does not follow a positive, linear trajectory (ICC_1= .876; ICC_2= .611; ICC_3= .411), revealing that as collaborative item review became deeper and more critical, item flagging became less consistent or predictable across reviewers.

Logistic Regression. A binary logistic regression analysis was used to examine the relationship between item characteristics and the likelihood that an item context would be flagged as containing a gender. Across raters, in general, item characteristics alone fail to predict item flagging (see Table 2.2 in Appendix 2 C), which provides good evidence that items are not systematically more likely to be problematic. Within raters, we see subtle differences in model fit, but overall, similar findings (Table 2.3 in Appendix 2 C). Time and pass rate were only significant at the dependent aggregate level, not at the reviewer model level. Reviewer 2 had a significant effect of the inclusion of a graphic, while Reviewer 3 responded inconsistently at time point 2. Reviewer 1's model showed no significant factors predicting rating an item as containing stereotypical gender representations. Overall, these results suggest that contextual bias is pervasive and non-discriminating. That is, subtle item characteristics generally do not influence or predict problematic item content; all items have similar potential.

2.6.2 *Content Analysis: Implicit Gendered Associations*

Analysis of the data revealed several interesting trends in how implicit gender bias is manifested in the sample of math story problems surveyed. While Yes/No ratings were not perfectly unanimous between panel raters, item-by-item, trends and themes were very stable across the item set. For the first question, *Does the context of the item reinforce gender stereotypes?,* 55% of items were flagged by reviewers 1 and 3, and 70% by reviewer 2. For the second question, *Does the context of the item assume that gender identity predicts interest or behavior?,* 42% of the items were flagged by reviewer 1, 69% by reviewer 2, and 57% by reviewer 3. Diagramming

was also not perfectly consistent, but revealed that there are distinct and problematic combinations of story problem contexts when it comes to the reproduction of gender stereotypes and patriarchal messages. For example, each time a character was presented as a teacher, the teacher was a man.

Warranting further exploration of the data were nascent patterns when it comes to the actions and interactions of characters depending on their implicit/explicit gender. Overall, feminine gendered people were more likely to be included in narratives involving food, exercise (as distinct from sports), and caretaking. Masculine gendered people were more likely to be included in narratives involving scientific exploration, athleticism, or the critique of another person's performance. These trends were consistent across the item sample.

Beyond simply what was there, consistent with Apple's (2015) "politics of absence", the review procedure was perhaps most interesting in terms of what wasn't present in the items. Sampled items presented no alternatives to the gender binary when it came to gender expression, such as the use of the singular *they,* and the inclusion of names which do not have familiar gendered connotations were sparse. Similarly, family structures were assumed to subscribe to a hetero-normative two-parent structure. Even when this was not explicit, references to other family structures and relational dynamics, including, but not limited to single-parent, grandparent headed, extended-family, and adoptive households were entirely absent.

Gender Stereotypes. Despite extensive bias and sensitivity review prior to release, there were still systematic associations between gender and behavior within the item sample. Only masculine people were depicted doing home repair or making significant purchasing decisions for a household such as laying carpet or buying a refrigerator, in authority and judgment roles like teacher, as physically disciplined, as knowledgeable and well-informed, and as very

autonomous. In contrast, items depicted feminine people performing service tasks and emotional labor such as packing snacks and candy for other people, as lacking in self direction, as relational as opposed to logical or rational, performing repetitive math tasks with no clear motivation such as spinning a spinner or completing a pattern, and yet are also depicted as less mathematically competent. We are repeatedly called on to critique the work of a feminine person, but not once that of a masculine person, as well as compare the work or behavior of a feminine person to that of a masculine person. As an example, one question explains that Sara "was asked" to draw a certain shape and the test taker is asked if the shape Sara drew was correct.

Patriarchal Messages. These associations laid a foundation for certain specific patriarchal messages to be conveyed by these items. One prominent message was that masculine people are the "head of the household" and are responsible for home repair and decision-making. Masculine people are never depicted in caregiving roles. One item portrays a teacher, Mr. Bell, but the context of the question has him acting as a decision-maker, not as a nurturer. Several items subtly imply that feminine people should be concerned with their weight and should be in shape in a manner distinct from the ways that masculine people are shown to be athletic and possessing physical prowess. A question about a masculine person riding his bike a certain number of miles per day over the course of a week is contrasted with another that depicts a feminine person riding a bike at a very slow pace (associated with fat loss) at only one time point.

Instrumental Role. The main way these messages were inscribed into the item texts were through the designation of masculine and feminine people as differentially instrumental. The complementary themes of agency and passivity are prominent. Feminine characters are consistently shown engaging in behaviors due to need and provocation or request ("needed to"; "has to"; "was asked to") compared to masculine characters, whose actions rarely include these

linguistic qualifiers. Feminine people need to be directed; masculine people just do things. Inexplicably, questions detailing someone's rate of pay or work schedule pertained to feminine characters. The few items that did not imply a specific gender nonetheless contained gendered cues that impacted the interpretation of the item (ex. "a plumber" produces a masculine prime despite no name or pronouns used). What this suggests is that even in the absence of explicit gender priming, ingrained ideas about gendered social roles are triggered in the mind of an item reviewer. Rather than assuming non-gendered items would be taken up as neutral by student test-takers, we discussed the various ways an individual student's identity might inform the assumptions *they* make about the person described in the item.

2.6.3 *The Collaborative Process*

One interesting and surprising result was the decreasing reliability of the review activity over time. It was expected that as the reviewers gained more experience with the task and as the discussion on item features became more nuanced, that we would attain greater consistency across items. However, the opposite was found. Over time, the intra-class correlation coefficients decreased from near-acceptable to not acceptable (by conventional standards). This raises interesting questions about the assumptions made about how item review is done. Rather than seeing expert review as a static judgment, it may be more reasonable to see item ratings as subjective, momentary impressions informed by a range of factors, from personal identity to the informal life experiences they had on the day they completed the task to the conversations had with colleagues or others.

2.7 A Call to Action

This research stream was initiated in an attempt to rethink the way bias in testing is understood by test takers, as well as test administrators and other relevant stakeholders. In order to ask these questions, it was necessary to problematize the current review processes and statistical procedures for identifying biased test questions. This led to a more fine-grained evaluation of individual test items in order to uncover patterns in narrative structure and framing that revealed underlying themes of cultural reproduction.

This early work radically challenges some fundamental assumptions made about how the measurement community evaluates test questions for bias and sensitivity. Further, it extends the conversation about testing, evaluation, and measurement within the larger educational community, especially when it comes to the role of scores in the promotion and placement of students, retention of teachers, and school district funding. This work represents a novel exploration of arguably the most important micro-level unit of analysis in measurement—the test item—as a means to expose macro-level social discord.

The findings reveal just how deeply ingrained these normalizing projects are. Beliefs about gendered behavior are persistent even in face of systematic review. Rather than placing the focus on the quantification of these patterns and inconsistencies, the goal is, on the one hand, to inspire the measurement community to rethink its position and role within the contemporary discourse on test culture and its relationship with identity development during childhood and adolescence, and on the other, to inspire a collective educational movement that promotes student and parent (as well as teacher) agency when it comes to participating in and understanding the impact of that test culture on their own lives and the lives of others.

As equity- and justice-minded education professionals, we are called upon in this Data Age to engage in difficult conversations about what the purpose of educational measurement is: Who does it serve? In what ways are our educational protocols flawed and/or harmful? How can we do better while still maintaining systems of accountability? It is not enough to simply study gaps in "achievement" and "opportunity" as mere facts and figures that describe inequity; it is essential that the direction be shifted to a more nuanced discourse about how our validation and bias-reduction methods interact with our ideals and how that in turn shapes and reshapes our collective understanding of why we collect data in the first place, as well as how they are used.

The goal of this research was to gain a more nuanced understanding of the gendered messages being sent to school-age young people, not only explicitly through direct socialization, but *implicitly* through the very texts they encounter on a daily basis. Nationally administered test items were selected because they, unlike typical teacher-designed tests and other informal assessments, have been subject to immense scrutiny already. To demonstrate that they are not impervious to the taint of deeply ingrained gender stereotypes is strong evidence that these linguistic patterns are pervasive in educational materials, from test questions to textbooks.

Math understanding, ability, and performance hold significant authority in society (Moses & Cobb, 2001). Given the fact that teachers and parents play an important role in shaping students' math beliefs and self-concepts (Amit & Fried, 2005; Tomasetto, Alparone, & Cadinu, 2011), the findings represent a valuable opportunity to leverage collaborartive learning approaches to transform students' experiences learning math. Taking the time to talk through the many ways different people can interpret a test question is one way of engaging these complex ideas with learners of all ages. While much needs to be done on the test development side, this is

one site for immediate intervention for test administrators and examinees. The goal is to get us all thinking more critically and reflectively about how identity is represented in tests.

In a recent study, Casad et al. (2017) argued that researchers should be focusing on the lived experiences of young girls to better understand how their early educational experiences impact their gendered beliefs and math trajectories. They're among a growing contingent of researchers calling for a shift in how stereotype threat and achievement/opportunity and recruitment/retention gaps are studied in girls, as well as in other marginalized student groups. Rather unfortunately, much of the research on gender bias in math testing fails to capture the essence of what it means to be a gendered person in the world (i.e. how gender identity is experienced in less clinical settings, how it interacts with other traits and social identities, and the all-encompassing nature of systemic oppression).

The findings of this study greatly call into question our current approaches to "sanitizing" and neutralizing test content for the purposes of accessibility. While they may become more homogenized in a sense, rather than representing everybody, the questions often come across as representing nobody. While it remains unclear how the measurement community will reconcile inevitable violations to the assumptions and limitations of traditional measurement models, starting a conversation about how math story problem contexts can more authentically reflect the true diversity of examinees and how that impacts our interpretation of student achievement seems like a critical next step.

2.8 LIMITATIONS

There were several limitations to the design and implementation of this study. In terms of design, the panel review session revealed some redundancy and lack of clarity within the item coding rubric. As an implied goal of this exercise was to do an informal evaluation of the instrument,

this is not inherently problematic, nor does it seem to bear on the broad findings. However, it also cannot be said that it did not have an impact on the process. For example, in the Intrinsic Bias section, the gender, family structure, and sexual orientation item pairs (ex. “Does the context of the item reinforce gender stereotypes?” and “Does the context of the item assume that gender identity predicts interest or behavior?”; “Does the context of the item favor one family structure over others?” and “Does the context of the item assume a common family structure for all people?”) were too similar. There were few instances in which a distinct description could be given for each without significant conceptual and referential overlap. Future iterations of the instrument will address these design issues.

2.9 Future Directions

While this initial exploration focused on bias and norms based on gender and sexual expression, the hope is to extend this procedure to the exploration of racial/ethnic bias, socio-economic bias, ability, and, ultimately a fully intersectional analysis of the subtle ways these elements interact to cultivate the conditions that uplift some while marginalizing others. Though this approach is time-consuming and therefore costly and highly impractical for large-scale item review, there remains a lot of potential for interdisciplinary collaboration towards automating some of the process. Natural Language Processing (NLP) (an extension of Machine Learning) techniques may offer some solutions. NLP is an algorithmic method for streamlining discourse analysis. It allows for much faster item evaluation for non-narrative review components and instrument optimization in a much shorter timeframe which could allow some of these principles to be implemented and evaluated more comprehensively.

In addition to exploring novel methods for item review, it would also be useful to extend our coding framework to be more accessible for use by reviewers not as deeply knowledgeable

about gender equity and test theory. Using the findings from the initial review and follow-up activity-by-gender analysis, it would be possible to create a new review rubric that first centers the item deconstruction (person, names, genders, activities) to reveal patterns amongst the relevant item features and then expands to a more nuanced consideration of how item contexts at the individual item- and whole test-level reproduce stereotypes and inequitable portrayals of people with differing identities. This process could significantly reduce the amount of previous theoretical knowledge required to engage in item review without having to sacrifice as much methodological efficiency or scope.

2.10 REFERENCES

Abedi, J., & Levine, H. G. (2013). Fairness in assessment of English learners. *Leadership, 42*(3), 26-38.

Ambady, N., Shih, M., Kim, A., & Pittinsky, T. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science, 12*(5), 385-390.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*(11), 1-16.

Apple, H. (2015). The $10,000 woman: Trans artifacts in the Pittsburgh queer history project archive. *TSQ: Transgender Studies Quarterly, 2*(4), 553-564.

Arendasy, M. E., & Sommer, M. (2012). Using automatic item generation to meet the increasing item demands of high-stakes educational and occupational assessment. *Learning and Individual Differences, 22*(1), 112-117.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258-267.

Au, W. (2011). Teaching under the new Taylorism: High-stakes testing and the standardization of the 21st century curriculum. *Journal of Curriculum Studies, 43*(1), 25-45.

Bhabha, H. K. (2004). *The location of culture.* Routledge.

Bolte, E., Brown, D., Segrist, D., & Dudley, M. (2015). Investigating the moderating role of gender collective self-esteem on the relationship between the experience of gender microaggressions and psychological distress in women, ProQuest Dissertations and Theses.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071.

Boyd, N., & Ramírez, H. N. R. (2012). *Bodies of evidence: The practice of queer oral history* (Oxford oral history series). New York, NY: Oxford University Press.

Bronfenbrenner, U. (1994). Ecological models of human development. In T. Husten & T. N. Postlethwaite (Eds), *International Encyclopedia of Education* (2nd ed., *Vol. 3*, pp. 1643–1647). New York: Elsevier Science.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Measurement methods for the social sciences series; 4). Thousand Oaks [Calif.]: Sage Publications.

Cooley, E., Payne, B., & Phillips, K. (2014). Implicit bias and the illusion of conscious ill will. *Social Psychological and Personality Science, 5*(4), 500-507.

Dee, T. (2005). A teacher like me: Does race, ethnicity, or gender matter? *The American Economic Review, 95*(2), 158-165.

Derrida, J. (1976). *Of grammatology* (1st American ed.). Baltimore: Johns Hopkins University Press.

Driskill, Q. *Asegi stories: Cherokee queer and two-spirit memory*. Tucson: University of Arizona Press, 2016.

Eccles, I., Midgley, C., & Adler, T.F. (1984). Grade-related changes in the school environment: Effects on achievement motivation. In G. Nicholls (Ed.), *Advances in Motivation and Achievement*, 3, 283-331. Greenwich, CT: JAI Press.

Educational Testing Service. (2014). ETS standards for quality and fairness. Princeton, N.J.: Educational Testing Service.

Galdi, S., Cadinu, M., & Tomasetto, C. (2014). The roots of stereotype threat: When automatic associations disrupt girls' math performance. *Child Development, 85*(1), 250-263.

Gallagher, A. M. (1998). Gender and antecedents of performance in mathematics testing. *Teachers College Record*, 100, 297-314.

Gallagher, A. M., Morley, M. E., & Levin, J. (1999). Cognitive patterns of gender differences on mathematics admissions tests. In *New Directions in Assessment for Higher Education* (GRE FAME Rep. 3; pp. 4-11). Princeton, NJ: Educational Testing Service.

Gillock, K. L., & Reyes, O. (1996). High school transition-related changes in urban minority students' academic performance and perceptions of self and school environment. *Journal of Community Psychology, 24*(3), 245-261.

Good, J., Woodzicka, J., & Wingfield, L. (2010). The effects of gender stereotypic and counter-stereotypic textbook images on science performance. *The Journal of Social Psychology, 150*(2), 132-47.

Gorski, Paul C. (2008). Good intentions are not enough: A decolonizing intercultural education. *Intercultural Education, 19*(6), 515-525.

Gorski, Paul C. (2012). Perceiving the problem of poverty and schooling: Deconstructing the class stereotypes that misshape education practice and policy. *Equity & Excellence in Education*, 45(2), 302-319.

Green, R., & Griffore, R. (1980). The impact of standardized testing on minority students. *The Journal of Negro Education, 49*(3), 238-252.

Greenwald, A. G.; Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*. 102, 4–27.

Greenwald, A. G.; McGhee, D. E.; Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*. 74, 1464–1480.

Haertel, E. (2013). Getting the help we need. *Journal of Educational Measurement*, *50*(1), 84-90.

Hendrix, K., Trotter, S., Counts, M., Penick, S., & Popkin, J. (2009). The impact of gender on standardized testing a meta-analysis of existing studies,1987-2007, ProQuest Dissertations and Theses.

Hernandez, R., Conoley, C. W., Israel, T., & Consoli, M. M. (2014). Responding to perceived racial microaggressions: Impacts on the mental health and academic persistence attitudes of Latina/o college students, ProQuest Dissertations and Theses.

Hidalgo-Montesinos, M. D., & Gómez-Benito, J. (2003). Test purification and the evaluation of differential item functioning with multinomial logistic regression. *European Journal of Psychological Assessment, 19*(1), 1–11.

Howard, T. (2010). *Why race and culture matter in schools: Closing the achievement gap in America's classrooms* (Multicultural education series (New York, N.Y.)). New York, N.Y.: Teachers College Press.

Ibarra, R. A. (2001). *Beyond affirmative action*. Madison, WI: University of Wisconsin Press.

Jamieson, J., & Harkins, S. (2012). Distinguishing between the effects of stereotype priming and stereotype threat on math performance. *Group Processes & Intergroup Relations*, *15*(3), 291-304.

Kan, A., & Bulut, O. (2014). Examining the relationship between gender DIF and language complexity in mathematics assessments. *International Journal of Testing, 14*(3), 245-264.

Kane, M. (2013). Validation. *National Conference of Bar Examiners*, 17-64.

Keith, J. (2016). The impacts of microaggressions on the performance of multiracial and monoracial college students. McNair Scholars Online Journal, 10(1), *McNair Scholars Online Journal, 10*(1).

Lan, M., & Li, Min. (2014). Exploring gender differential item functioning (DIF) on eighth grade mathematics items for the United States and Taiwan. Seattle: University of Washington.

Lester, J., Yamanaka, A., & Struthers, B. (2016). Gender microaggressions and learning environments: The role of physical space in teaching pedagogy and communication. *Community College Journal of Research and Practice,* 1-18.

Levine, M., & Levine, A. (2013). Holding accountability accountable: A cost-benefit analysis of achievement test scores. *American Journal of Orthopsychiatry, 83*(1), 17-26.

Levy, R. (2008). 'Third spaces' are interesting places: Applying 'third space theory' to nursery-aged children's constructions of themselves as readers. *Journal of Early Childhood Literacy*. *8*(1), 43–66.

Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of mathematics items associated with gender dif. *International Journal of Testing, 4*(2), 115-136.

Lomax, R., West, M., Harmon, M., Viator, K., & Madaus, G. (1995). The impact of mandated standardized testing on minority students. *The Journal of Negro Education, 64*(2), 171-185.

Los Angeles Community College District. (2012). Investigations for bias, offensiveness and sensitivity. Prepared for the California Community Colleges Chancellor's Office Assessment Validation Training. https://www.laccd.edu/Departments/EPIE/Documents/TestBias.pdf

Madaus, G. F., & Clarke, M. (2001). *The adverse impact of high stakes testing on minority students: Evidence from 100 years of test data.*

Maniotes, L. K. (2005). The transformative power of literary third space. Doctoral dissertation, School of Education, University of Colorado, Boulder.

Mattarella-Micke, A., & Beilock, S. (2010). Situating math word problems: The story matters. *Psychonomic Bulletin & Review, 17*(1), 106-111.

Mayes, C. (2007). *Understanding the whole student: Holistic multicultural education*. Lanham: Rowman & Littlefield Education.

Mendes-Barnett, S., & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education, 19*(4), 289-304.

Messick, S. (1987). Validity. *ETS Research Report Series*, (2), 13-208.

Medina, N., Neill, D. Monty, & National Center for Fair & Open Testing. (1990). *Fallout from the testing explosion: How 100 million standardized exams undermine equity and excellence in America's public schools* (Rev. 3rd ed.). Cambridge, MA: National Center for Fair & Open Testing (FairTest).

Millsap, R. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research*, 30(4), 577-605.

Mizelle, N. B. & Irvin, J. L. (2000). Transition from middle school into high school, *Middle School Journal*, (*31*)5, 57-61.

Nadal, K. (2013). *That's so gay! : Microaggressions and the lesbian, gay, bisexual, and transgender community* (1st ed., Perspectives on sexual orientation and gender diversity series). Washington, D.C.: American Psychological Association.

Office of the Superintendent of Public Instruction (OSPI). (2011). Bias and sensitivity review of the common core state standards in English language arts and mathematics (implementation recommendations report). Prepared by Relevant Strategies.

Pérez, E. (1999). *The decolonial imaginary: Writing Chicanas into history* (Theories of representation and difference). Bloomington: Indiana University Press.

Picho, K., & Schmader, T. (2017). When do gender stereotypes impair math performance? A study of stereotype threat among Ugandan adolescents. *Sex Roles*.

Pierce, C. (1970). Offensive mechanisms. In F. Barbour (Ed.), *The Black Seventies* (pp. 265-282). Boston, MA: Porter Sargent.

Prather, C., & Ruiz, John. (2015). *Nice dissertation, for a girl: Cardiovascular and emotional reactivity to gender microaggressions,* ProQuest Dissertations and Theses.

Rawson, KJ. (2015). Introduction: 'An inevitably political craft.'[Special Issue on Archives and Archiving] *TSQ: Transgender Studies Quarterly 2*(4), 544-552.

Rolfe, G. (2004). Deconstruction in a nutshell. *Nursing Philosophy*, *5*(3), 274-276.

Roth, W. (1996). Where IS the context in contextual word problem?: Mathematical practices and products in grade 8 students' answers to story problems. *Cognition and Instruction, 14*(4), 487-527.

Ryan, K. E., & Fan, M. (1996). Examining gender DIF on a multiple-choice test of mathematics: A confirmatory approach. *Educational Measurement: Issues and Practice, 15*(4), 15-20, 38.

Ryan, K. E., & Ryan, A. M. (2005). Psychological processes underlying stereotype threat and standardized math test performance. *Educational Psychologist, 40*(1), 53-63.

Schley, D. R., & Fujita, K. (2014). Seeing the math in the story: On how abstraction promotes performance on mathematical word problems. *Social Psychological and Personality Science, 5*(8), 953-961.

Schreier, M. (2014). Qualitative content analysis, In Flick, U., Metzler, Katie, & Scott, Wendy (Eds.), *The SAGE Handbook of Qualitative Data Analysis* (pp. 170-183). Los Angeles; London [England]: SAGE.

Skerrett, A. (2010). Lolita, Facebook, and the third space of literacy teacher education. *Education Studies*. *46*(1), 67–84.

Smarter Balanced Assessment Consortium (2012). Bias and sensitivity guidelines. UC-Santa, Cruz, CA: Smarter Balanced Assessment Consortium.

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.

Steele, C. (2010). *Whistling vivaldi: And other clues to how stereotypes affect us* (1st ed., Issues of our time (W.W. Norton & Company)). New York: W.W. Norton & Company.

Steele, C., Aronson, J., & Kruglanski, Arie W. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797-811.

Sterzing, P., Gartner, R., Woodford, M., & Fisher, C. (2017). Sexual orientation, gender, and gender identity microaggressions: Toward an intersectional framework for social work research. *Journal of Ethnic & Cultural Diversity in Social Work, 26*(1-2), 81-94.

Sue, D. (2010). *Microaggressions in everyday life: Race, gender, and sexual orientation*. Hoboken, N.J.: Wiley.

Sue, D. (2010). *Microaggressions and marginality: Manifestation, dynamics, and impact.* Hoboken, N.J.: Wiley.

Taie, S., & Goldring, R. (2017). Characteristics of public elementary and secondary school teachers in the United States: Results from the 2015-16 national teacher and principal survey. First look. National Center for Education Statistics, 2017.

Taylor, C., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education, 25*(3), 246-280.

Tomasetto, C., Alparone, F. R., & Cadinu, M. (2011). Girls' math performance under stereotype threat: The moderating role of mothers' gender stereotypes. *Developmental Psychology, 47*(4), 943-949.

Warne, R., Yoon, M., Price, C., Zárate, M. A., & Lee, R. M. (2014). Exploring the various interpretations of “test bias”. *Cultural Diversity and Ethnic Minority Psychology, 20*(4), 570-582.

Zumbo, B. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. Language Assessment Quarterly, *4*(2), 223-233.

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. ETS Research Report Series, *2012*(1), 1-30.

2.11 APPENDIX 2 A

NAEP Item Coding Rubric for Panel Review Use

Item ID:

Directions: Read the story problem carefully. After you have read each one, please respond to the questions below for *only* the item you indicated above.

Part I: Intrinsic Bias

(*Circle Y or N*)

Does the context of the item reinforce gender stereotypes?	Y	N
If you said yes, please describe.		
Does the context of the item assume that gender identity predicts interest or behavior?	Y	N
If you said yes, please describe.		
Does the context of the item favor one family structure over others?	Y	N
If you said yes, please describe.		

Does the context of the item assume a common family structure for all people?	Y	N
If you said yes, please describe.		
Does the context of the item favor one sexual/romantic orientation over others?	Y	N
If you said yes, please describe.		
Does the context of the item assume a common sexual/romantic orientation for everyone?	Y	N
If you said yes, please describe.		

Part II. Item Diagraming

How many individual people are represented in the item? (*Enter a number*) ☐

Now, please describe each person represented in the question. Some questions may provide this information explicitly, but others may not. If you are not sure of an answer, do not guess;

leave it blank, but make note of your uncertainty in the "Additional Details" section, in which you should add other descriptors used for that person, as applicable.

Person 1 Gender:	
Person 1 Name:	
Person 1 Age:	
Additional details about Person 1:	
Person 2 Gender:	
Person 2 Name:	
Person 2 Age:	
Additional details about Person 2:	
Person 3 Gender:	
Person 3 Name:	
Person 3 Age:	
Additional details about Person 3:	
Person 4 Gender:	
Person 4 Name:	
Person 4 Age:	
Additional details about Person 4:	
Person 5 Gender:	
Person 5 Name:	
Person 5 Age:	
Additional details about Person 5:	

What are they doing?

Please describe the action or behavior of each person represented in the item. If there are multiple people represented, please refer to each as Person 1, Person 2, etc. consistent with your answers to the previous question. (For example: "Person 1 is doing an experiment." OR "Person 1 and Person 2 are playing tennis.")

You are encouraged to include as many details as you see fit.
(For example: "Person 1, Person 2, and Person 3 are in the school band. Person 1 is playing guitar, Person 2 is singing, and Person 3 is playing the drums.")

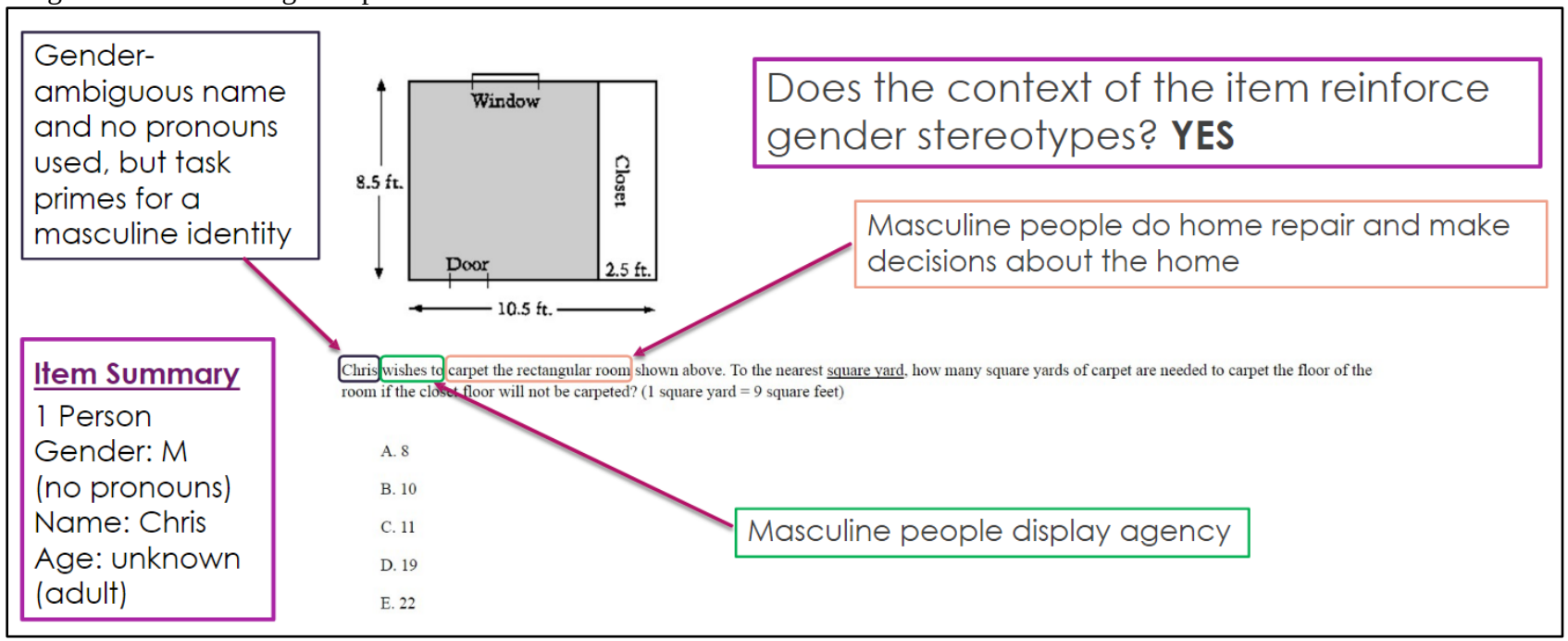
If the answer involves a comparison or competition between two or more people (e.g. a race or a contest), what was the outcome?

Please describe the nature of the competition and indicate who "won," had the advantage, or otherwise "came out on top," referring to each as Person 1, Person 2, etc. consistent with your answers to the previous question. (For example: "Person 1 and Person 2 are competing to see who can sell more chocolate bars for a school fundraiser. Person 2 sold more chocolate bars.")

Thanks! You're all done!

2.12 APPENDIX 2 B

Diagram of Item Coding Components



Example Reviewer Feedback

Reviewer 1	Reviewer 2	Reviewer 3
"Yes. Males are associated with partaking in *house remodeling* projects especially carpentry. Also patriarchal message of men making the big *household decisions; 'man of the house*'. (If Chris is male in the example.)"	"Hard to know if Chris is a masc or femme or neutral name; no pronouns used, but masculine primed; the activity (laying carpet) primes the reader for masc identity; trend of masc people doing *home repair*."	"Chris would be assumed a normal male name, even if it doesn't explicitly say so. And, it is associated with carpeting the floor, which would be perceived as *hardcore masculine work for a household*."

2.13 APPENDIX 2 C

Table 2.2

Summary of Stepwise Regression Analysis for Variables Predicting Gender Stereotype Flagging (N = 352)

	Model 1			Model 2			Model 3			Model 4		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE(B)*	*β*
Collection Wave												
2	-0.89	0.38	.41*	-0.91	0.38	.40*	-0.92	0.38	.40*	-1.01	0.42	.36*
3	-1.07	0.37	.35**	-1.10	0.37	.34**	-1.15	0.38	.32**	-1.18	0.42	.31**
Reviewer												
2				0.88	0.39	2.4*	0.89	0.39	2.4*	.923	0.40	2.5*
3				-0.57	0.30	.56	-0.58	0.30	.56**	-0.14	0.34	.87
Character Gender (F)							0.42	0.30	1.52	-0.42	0.31	.66
Pass Rate										-1.80	0.77	.17*
Graphic (Yes)										-0.32	0.32	.73
Year												
2003 to 2007										0.21	0.40	1.2
2009 to 2013										0.29	0.39	1.3
Content Area												
Data Analysis, Statistics, & Probability										-0.01	0.42	.99
Algebra										0.33	0.44	1.4
Geometry										-0.38	0.57	.68
Measurement										-0.05	0.51	.96
Type (Constructed)										0.10	0.33	1.1
R^2		.05			.10			.11			.15	

Note:. *$p < .05$. **$p < .01$

Table 2.3

Summary of Stepwise Regression Analysis for Variables Predicting Gender Stereotype Flagging by Reviewer

	Reviewer 1			Reviewer 2			Reviewer 3		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Collection Wave									
2	-1.27	0.71	.28	-1.34	0.93	.26	-0.86	0.72	.43
3	-1.12	0.74	.33	0.00	0.97	1.0	-2.04	0.71	.13*
Character Gender (F)	-0.16	0.54	.85	-0.87	0.73	.41	-0.38	0.53	.69
Pass Rate	-2.64	1.51	.07	-1.71	1.62	.18	-1.42	1.29	.24
Graphic (Yes)	-0.21	0.56	.81	-1.73	0.76	.18*	0.42	0.56	1.5
Year (vs 1990 to 1996)			.						
2003 to 2007	0.26	0.71	1.3	-0.35	0.82	.70	0.32	0.68	1.4
2009 to 2013	-0.40	0.68	.67	0.94	0.95	2.6	0.70	0.66	2.0
Content Area									
Data Analysis, Statistics, & Probability	-0.28	0.71	.75	0.71	0.94	2.0	0.19	0.74	1.2
Algebra	-0.34	0.71	.71	2.30	1.34	10	0.32	0.72	1.4
Geometry[+]				-0.20	1.05	.82	-1.65	0.99	.19
Measurement	-0.78	0.87	.46	0.81	1.28	2.3	0.16	0.85	1.2
Type (Constructed)	0.16	0.56	1.2	0.55	0.72	1.7	-0.35	0.56	.70
R^2	0.14			0.14			0.17		

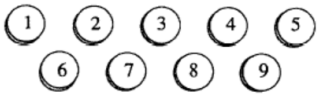
Note:. *$p < .05$. **$p < .01$.

[+] Geometry was excluded from Reviewer 1 analysis due to small subgroup n.

2.14 APPENDIX 2 D

NAEP Item examples

NAEP ID	Item Text
M110901	Ms. Thierry and 3 friends ate dinner at a restaurant. The bill was $67. In addition, they left a $13 tip. Approximately what percent of the total bill did they leave as a tip? A. 10% B. 13% C. 15% D. 20% E. 25%
M1663E1	Bags of Healthy Snack Mix are packed into small and large cartons. The small cartons contain 12 bags each. The large cartons contain 18 bags each. Meg claimed that she packed a total of 150 bags of Healthy Snack Mix into 2 small cartons and 7 large cartons. Could Meg have packed the cartons the way she claimed? ◯ Yes ◯ No Show the computations you used to arrive at your answer.
M1687E1	Tyler drinks 24 fluid ounces of milk each day for 7 days. How many quarts of milk does he drink in the 7 days? Do not round your answer. (1 quart = 32 fluid ounces) Answer: _____ quarts
M013031	<table><tr><th>Score</th><th>Number of Students</th></tr><tr><td>90</td><td>1</td></tr><tr><td>80</td><td>3</td></tr><tr><td>70</td><td>4</td></tr><tr><td>60</td><td>0</td></tr><tr><td>50</td><td>3</td></tr></table> The table above shows the scores of a group of 11 students on a history test. What is the average (mean) score of the group to the nearest whole number? Answer:_____

M016101	 The nine chips shown above are placed in a sack and then mixed up. Madeline draws one chip from this sack. What is the probability that Madeline draws a chip with an even number? A. $\frac{1}{9}$ B. $\frac{2}{9}$ C. $\frac{4}{9}$ D. $\frac{1}{2}$ E. $\frac{4}{5}$
M018401	Six students bought exactly enough pens to share equally among themselves. Which of the following could be the number of pens they bought? A. 46 B. 48 C. 50 D. 52
M018501	Jim has 3/4 of a yard of string which he wishes to divide into pieces, each 1/8 of a yard long. How many pieces will he have? A. 3 B. 4 C. 6 D. 8
M020401	Peter wrote down a pattern of A's and B's that repeats in groups of 3. Here is the beginning of his pattern with some of the letters erased. Fill in the missing letters. A B _ A _ B _ _ _

Chapter 3. AUTOMATION FRUSTRATION: CAN MACHINE LEARNING TECHNIQUES OFFER A SHORTCUT TO CONTEXTUAL BIAS IDENTIFICATION IN MATH ASSESSMENT?

FARAH NADEEM, GABRIELLA SILVA GORSKY, & NIXI WANG

3.1 ABSTRACT

This paper explores the potential for applying Machine Learning (ML) and Natural Language Processing (NLP) methodologies to the process of math test item bias and sensitivity review. Earlier work revealed some serious limitations to the widespread implementation of a human-coded protocol, particularly in terms of how long review takes and in the depth of knowledge required. By teaching an algorithm to identify, categorize, filter, and sort various structural (i.e., word count) and content components of math story problem texts such as gender pronouns, names, keywords, verbs, and nouns, a large portion of the item screening work can potentially be done more quickly, from initial word problem identification to the assignment of items to different activity categories. Results from the implementation of a text classifier identifying items by character gender and activity category for two different data sets reveal stereotypical gender-by-activity associations in two different data sets. Nonetheless, one of the lingering challenges to automation is the perceptual component to identifying and describing patterns in contextualized gender representations in math story problems.

3.2 INTRODUCTION

While the use of algorithms to parse out complex textual and numeric data is certainly well-established, recent advancements in Machine Learning (ML) and Natural Language Processing (NLP) have meant a proliferation of studies exploring the range of applications. With a few exceptions (Luckin, 2018; Ma, 2017; Spikol et al., 2016), a bridge is still being built between these novel methods and the range of potential uses in educational settings. Assessment provides a rich context for applying these techniques, but the research is lacking outside of certain specific application s such as Computer Adaptive Testing (CAT) and automatic scoring of textual test data like essays (CAT- Attali, 2018; Arendasy & Sommer, 2012; Gierl & Lai, 2012; Gierl, Lai & Chen, 2018; auto-scoring- Powers et al., 2002; Singley & Bennet, 1998; Srihari et al., 2008). The use of NLP for screening items for social bias is one historically relevant, but underexplored potential use.

Earlier work used human coders to evaluate grade 8 contextualized math assessment items (i.e. "word problems") for the presence of gender stereotypes and distinct patterns in gender representations (Gorsky, Wang, and Phelps, 2019, not yet published). Overall, the coding protocol was effective at identifying linguistic, contextual, and otherwise embedded associations between gender identity and item setting. However, it was also time-consuming and labor-intensive, rendering the use of this approach to item screening less desirable or feasible from a cost-benefit standpoint. Nonetheless, the task itself and the findings of the earlier work suggest that this is not a procedure to simply skip. The findings provide significant evidence that our current methods for item sensitivity and bias review do not fully neutralize the item pool. In fact, they call into question if this even possible.

By applying NLP, there is a potential to automate the coding process. Techniques such as automatic text classification can be used to learn to identify different aspects of the coding protocol, and can provide both the coding and the "confidence" of the predictions for items. This can in turn considerably reduce the amount of human effort required for coding, where the focus can just be on refining the coding for items that are assigned a low confidence score by the NLP algorithm. This can also allow for a "first-pass" at the set of items under study, so that items that can benefit from more thorough analysis can be flagged automatically.

3.3 Framing the Problem and Methodology

A few bodies of research inform this study. Specifically, this work sits at the intersection of Artificial Intelligence methodologies, culturally responsive pedagogy, and equity in assessment. While some educational applications of these methodologies exist, this study seeks to extend those applications. Bias in assessment is a long-studied phenomenon (Au, 2011; Green & Griffore, 1980; Lomax et al., 1995; Madaus & Neill, 1990). but computer automation has not been widely applied to addressing sociological features of test content. However, new research on the presence of bias in text- classifying and text-producing algorithms suggests that there is a unique opportunity to leverage equity in both the methodological and substantive research spaces (Bolukbasi et al., 2016; Garg et al., 2018).

3.3.1 *Machine Learning and Natural Language Processing*

Machine learning is now becoming common for applications across a multitude of contexts, including education (Luckin, 2018; Ma, 2017; Spikol et al., 2016). Natural language processing, or NLP, is the intersection of computational linguistics and machine learning. Within NLP, text classification is a well explored problem. Typical systems are trained to classify texts based on

topic (Yang et al., 2016), sentiment (Liu et al., 2015) and other aspects such as level of coherence (Lai et al., 2018). NLP provides tools to automatically identify different aspects of texts, such as named entities, pronouns, topics, etc. Recent work (Sap et al., 2017) has used NLP to explore gender roles and agency in movie scripts, analyzing how gender interacts with given roles and dialogues. This technology allows for the quick parsing of textual data for further exploration.

3.3.2 *ML/NLP applications in Education*

There has been extensive work on data mining in education (Napolitano et al., 2015, Sheehan et al., 2013). There has been recent focus on automatic evaluation of student responses, especially in the context of online courses and large scale testing (Attali & Burst, 2006), reflecting a demand for algorithmic systems for assessment use. Work by Nadeem and Ostendorf (2017) was the first work exploring the automatic creation of a Q-Matrix for science assessment items using automatic neural network-based text classification. This work explored the difficulty of science tests, applying a more accurate neural net approach than traditional linear model approaches. The advantage of automatic classification, as opposed to manual coding, is that it is faster to label and re-label large sets of items without the cost of domain expert annotation. Additionally, there has also been work exploring automatic readability analysis of science texts (Sheehan et al., 2013). While readability analysis is done extensively in the context of language learning, there has been less focus on science and math, particularly assessment items. Work by Nadeem and Ostendorf (2018) explored automatically quantifying language difficulty of science texts, including shorter texts like assessment items. These works indicate that NLP can be used in the context of STEM assessment items.

3.3.3 *Bias in Algorithms*

Machine learning algorithms are trained on large data sets, and recent work has shown that these algorithms pick up social biases, both implicit and explicit, as well as stereotypes. This has been observed in natural language processing, where Bolukbasi et al. (2016) demonstrated that algorithms can amplify the biases in certain cases. An aspect that has been explored by Garg et al. (2018) is how word representations learned from data can help quantify biases observed in society. While the research in this area is relatively new, there is evidence that any machine learning algorithm can generate biased results. As educational applications of machine learning emerge, we can expect to see similar trends emerge that reflect the biases in the data. However, as demonstrated by Garg et al. (2018), we can use machine learning algorithms to quantify biases. In the context of educational assessment, this presents a unique opportunity to leverage a controversial method in the service of exposing ingrained social biases.

3.3.4 *Culturally Responsive Assessment*

There are a number of ways test items are assessed for appropriateness of use. In particular, tests and test items are consistently questioned in terms of their cultural responsiveness or adaptivity. However, examining the literature shows that sparse efforts have been given to address patterns of linguistic and cultural communication (Solano-Flores & Nelson-Barber, 2001), and cultural assumptions (Parrott et al., 2000) in assessment items. Research indicates there is a need for procedures that help researchers and assessment developers to examine and calibrate cultural validity of assessments. While some important work has been done in this area, especially to address the needs to English language learners (ELLs), students with disabilities, and other marginalized students, the specific study of gendered language is less researched (Abedi et al., 2001; Abedi & Levine, 2013; Mattarella-Micke & Beilrock, 2010). Gender represents only a

single facet of identity, but it serves as a salient starting point for exploring broader machine learning applications to bias and cultural representation review. The application of ML/NLP to the exploration of cultural responsiveness in assessment offers one way to explore cultural validity across a large number of assessment items.

3.3.5 *Impacts of Gender Bias in Assessment*

Conceptualizing gender bias in terms of the forms and presence of implicit bias in math item is necessary to unveil the underlying mechanism--how it is facilitated and perpetuated by testing items and practice—and open up the discussion beyond achievement scores onto social consequences of testing. Such a conceptualization not only holds normative and expressive value, but it also creates the space to talk about what it would look like to build assessment tools carefully designed to combat discrimination in its modern forms.

Decades of research in social psychology has confirmed that performance can be undermined when a person is triggered a negative stereotype about their identity group when that identity is salient. Theories undergirding research would include prescriptive gender stereotypes (Prentice & Carranza 2002), discrimination in the form of either benevolent sexism or hostile sexism (Glick & Fiske, 2001), stereotype and social identity threat (Steele & Aronson, 1995; Murphy et al., 2007; Sekaquaptewa & Thompson 2002), and even unconscious gender-stereotypical cues in the environment (Cheryan et al., 2009). Moreover, the formation and function of gender stereotypes, norms, and roles in assessment show that testing as a social instrument fails to reflect and espouse those same gender equality values (Gayles, 2011; NSF, 2016; Shapiro & Williams, 2012; Smeding, 2012). For example, students who encounter stereotype type threats may hold less motivation in answering a correct response or simply

experience cognitive depletion, which can lead to underperformance of woman and girls of color in achievement tests.

The science education and testing literature is well-aligned with math in terms of addressing gender-subject interaction, be it about representation in industry or academic spaces. Alignment with stronger gender–science stereotypes (implicit associations and endorsement of male superiority in science) leads to women less identified with science and, in turn, weaker science career aspirations (Cundiff et al., 2013). Similarly, stronger implicit math-male stereotypes corresponded with more negative implicit and explicit math attitudes for women than positive attitudes for men (Nosek, Banaji, & Greenwald, 2002). In other words, women were negatively impacted by the stereotypical representations to a greater extent than men were positively impacted, suggesting that the target group, in this case women, is more susceptible to the effect of bias.

But the impact of these associations isn't limited to impressions. When women's gender identity was linked to their performance on a math test, women with higher levels of gender identification performed worse than men, but those with lower levels of gender identification performed equally to men (Schmader, 2002). Essentially, women and girls who possess strong implicit gender-math stereotypes have these stereotypes chronically accessible and therefore may activate gender-math stereotypes even in the absence of stereotypic cues within the test-taking environment. This has major implications for the ways we understand the impact of test content vis-a-vis the representation of people of different genders.

3.4 Research Rationale

This research focuses on the potential for applying Machine Learning (ML) and Natural Language Processing (NLP) methodologies to the process of math test question review. Earlier

work revealed some serious limitations to the widespread implementation of an evaluatory protocol like the one used for human coding, with the greatest burden being time-intensity. The hand-coded approach is slow, collaborative, and iterative, and therefore costly, unfortunately making it a less desirable intervention. If there was a way to automate even some parts of evaluative process, the feasibility of implementing additional bias screening procedures improves dramatically. ML and NLP approaches provide a potential way to streamline the application of my rubric to large numbers of items.

By teaching an algorithm to identify, categorize, filter, and sort various structural and content components of math story problem texts such as gender pronouns, names, keywords, verbs, and nouns, a large portion of the item screening work can potentially be done more quickly, from initial story problem identification to the assignment of items to different categories of representation. Nonetheless, one of the lingering challenges to automation is how to capture the perceptual component to identifying and describing problematic patterns in math story problems. That is, how do we translate subtle human interpretations of historically and culturally-embedded knowledge into a series of straightforward tasks that can be programmed algorithmically? Similar to the specificity of DIF analysis required for flagging systematically biased items, meeting a certain metric criteria doesn't necessarily mean bias has been detected. Oftentimes, the *why* in "why is this problematic?" is unclear, even to human reviewers. Given this, we hold some skepticism that algorithmic approaches can *fully* replace human review. In order to test this out, we can use data from an ML/NLP application of the original principles of the human-coded rubric to identify differences between computerized and manual item coding/categorization.

Through a cross-disciplinary collaboration, we aimed to compare the efficacy/efficiency of a gender-by-activity identification algorithm to Gorsky, Wang, and Phelps' earlier work in terms of computer-human alignment, relative nuance of categorization, and the potential for large(r)-scale implementation. Specifically, we wanted to apply a similar coding logic to that used in previous research but using ML/NLP methods to automate the item review, identify strengths, weaknesses, and limitations of the ML/NLP approach compared to manual coding, and articulate the potential applications of ML/NLP approaches to item review given the findings of the algorithmic pilot.

3.4.1 *Research Questions*

Given the theoretical and methodological challenges to doing a more comprehensive, equity-focused item review using human coders, we articulated the following research questions:

1. Can crowdsourced[5] annotation be used to train efficient NLP systems in the context of math assessment items?
2. How well does the text classifier sort and organize math assessment items by gender and activity type for a set of training items as compared to a set of control items?

3.5 METHODS

This work involved 4 main stages: manually coding 20% of items with a researcher-defined "correct answer" for activity category, crowdsourcing item activity category annotations for a sample of math questions, training an automatic text classifier using the activity annotations from the first stage for activity classification and existing programs for gender classification, and applying the classifier to the original set of training items and to a "control" set of benchmark items from a large-scale nationally-administered standardized test. In the first stage, we

[5] *Crowdsourcing* in this context refers to the use of a public paid task performance platform to recruit anonymous individuals to do a specific set of pre-defined actions. It is used extensively in ML research.

crowdsource to get annotations for activity categories for a set of 2,135 items. These items are now considered "labelled" with the "true" activity category. Then we use the classifier to find activity categories for labeled items (for verification of accuracy) *and* unlabeled items (to test the application with other data), thus cutting back substantially on the hand coding required.

3.5.1 *Math Item Sources*

For training the classifier we use the Algebra Question Answering with Rationales (AQuA) data set, which is a publically available set of about 100,000 algebra questions (not identified by grade or difficulty) (Ling et al., 2017) mined from a broad range of educational resources. Control "benchmark" items consisted of 440 released NAEP grade 8 math items spanning 1993 to 2013. Items in this set are not limited to algebra content (23% of NAEP items used) and contain geometry (21%), data analysis, statistics, and probability (13%), measurement (17%), and number properties and operations (26%). While the NAEP items are known to be rigorously reviewed pre- and post-test-administration and contain overall familiar contexts, the AQuA items contain a wider range of contexts, including many that would not be included in NAEP due to being inaccessible to all test takers (NCES, 2016). For example, the AQuA set contains references to non-U.S. number and monetary systems, explicitly traditional marital roles, and warfare. In a way, the two sets represent opposing ends of the range of math-related educational materials a student might actually see in real academic life.

3.5.2 *Linking Gender Roles to Activities*

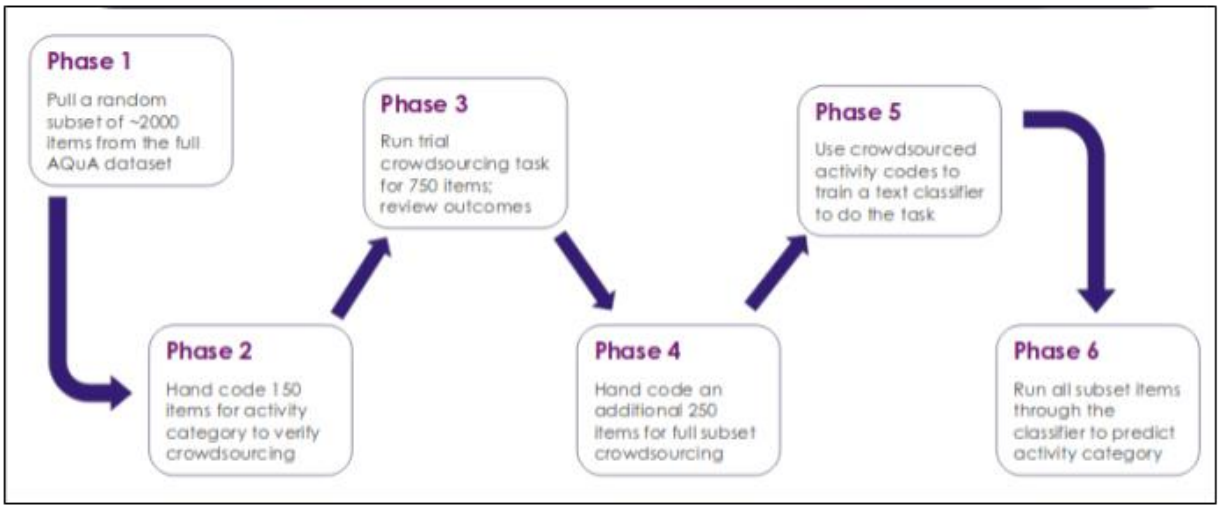
For all the items that are labelled using crowdsourcing, we automatically identify all personal pronouns using a Python based NLP toolkit, the Natural Language Toolkit (NLTK). Once the pronouns have been extracted, they are then automatically mapped to predefined groups that

represent gender identities via personal pronoun use. *Masculine* identity was defined as ‘he,’ ‘him,’ ‘his,’ or ‘himself’; *feminine* was defined as ‘she,’ ‘her, ‘hers,’ or ‘herself; *neutral-self* as ‘I,’ ‘me,’ ‘we,’ ‘us,’ or ‘ourselves’; *neutral-other* as ‘you,’ ‘they,’ ‘them,’ or ‘themselves’; *non-person* as ‘it’ or ‘itself.’ This then allows for studying the linkage of gender roles to the activities used as context in the assessment items.

3.5.3 *Crowdsourcing Item Annotations*

Crowdsourcing is a data collection method which uses human raters to perform pre-defined tasks quickly, typically through an online task performance platform, and it is an option used extensively for annotating large publicly available datasets. In this work we use crowdsourcing to quickly annotate the subset of AQuA items we use to train the classifier, which eliminates the need for hand-coding by the researcher and speeds up the process significantly (for reference, all 2,135 were annotated within only a few hours of task publication). The crowdsourcing platform we use is Figure Eight, which allows more built-in control over the quality of annotation compared to Mechanical Turk, which has been criticized for its susceptibility to data contamination (Kees et al., 2017; Thomas & Clifford, 2017). As opposed to having domain expert(s) hand code all or part of the data set, we use the following approach to automate this process. Crowdsourcing item activity categories occurred iteratively and the steps are summarized below and in Figure 3.1.

Figure 3.1. Stages of item coding and crowdsourcing.



Crowdsourcing, though a major component of the process, represents only the first step in building and evaluating the classifier. First, the data to be annotated (item texts) is uploaded to the chosen platform as a *name*.xls or *name*.csv file. The task is then made available to the annotators who are signed up with the platform. Annotators can label any number of samples. Quality control is maintained through "test" items, which are annotated by the creators of the task, and interspersed in the annotation questions. Annotators have to maintain a preset level of accuracy on these test items (in our case 80%) to be able to continue the task. A limitation of crowdsourcing is that there is not a lot of control over who annotates the data. We set the geographical location of annotators to North America, however we did not have additional control over who annotated the data, or how many samples are annotated by each annotator. For the Figure Eight platform it is not possible to get the demographic information for annotators.

For the crowdsourcing, a set of activity categories is defined for items. We define 11 categories and include a "None" category in case the question has no activity (Table 3.1). The list of categories is by no means exhaustive, but captured the majority of item contexts in both the NAEP and AQuA sets. Categories originated from those identified in the human-coding task

that catalyzed the present study and were adapted to better fit the training set. Some items could be cross-coded for multiple categories, but by looking at overall annotator agreement, we somewhat mitigate the extent to which individual item-category salience is driving item categorization across coders.

Table 3.1.
Item Context Categories and Definitions with AQuA Examples

Category	**Definition**	**Example**
Academics/ School	students, teachers, classrooms, courses, majors, grades, tests	"The average age of students of a class is 15.8 years. The average age of boys in the class is 16.4 years and that of the girls is 15.4 years. The ratio of the number of boys to the number of girls in the class is?"
Animals/ Pets	domestic/wild animals, pets, live animals, cats, dogs, fish	"John has 1,210 Dogs in his house. If exactly 40 percent of all dogs in home are tagged, what percent of the untagged dogs must be tagged so that half of all dogs in the home are tagged?"
Arts/ Creativity	expression, music, jewelry, painting, crafts, singing, CDs	"To fill an art exhibit, the boys in an art course are assigned to create one piece of artwork each in the following distribution: 1/3 are sculptures, 1/8 are oil paintings, 1/2 are watercolors, and the remaining 10 pieces are mosaics. How many boys are in the art class?"
Business/ Money	investment, banking, finances, retail sales, income/salary	"Anthony and Michael sit on the six member board of directors for company X. If the board is to be split up into 2 three-person subcommittees, what percent of all the possible subcommittees that include Michael also include Anthony?"
Cars/ Vehicles	Cars, driving, transportation, trains, ships, travel	"How many seconds will a 220 meter long train take to cross a man running with a speed of 8 km/hr in the direction of the moving train if the speed of the train is 80 km/hr?"
Exercise/ Sports	walking/running, team sports, cycling, coaching, fitness	"A and B go around a circular track of length 600 m on a cycle at speeds of 18 kmph and 48 kmph. After how much time will they meet for the first time at the starting point?"

Food/ Drink	food items, groceries, dining, cooking, snacks, beverages	"John's Ice Cream Shop sells ice cream at m cents a scoop. For an additional n cents, a customer can add 2 toppings to his or her sundae. How much would a sundae with 2 scoops and 2 toppings cost, in terms of m and n?"
Home Maintenance	home repair, non-commercial building/construction, appliances	"Thomas bought a table for Rs. 2000 and a chair for Rs. 1000. After one month of usage he sold table to David for Rs. 1800 and chair to Michael for Rs. 900. What is the total percentage of loss incurred to Thomas?"
Industry/ Trades	manual labor or industrial work setting, machine work	"If 12 welders work at a constant rate, they complete an order in 8 days. If after the first day, 9 welders start to work on the other project, how many more days the remaining welders will need to complete the rest of the order?"
Science/ Research	lab science, space, experimentation, survey research, -ology	"From a group of 16 astronauts that includes 7 people with previous experience in space flight, a 3-person crew is to be selected so that exactly 1 person in the crew has previous experience in space flight. How many different crews of this type are possible?"
Social Relationships	social behavior or interactions, family dynamics, friendship	"My grandson is about as many days as my son in weeks, and my grandson is as many months as I am in years. My grandson, my son and I together are 140 years. Can you tell me my age in years?"
None	no setting, no context, dis-identified "actors"	"If n is a three-digit prime number and j is an integer, which of the following is NOT a possible value of k, where k is the smallest positive integer such that $n - 5j = k$?"

We then create examples for annotators for each category, and code 20% of AQuA items with 'gold' labels (researcher-defined "correct" answers for which activity category each item fit into best; 400 total; refers to phases 2 to 4 in Figure 3.1) to serve as test questions for annotators during the crowdsourcing. In this step, each item was assigned an activity category that fits best

by one of the researchers and then informally cross-checked by the others. Then, for each item, we collect the judgment from three annotators per item (see Appendix 3 B for task example). Using as the final label the one identified as the majority vote, we then use this labeled set of items for training the classifier.

3.5.4 *Training the Classifier*

Text classification is a well-established problem in NLP. For this work, since we do not have a large amount of annotated data, we make use of pre-trained model that can leverage information learned from other tasks, such as prediction of the next sentence, on other larger data sets. The model we specifically use is Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2018). This a neural network architecture pre-trained on two language based tasks, predict a word given the context words, and given a set of sentences, predict whether a sentence is the next sentence or not. For our work, we take the pre-trained BERT model, and then fine-tune it to predict the activity category. For this work, which is a pilot, we assume that each question has one associated activity, so the classifier is trained to predict one of the categories as being the most likely for each question. This can easily be extended to a multi-class problem, where we predict multiple categories for each question. The classifier (BERT) is trained using labeled data. For the training process, each item is mapped to a vector via the neural network, starting with dense 768 dimensional vectors representing each "token" in an item (stemmed words, punctuations etc.).

These vectors are then passed through the neural network, which performs a set of linear (addition and multiplication) and non-linear operations to generate one final vector x for the item. This vector x is then multiplied with a weight matrix w, and added to a bias vector b, resulting in a vector of n number, where n is the number of categories we are classifying into.

This vector is then transformed into a set of probabilities that sum to one across the n categories. The category with the highest probability is selected. Then the loss is computed using a cross-entropy loss based on the true label. The network is then trained, or fine-tuned, to minimize the loss, so that for the training data, the generated probability is highest for the true label. Once the network is trained, it can then be used to take a new, unlabeled item, and generate a probability for that item across the item categories. The category with the highest probability is then selected as the true category. Since the loss is not computed for the new unlabeled data once the network has been trained, we don't need labels, and can generate labels for unlabeled items. For evaluating how well the trained network performs, a set of labelled items is held out during training, and we use the network to generate labels and then compute the accuracy, or the number of items the classifier labelled correctly.

For training, we use a loss function, log likelihood, and use gradient descent to update the neural network to minimize the loss for each instance of training. A set of labeled items is held out and used as a development set. We took a random sample of 10% of items in each set of data (AQuA and NAEP). The purpose of the development is to find the optimal parameters that give the best classifier performance, including indicating the point at which the network has been sufficiently trained. Once the network has been trained, it can then be used to predict labels for new unseen items.

For a trained classifier, the output is a set of probabilities that the item input to the classifier belongs to each of the n activity categories. The category with the highest probability is selected, with the associated probability being the confidence score. For example, if we have two classes *a* and *b*, the classifier will output two probabilities for the item belonging to *a* or *b*, which sum to one. In a confident case, *a* can be assigned a probability of 0.99 and *b* a probability of

0.01. We can have the same decision is a less confident case, where *a* gets a probability of 0.51 and *b* gets a probability of 0.49. Confidence can be computed for each item that the classifier is used on.

3.6 RESULTS

3.6.1 *Crowdsourcing Task*

As noted, crowdsourcing occurred in multiple waves. At each stage, a subset of items was hand-coded and released for crowdsource coding, until all 2,135 items were released. The first wave (750 items total, 150 ‘golden’) produced item label agreement of 67% or more on 75% percent of the golden items. This means that at least 2 out of 3 coders agreed with the ‘golden’ activity label identified by us in the hand-coded subset. The second wave (2135 items total, 400 ‘golden’) produced 65% item label agreement of 81.4 % of the golden items. This represented a significant improvement over the smaller subset and given that this is not a standard classification task, and there are no standard metrics for evaluating the performance of the annotation task, these results were considered adequate to proceed to writing the classifier script.

3.6.2 *Text Classifier Outcomes*

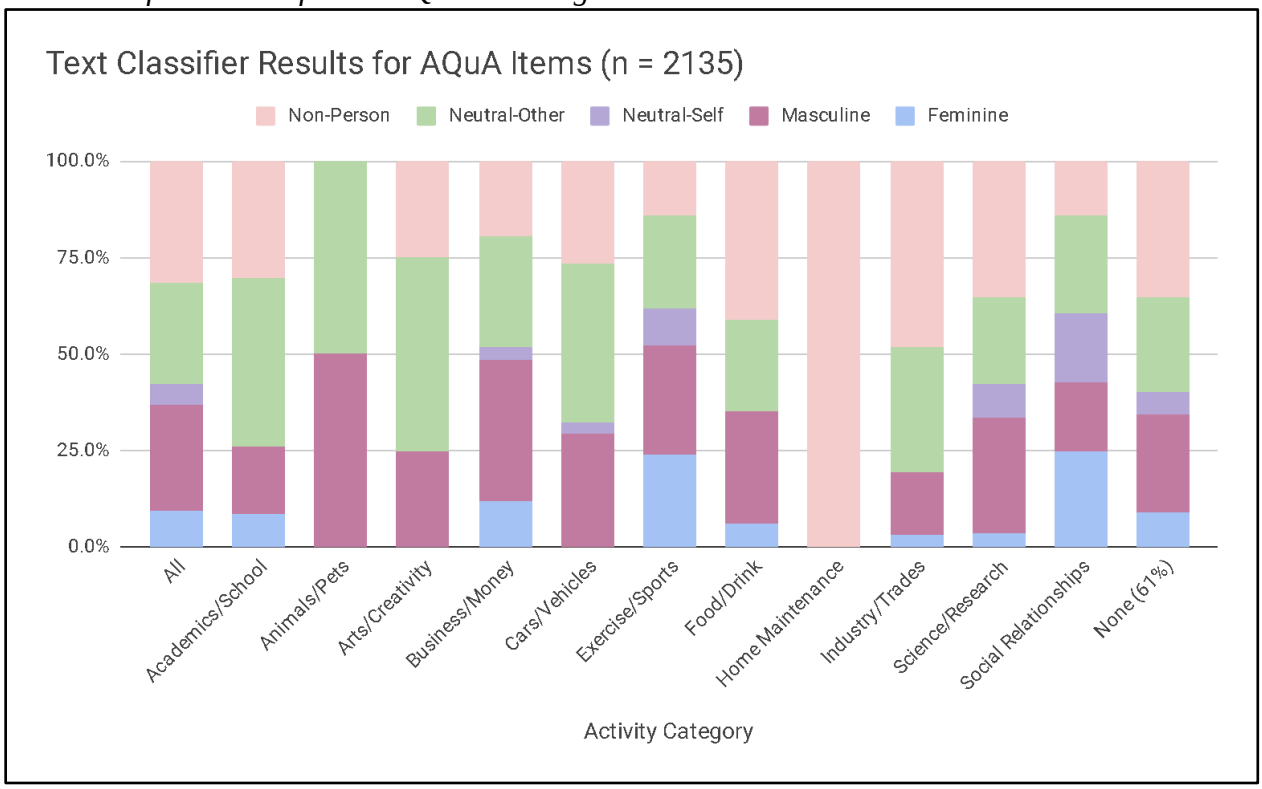
The text classifier produced interesting, but unsurprising results when it came to gender-by-activity mapping. The following pronouns were extracted: 'me', 'you', 'it', 'itself', 'he', 'them', 'i', 'himself', 'him', 'themselves', 'us', 'she', 'we', and 'they'. We identified 5 major categories for organizing pronoun groups: masculine (he, himself, him), feminine (she), neutral-self (I, me, we, us), neutral-other (you, them, themselves), and non-person (it, itself). There is a distinct absence of the pronouns “his,” “hers,” and “herself” in the item texts. Figure 3.2 shows the distribution of items in each pronoun category for the overall test set of 2,135 items used to train the classifier.

In general, there is over-representation of masculine pronouns across the item groups. This pattern is evident within context categories, as well. Figure 3.3 similarly shows this distribution for the benchmark items. It is important to note that the "None" (no setting or context) classification over-captured item contexts based on our annotations. This means that items are being assigned to the "None" category when they should be in one of the 11 categories (or, perhaps in a not-yet-defined activity category, which is addressed further in the discussion).

Distribution of Personal Pronouns Across and Within Training Item Context Categories

From the ML-based automatic text classification of gender groups and social activities output from the classifier, we can see that differentiated gender expectations and gender presence are established throughout the 2,135 math items (Figure 3.2), with masculine frequency (13.0%)

Figure 3.2.
Text Classifier Results for the AQuA Training Items



Note. In the chart above, the 61% next to "None" refers to the percent of all items that were placed in this category.

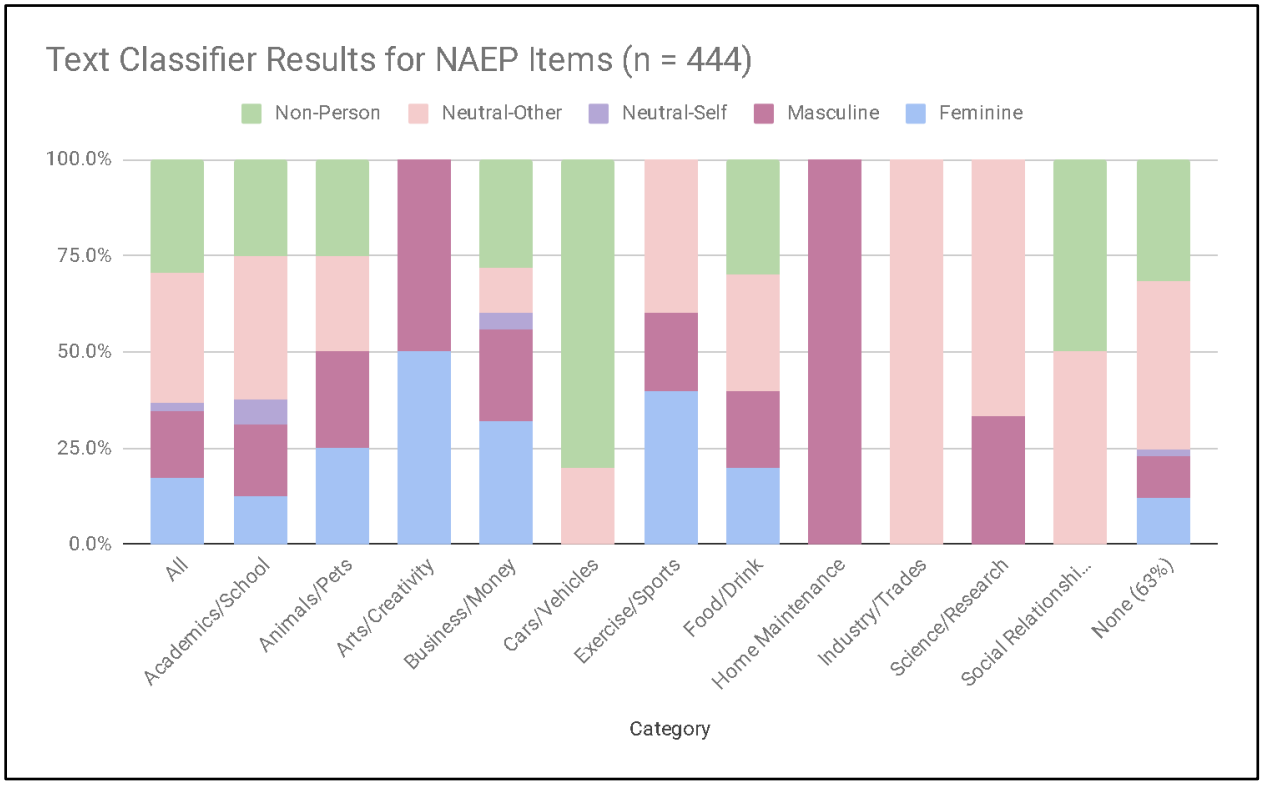
being almost three times greater than feminine (4.5%). Disparities of gender representations exist in manifestations of social activities such as “Industry/Trade,” “Cars/Vehicles,” and “Science/Research” in our sample. The weighted masculine/feminine ratio is the largest in the “Business/Money” category, while “Social Relationships” is the only category that we see more feminine features being identified than masculine features (8.2% compared to 5.9%), and “Exercise/Sports” is the category where women make up the highest percentage in the subtotal items (12.8%).

Distribution of Personal Pronouns Across and Within Benchmark Item Context Categories

We also tested the relative accuracy of the classifier on the NAEP items (Figure 3.3) Overall there was a balance between masculine (5.2%) and feminine (5.2%) pronouns used

Figure 3.3.

Text Classifier Results for the NAEP Classification Task



Note. In the chart above, the 63% next to “None” refers to the percent of all items that were placed in this category.

throughout the item set across categories. However, certain activity categories show clear overrepresentation of certain genders. In particular, two activity categories contained only masculine or neutral: home maintenance and science research. Consistent with the AQuA classification results, the relevance of these gendered messages in terms of occupation and income becomes more obvious when we begin to look at the relative absence of women in STEM fields (Gayles, 2011; Settles et al., 2016; Steffens, Jelenec, & Noack, 2010).

"Arts/Creativity" and "Animals/Pets" were the only categories which showed gender balance (33.3% and 11.1%, respectively). Interestingly, "Business/Money" and "Exercise/Sports" were marked by feminine pronoun overrepresentation. While this might suggest a deviation from stereotypical representations of women, it is actually consistent with earlier NAEP review (Gorsky, Wang, & Phelps, not yet published) that revealed an association between questions featuring women and girls that contained or were focused on earning or saving money. Relatively fewer of these items featured masculine characters.

3.7 IMPLICATIONS

The primary goal of this research was to explore the potential use of Machine Learning (ML) and Natural Language Processing (NLP) in the assessment item review process. The results of this novel exploration providing encouraging evidence that this or similar processes may be considered as a viable method for item evaluation that is neither a replacement for, nor a duplicate of existing bias identification methods. This exploratory study begins to demonstrate the potential and limitations of computer-assisted item evaluation in assessment creation and evaluation process.

It has significant takeaways for test creators and the hope is that testing companies will use this and other similar work as a model for addressing patterns in item representation without

sacrificing efficiency of process. Expert human reviewers should never be fully removed from the procedure, but this approach can be used to streamline existing sorting/organizational procedures. Perhaps the greatest takeaway for test developers is the fact that this can be done on items pre- or post- administration. This sort of detailed technical review is not usually possible pre-administration and could help shorten post-administration procedures.

3.8 Discussion

Test developers face a number of challenges in the process of creating reliable, accessible, informative, and fair assessment items. The two primary considerations, *structure* and *content*, can seem straightforward to address as a test developer, but complex dilemmas arise when trying to create tests for an increasingly heterogeneous student population. Despite efforts to make items that are "universally accessible," the current go-to methods remain limited and rely heavily on simply eliminating items that "perform poorly" in post-administration item analysis. This means that a) students are still seeing those items, b) there's no holistic (test- , department-, or organization-level exploration of what themes are coming up as problematic and why, and c) the assumption remains that a culturally responsive test is a culturally neutral test. When it comes to representation of identity in math tests, it is clear that there is a disconnect between the goals of item review for cultural relevance and accessibility, the methods used to explore these relationships, and the impact this has on students' self-concept.

Given the persistent gender differences in math achievement and STEM disciplinary representation, and extensive research on the factors underscoring such differences [including math assessment item contexts (Kan et al., 2014; Lan & Li, 2014; Li et al., 2004; Mendes-Barnett & Ercikan, 2006; Taylor & Lee, 2012)], one might think that we have made substantial strides when it comes to how we represent opportunities and create narratives of equality for

girls and other marginalized groups. However, the results of our classifier of test item narratives is telling a different story. Gender-based social activities are the most prevalent and deep-rooted inequality norms that despite egalitarian values, remain as a challenge for consistently entrenching gender-based stereotypes (Kerr & Multon, 2015; Martin & Ruble, 2004). The distribution of gender pronouns clearly show that those gender bias exist both at a structural, cultural level of testing, but also within items in interactions with represented social identities. Because gender biases are cultural rather than individual, those cues within the testing environment will affect whether implicit biases are experienced among this generation of students who take tests.

In terms of performance, using the crowdsourced annotations to train our classifier, we were able to parse items based on activity category. Using automated tools for identifying pronouns gave an efficient way to identify gender roles in the items, which were then linked to the activity categories defined in the crowdsource annotation task. While this represents a great success in terms of streamlining item review compared to human coding, there are definite limitations to the level of nuance and subtlety captured compared to human coders. In the human coding task, reviewers were able to give supportive qualitative evidence for why items were problematic, which was a significant advantage over other score-based methods like differential item function (DIF) procedures.

While human perception is necessary to fully understand the patterns of representation that arose in our classification study, much can be discerned about gender representation in math items more generally when the overall distribution of items is observed. Further, whereas in large-scale test development, items need to be evaluated more quickly, our method offers a relatively simple way of identifying over- and under-representation of identity and activity

associations. This is not a comprehensive solution to the issue of assessment bias more broadly, but it does offer a way to utilize a new technology in service to equity.

Rather than offering our approach as a replacement for existing methods, we see it as an opportunity to expand what constitutes current best practice and challenge test developers, computer scientists, and math educators to see the item review process differently. We can and should continue to use these various methods in tandem to help us better understand both what is happening in tests and our own evolving framework for identifying and addressing inequity in educational contexts.

One question we are still grappling with is how to use this approach in a way that truly serves the greater mission of envisioning assessment that is not only culturally *responsive*, but also culturally relevant. Large-scale, standardized testing is, itself, under increasing scrutiny as the ideal or even preferred way to measure student learning in our pluralistic global society, with the negative consequences of exhaustive focus on assessment and repeated exposure to the stereotypical gender representations in test-related materials overshadowing potential positives, especially for students holding marginalized identities (Au, 2007, 2011; Green & Griffore, 1980; Howard, 2010; Lomax et al., 1995). It can be challenging to imagine them functioning in a way that disrupts the existing model of educational accountability, but few alternatives seem viable. In a small way, our use of an NLP classifier for gender and activity on math test items creates a liminal space for thinking about the item texts as cultural products that are historically and politically situated, effectively "de-neutralizing" them.

3.9 LIMITATIONS

There were three primary limitations to our study: training data item categories, activity category scope and intuitiveness, and overall item quality. The original set of human-coded items that

sparked this project only contained 120 texts. Since those were needed for testing the classifier, an additional corpus of items was needed for crowd-coding. We were able to find a rich and extensive item set to use, but it also meant compromising perfect alignment with the original set of "benchmark" questions. As a specific example, the training set tended to have fewer items describing individuals performing tasks, favoring those describing contextualized, but non-social themes, such as finance or transportation. While the classifier did a good job assigning AQuA corpus and NAEP benchmark items to appropriate categories, incongruity between training and benchmark items in terms of structure and contexts inevitably played a role.

In addition, the activity categories were closely, but not perfectly mapped to the training set in order to create a classifier that would work well on the benchmark items, too. This meant that many items were not easily placed into one of the 11 identified categories, as was evidenced by the large number of items ultimately classified as 'None.' It is unclear how much this can be attributed to the contextual scope of training set or the intuitiveness of item categories to the crowd coders, but we may be able to discern the answer by returning to the training data output to and perhaps tweaking and re-administering the coding task. Regardless, the "None" category ended up playing a significant role in how we were able to analyze and interpret the classifier data. The expectation is that with more crowdsourcing, we can improve the cross-classification problem with regards to the "None" category. Additionally, we can create a subtask that asks crowdsourcers to label item as "containing a context" versus "containing no context" (the 'true' "None") to begin parsing out the items that are misclassified from the ones that need an activity category not already outlined. Crowdsourcing for additional categories would improve the classification task, as well.

3.10 Future Research

This study provides strong empirical support for the use of ML/NLP methodologies alongside existing large-scale assessment item review procedures. Future work would seek to resolve some of the main limitations to the present study. What our classifier lacked most compared to human coders was subtlety. In this iteration, it could only identify gender and activity category. Through further crowd-sourcing, future work will extend the classifier to account for more contextual nuance, such as other sorts of identities (race, ability, class) and other pertinent linguistic features, such as representations of agency vs. passivity (Sap et al., 2017).

It would also be worthwhile to examine how this might be integrated with other test-development algorithmic approaches, such as by applying a similar, and perhaps more rigorous, logic to computer-generated items, which are becoming more standard, especially in Computer Adaptive Testing (CAT) settings (Attali, 2018; Arendasy & Sommer, 2012; Gierl & Lai, 2012; Gierl, Lai & Chen, 2018). Computer generated items are notoriously inconsistent in terms of overall "natural-language" quality and sets of computer-generated items often contain many near-duplicate texts (as we saw in our training set); both problems are often the product of low quality source data. Our procedure sets the groundwork for an application that would reveal gaps in target content. A long-term testing industry goal would be to synthesize these two processes to automatically correct problematic patterns in item representations. That is, when a set of human- or machine-written items is passed through the classifier, if item contexts are not more or less randomly distributed by activity and identity, the tool would revise the item set to more equitably distribute representational parameters.

3.11 REFERENCES

Attali, Y. (2018). Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10947, 17-29.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment, 4*(3).

Au, W. (2011). Teaching under the new Taylorism: High-stakes testing and the standardization of the 21st century curriculum. *Journal of Curriculum Studies, 43*(1), 25-45.

Arendasy, Martin E., & Sommer, Markus. (2012). Using automatic item generation to meet the increasing item demands of high-stakes educational and occupational assessment. *Learning and Individual Differences, 22*(1), 112-117.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems* (pp. 4349-4357).

Cheryan, S., Plaut, V. C., Davies, P. G., & Steele, C. M. (2009). Ambient belonging: How stereotypical cues impact gender participation in computer science. *Journal of Personality and Social Psychology, 97*, 1045–1060.

Cundiff, J. L., Vescio, T. K., Loken, E., & Lo, L. (2013). Do gender–science stereotypes predict science identification and science career aspirations among undergraduate science majors? *Social Psychology of Education, 16*(4), 541–554.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences, *115*(16), E3635-E3644.

Gayles, J. (2011). Attracting and retaining women in STEM (New directions for institutional research; no. 152). San Francisco: Jossey-Bass.

Gierl, M. J., & Lai, H. (2012). The role of item models in automatic item generation. *International Journal of Testing, 12*(3), 273-298.

Gierl, M., Lai, H., & Chen, Y. (2018). Using automatic item generation to create solutions and rationales for computerized formative testing. *Applied Psychological Measurement, 42*(1), 42-57.

Glick, P., & Fiske, S. T. (2001). An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. American Psychologist, 56(2), 109–118.

Gorsky, G. S., Wang, N., & Phelps, D. Rethinking consequential validity: Linguistic deconstruction of 'gold-standard' math story problems. In review.

Green, R., & Griffore, R. (1980). The impact of standardized testing on minority students. *The Journal of Negro Education, 49*(3), 238-252.

Kan, A., & Bulut, O. (2014). Examining the Relationship between Gender DIF and Language Complexity in Mathematics Assessments. *International Journal of Testing, 14*(3), 245-264.

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of Advertising, 46*(1), 141-155.

Kerr, B., & Multon, K. (2015). The development of gender identity, gender roles, and gender relations in gifted students. *Journal of Counseling & Development, 93*(2), 183-191.

Lai, A., & Tetreault, J. (2018). Discourse coherence in the wild: a dataset, evaluation and methods. arXiv preprint arXiv:1805.04993.

Lan, M., & Li, Min. (2014). Exploring gender differential item functioning (DIF) on eighth grade mathematics items for the United States and Taiwan. Seattle: University of Washington.

Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of Mathematics Items Associated with Gender DIF. *International Journal of Testing, 4*(2), 115-136.

Ling, W., Yogatama, D., Dyer, C., & Blunsom, P. (2017). Program induction by rationale generation: Learning to solve and explain algebraic word problems. In Proc. ACL.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Lomax, R., West, M., Harmon, M., Viator, K., & Madaus, G. (1995). The Impact of Mandated Standardized Testing on Minority Students. *The Journal of Negro Education, 64*(2), 171-185.

Luckin, R. (2018). *Machine learning and human intelligence: The future of education for the 21st century*. London: UCL Institute of Education Press.

Ma, A. (2017). The Application of Machine Learning to Education. *Journal of Student Science and Technology, 10*(1).

Madaus, G. F., & Clarke, M. (2001). *The adverse impact of high stakes testing on minority students: Evidence from 100 years of test data.*

Martin, C. & Ruble, D. (2004). Children's search for gender cues. Cognitive Perspectives on gender development. *Current Directions in Psychological Science, 13*(2), 67-70.

Mendes-Barnett, S., & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education, 19*(4), 289-304.

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

Nadeem, F., & Ostendorf, M. (2017). Language based mapping of science assessment items to skills. In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications (pp. 319-326).

Nadeem, F., & Ostendorf, M. (2018). Estimating linguistic complexity for science texts. In Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 45-55).

Napolitano, D., Sheehan, K., & Mundkowsky, R. (2015). Online readability and text complexity analysis with TextEvaluator. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (pp. 96-100).

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math = male, me = female, therefore math ≠ me. *Journal of Personality and Social Psychology, 83*(1), 44–59.

Parrott, L., Spatig, L., Kusimo, P. S., Carter, C. C., & Keyes, M. (2000). Troubled waters: Where multiple streams of inequality converge in the math and science experiences of nonprivileged girls. *Journal of Women and Minorities in Science and Engineering*, *6*(1).

Powers, D., Burstein, J., Chodorow, M., Fowles, M., & Kukich, K. (2002). Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research, 26*(4), 407-425.

Prentice, D. A., & Carranza, E. (2002). What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of Women Quarterly, 26*(4), 269–281.

Ryan, K. E., & Fan, M. (1996). Examining Gender DIF on a Multiple-Choice Test of Mathematics: A Confirmatory Approach. *Educational Measurement: Issues and Practice, 15*(4), 15-20, 38.

Sap, M., Prasettio, M. C., Holtzman, A., Rashkin, H., & Choi, Y. (2017, September). Connotation frames of power and agency in modern films. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 2329-2334).

Schmader, T. (2002). Gender Identification Moderates Stereotype Threat Effects on Women's Math Performance. *Journal of Experimental Social Psychology*, 38(2), 194–201.

Sekaquaptewa, D., & Thompson, M. (2002). Solo status, stereotype threat, and performance expectancies: Their effects on women's performance. *Journal of Experimental Social Psychology*, 39, 68–74.

Settles, I., O'Connor, R., & Yap, S. (2016). Climate perceptions and identity interference among undergraduate women in stem: The protective role of gender identity. *Psychology of Women Quarterly, 40*(4), 488-503.

Sheehan, K. M., Flor, M., & Napolitano, D. (2013). A two-stage approach for generating unbiased estimates of text complexity. In Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility (pp. 49-58)

Singley, M., & Bennett, R. E. (1998). Validation and extension of the mathematical expression response type: Applications of schema theory to automatic scoring and item generation in mathematics (GRE Board professional report; no. 93-24P). Princeton, NJ: Educational Testing Service.

Solano-Flores, G., & Nelson & Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 38(5), 553-573.

Spikol, D., Avramides, K., Cukurova, M., Vogel, B., Luckin, R., Ruffaldi, E., & Mavrikis, M. (2016). Exploring the interplay between human and machine annotated multimodal learning analytics in hands-on STEM activities. Proceedings of the Sixth International Conference on Learning Analytics & Amp Knowledge; 522-523.

Srihari, S., Collins, J., Srihari, R., Srinivasan, H., Shetty, S., & Brutt-Griffler, J. (2008). Automatic scoring of short handwritten essays in reading comprehension tests. *Artificial Intelligence, 172*(2), 300-324.

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.

Steffens, M., Jelenec, P., & Noack, P. (2010). On the leaky math pipeline: Comparing implicit math-gender stereotypes and math withdrawal in female and male children and adolescents. *Journal of Educational Psychology, 102*(4), 947-963.

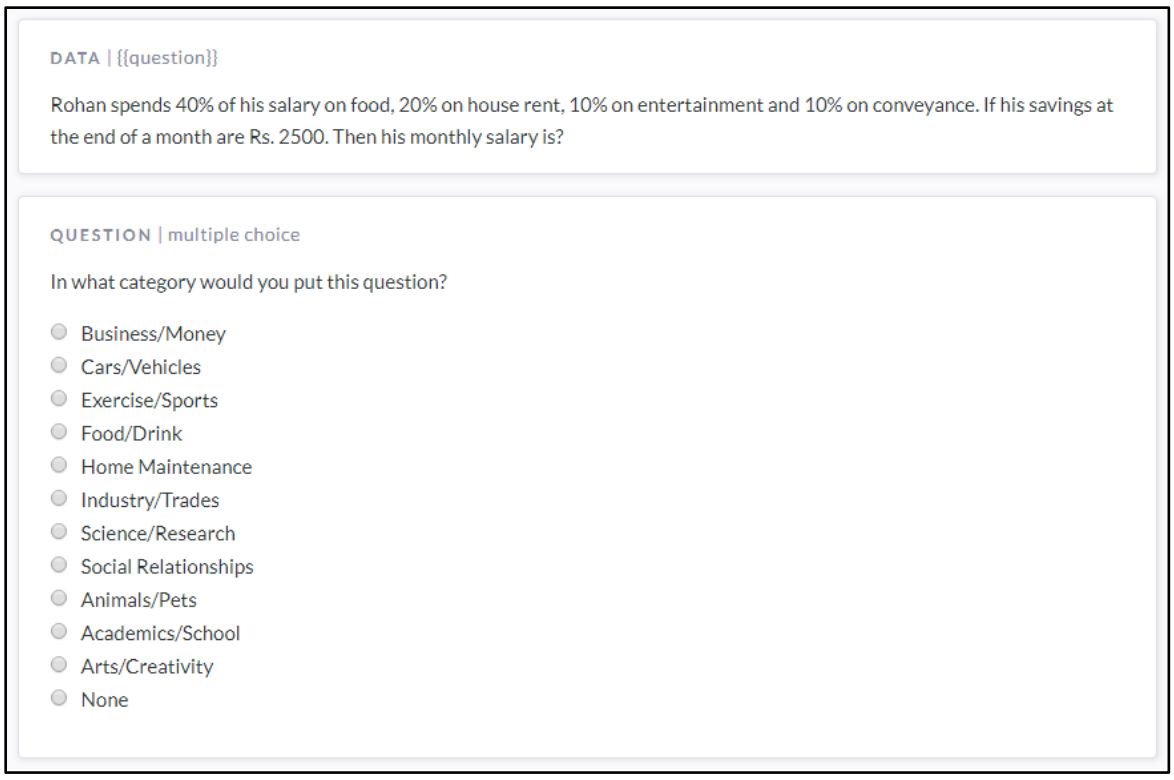
Taylor, C., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education, 25*(3), 246-280.

Thomas, K. & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior, 77*, 184-197.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1480-1489).

3.12 APPENDIX 3 A

Crowdsource Task Example Question with Activity Categories for Selection



3.13 APPENDIX 3 B

AQuA Training Item Classes

Category (n)	Fem-inine	Mas-culine	Neutral-Self	Neutral-Other	Non-Person	No PN Used
All Items (2135)	4.5%	13.0%	2.6%	12.6%	15.1%	52.2%
Academics/School (69)	2.9%	5.8%	0.0%	14.5%	10.1%	66.7%
Animals/Pets (7)	0.0%	14.3%	0.0%	14.3%	0.0%	71.4%
Arts/Creativity (4)	0.0%	25.0%	0.0%	50.0%	25.0%	0.0%
Business/Money (362)	6.1%	18.5%	1.7%	14.4%	9.9%	49.4%
Cars/Vehicles (57)	0.0%	17.5%	1.8%	24.6%	15.8%	40.4%
Exercise/Sports (39)	12.8%	15.4%	5.1%	12.8%	7.7%	46.2%
Food/Drink (34)	2.9%	14.7%	0.0%	11.8%	20.6%	50.0%
Home Maintenance (6)	0.0%	0.0%	0.0%	0.0%	16.7%	83.3%
Industry/Trades (47)	2.1%	10.6%	0.0%	21.3%	31.9%	34.0%
Science/Research (127)	1.6%	13.4%	3.9%	10.2%	15.7%	55.1%
Social Relationships (85)	8.2%	5.9%	5.9%	8.2%	4.7%	67.1%
None (1298)	4.4%	12.1%	2.8%	11.6%	16.9%	52.2%

NAEP Benchmark Item Classes

Category (n)	Fem-inine	Mas-culine	Neutral-Self	Neutral-Other	Non-Person	No PN Used
All (444)	5.2%	5.2%	0.7%	10.1%	8.8%	70.0%
Academics/School (26)	7.7%	11.5%	3.8%	23.1%	15.4%	38.5%
Animals/Pets (9)	11.1%	11.1%	0.0%	11.1%	11.1%	55.6%
Arts/Creativity (3)	33.3%	33.3%	0.0%	0.0%	0.0%	33.3%
Business/Money (46)	17.4%	13.0%	2.2%	6.5%	15.2%	45.7%
Cars/Vehicles (16)	0.0%	0.0%	0.0%	6.3%	25.0%	68.8%
Exercise/Sports (10)	20.0%	10.0%	0.0%	20.0%	0.0%	50.0%
Food/Drink (20)	10.0%	10.0%	0.0%	15.0%	15.0%	50.0%
Home Maintenance (8)	0.0%	25.0%	0.0%	0.0%	0.0%	75.0%
Industry/Trades (4)	0.0%	0.0%	0.0%	25.0%	0.0%	75.0%
Science/Research (16)	0.0%	6.3%	0.0%	12.5%	0.0%	81.3%
Social Relationships (5)	0.0%	0.0%	0.0%	20.0%	20.0%	60.0%
None (281)	2.5%	2.1%	0.4%	8.9%	6.4%	79.7%

3.14 APPENDIX 3 C

AQuA Item Examples

Question
Rohan spends 40% of his salary on food, 20% on house rent, 10% on entertainment and 10% on conveyance. If his savings at the end of a month are Rs. 2500. Then his monthly salary is?
On a certain transatlantic crossing, 20 percent of a ship's passengers held round-trip tickets and also took their cars abroad the ship. If 40 percent of the passengers with round-trip tickets did not take their cars abroad the ship, what percent of the ship's passengers held round-trip tickets?
A basketball coach is making the starting lineup for his team of 10 players. There are 4 guards on the team and 6 post players. How many different groups of 5 players could he choose assuming he must pick 2 guards and 3 post players?
John's Ice Cream Shop sells ice cream at m cents a scoop. For an additional n cents, a customer can add 2 toppings to his or her sundae. How much would a sundae with 2 scoops and 2 toppings cost, in terms of m and n?
Caleb and Kyle built completed the construction of a shed in 10 and half days. If they were to work separately, how long will it take each for each of them to build the shed, if it will take Caleb 2 day earlier than Kyle?
12 welders work at a constant rate they complete an order in 8 days. If after the first day, 9 welders start to work on the other project, how many more days the remaining welders will need to complete the rest of the order?
A certain galaxy is known to comprise approximately 2x 10^11 stars. Of every 50 million of these stars, one is larger in mass than our sun. Approximately how many stars in this galaxy are larger than the sun?
Three years ago the average age of a family of six members was 19 years. A boy have been born, the average age of the family is the same today. What is the age of the boy?
There are 40 balls which are red, blue or green. If 15 balls are green and the sum of red balls and green balls is less than 25, at most how many red balls are there?
A Tiger walks an average of 500 meters in 12 minutes. A Horse walks 20% less distance at the same time on the average. Assuming the horse walks at her regular rate, what is its speed in km/h?

NAEP Item Examples

Question
There are 50 hamburgers to serve 38 children. If each child is to have at least one hamburger, at most how many of the children can have more than one? A. 6 B. 12 C. 26 D. 38
Jill needs to earn $45.00 for a class trip. She earns $2.00 each day on Mondays, Tuesdays, and Wednesdays, and $3.00 each day on Thursdays, Fridays, and Saturdays. She does not work on Sundays. How many weeks will it take her to earn $45.00? Answer:_____
Amanda wants to paint each face of a cube a different color. How many colors will she need? A. Three B. Four C. Six D. Eight
A club needs to sell 625 tickets. If it has already sold 184 tickets to adults and 80 tickets to children, how many more does it need to sell? Answer: _____
Marty has 6 red pencils, 4 green pencils, and 5 blue pencils. If he picks out one pencil without looking, what is the probability that the pencil he picks will be green? A. 1 out of 3 B. 1 out of 4 C. 1 out of 15 D. 4 out of 15
A plumber charges customers $48 for each hour worked plus an additional $9 for travel. If h represents the number of hours worked, which of the following expressions could be used to calculate the plumber's total charge in dollars? A. $48 + 9 + h$ B. $48 \times 9 \times h$ C. $48 + (9 \times h)$ D. $(48 \times 9) + h$ E. $(48 \times h) + 9$

Chapter 4. ‘TEACHING TO THE TEST’ REIMAGINED: A NOVEL APPROACH TO CONFRONTING STEREOTYPES IN MATH STORY PROBLEMS

4.1 ABSTRACT

This paper had two aims: expand current “best practice” in test equity and identify teacher perspectives on gender bias in math, math assessment, and math instruction. At the core of this research is a deep concern for the experiences of teachers and students who are the final users of test materials. I offer guidelines and recommendations for writing high-quality items, as well as for evaluating and modifying existing tests and other instructional materials, with a focus on minimizing biased patterns of gender representation (given that eliminating it altogether is close to impossible), integrating intersectional character representation, and reducing potential harm to all students, especially those who are already the most vulnerable. The long-term goal is to begin to articulate a culturally-adaptive approach to teaching test literacy and identify ways to facilitate questions and conversations with students about the messages we receive from a test-focused educational culture, as well as from the test content itself.

4.2 INTRODUCTION

Researchers, educators, and test developers have long known that unintended negative (or positive) consequences pose a real threat to the validity of large- and even small-scale assessment, but have struggled to articulate, identify, and verify the mechanisms through which

this impact occurs. While procedures are in place for minimizing the amount of explicit bias and group-based disparagement in test questions, there is little transparency regarding the non-statistical procedures utilized towards this end (LACCD, 2012; NCES, 2017; WA OSPI, 2011; Smarter Balanced, 2012). Further, beyond this stage, test/item 'fairness' becomes a matter of probabilities through the focus on statistical approaches to fairness (Camilli, & Shepard, 1994; Zumbo, 2007; Zwick, 2012. In order to attempt to address this hole in our understanding of how the process of being tested impacts students academically as well as internally and interpersonally, I developed a novel methodology for reviewing and cataloguing math story problem texts for the existence of harmful gender stereotypes.

In my preliminary research, 120 8th grade math story problems drawn from the item bank of the National Assessment of Educational Progress (NAEP), a nationally representative standardized test, were reviewed using a coding rubric I developed to catalogue items in two ways: 1) identify how item contexts reproduce stereotypes about gender roles, family structure, and sexual orientation, and 2) reveal patterns of associations between gender, family, or sexuality and behavior across the sample of items, as a 3-part collaborative coding project (Gorsky, Wang, & Phelps, 2019, in review). Coding revealed pervasive bias across the items, from more obvious behavioral associations (women and femmes are consistently featured in contexts involving food preparation) to more concerning schematic trends (masculine people are described using more words that convey agency and self-determination). Indeed, the collaborative process itself was found to be critical to developing an understanding of how bias gets inscribed into tests and reinforced at each stage of test development. In short: representation matters.

That work was predicated on a Bias-Impact Framework (BIF) (see graphic model below) that examines bias exposure, reinforcement (of biased ideologies), and identity development. Within this framework, students are exposed to test items and other educational materials that reinforce gender-biased, gender-stereotyping, and microaggressive gender characterizations (represented as a small subset of the myriad ways young people are exposed to gender stereotypes, in general). Such messages are reinforced through repeated exposure to these problematic items via in-school and out-of-school test preparation and administration. Unsurprisingly, little if any time is given in typical test preparation or math instruction to

Figure 4.1

Graphical Model Representing the Bias Impact Framework (BIF)

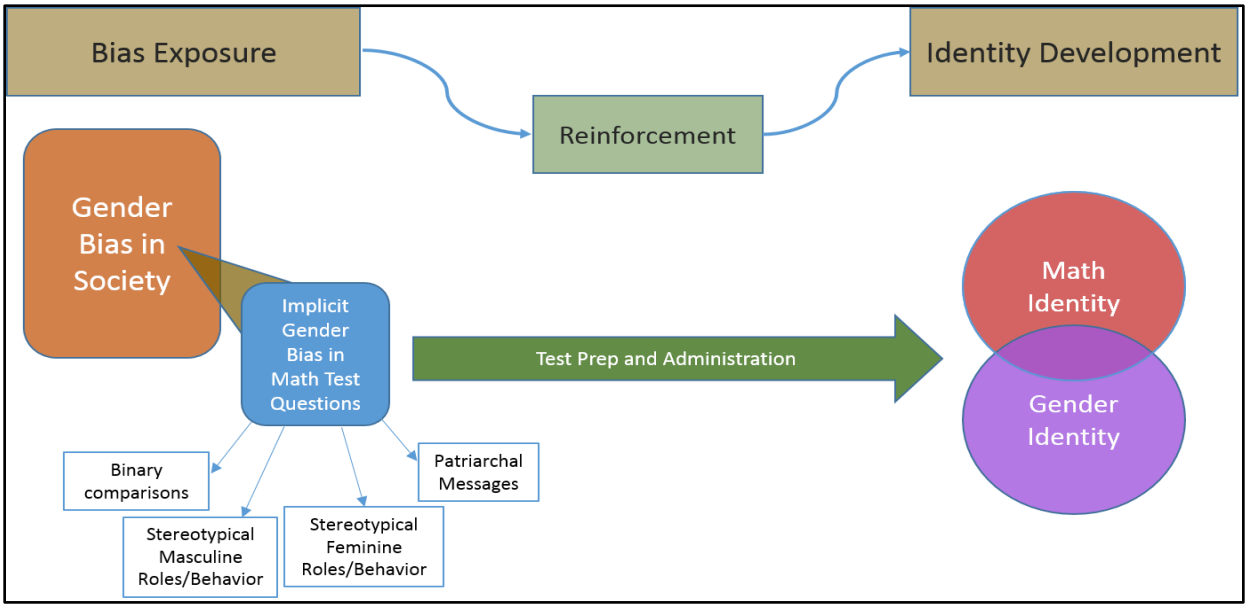


Figure 1. The BIF graphic shows the way item themes are reinforced and integrated into identity.

exposing and unpacking the alienating and destructive social messages being sent and reiterated week after week, quarter after quarter, year after year in test-related curriculum and print materials and how this is ultimately harmful to students (Good et al., 2010;Hamdan, 2010; Osaďan, Belešová, & Szentesiová, 2018; Woyshner, 2006). Over time, these ideas become

internalized and inform students' identity development, as well as how they perceive others, both in terms of gender identity, math identity, and their intersection. The long-term goal of this research is to explore this process, better understand its mechanisms, and uncover ways to intervene in the harmful cycle of misrepresentation.

This paper offers a framework and approach for identifying and addressing biased patterns of gender representation, especially by integrating more dynamic character representation, in tests and other instructional materials. The long-term goal of this work is to begin to articulate a culturally-situated approach to teaching test literacy and identify ways to facilitate questions and conversations with students about the messages we receive from a test-focused educational culture, as well as from the test content itself.

4.3 UNPACKING *BIAS*, *FAIRNESS*, AND *SENSITIVITY*

When it comes to student tests, the term ‘bias’ takes on two distinct, but equally important meanings. When statisticians or test developers talk about *bias,* we are often talking about *measurement bias,* defined as systematic error in how an instrument measures a trait in a specific group (Camilli and Shepherd, 1994). This means that one group (i.e., students with limited English proficiency) consistently underperform relative to another group (native English speakers), *regardless of ability.* This last piece is really important. When there is systematic measurement bias, students with the same theoretical underlying skill level do not earn comparable test scores.

This is a huge problem and test developers know it. Most large-scale, state-wide or nationally-administered tests go through rigorous evaluation for measurement bias at the test and item level. There are a number of methods to test for this, from less rigorous, but more accessible Classical Test Theory (CTT) approaches that use item difficulty and true-score/observed score

differences to compare performance between groups (Camilli and Shepherd, 1994) to Item Response Theory (IRT) and Regression approaches that use probabilistic parametric model-building to estimate group differences at the item or test level (Camilli & Shepard, 1994; Zumbo, 2007; Zwick, 2012).

There is, however, another way we use the term 'bias,' even as measurement professionals. Often, when we say *bias,* what we really mean is *fairness,* which comes at the intersection between measurement bias (how our assessment systems differentially assess students from different groups) and social bias (the beliefs and ideas we have about people from those groups). That is, we are interested in the social impact of test administration, use of test scores, or of the test content itself. In this case, we are interested not only in how students will perform on tests, but also in how taking tests feels for kids and how it impacts them short and long term. Examples of statistical bias, social bias, and fairness are easily found in tests and other instruments across subjects and contexts, especially on the basis of group membership/identity (i.e., gender, race, social class, and disability, among others) (Abedi & Levine, 2013; Green & Griffore, 1980, McGraw, Lubienski, & Strutchens, 2006).

While attention is given by test developers to both types of bias described above, the way items are evaluated for social bias is quite interesting from an equity perspective. Items created for large-scale, high-stakes administration are typically put through what is called a "sensitivity review." The goal of a bias and sensitivity review as part of the test development process is to consider the impact test content may have on a variety of examinees given their unique experiences and identities. Typically this begins with identification of groups that "require special consideration" (WA OSPI, 2011). This includes students who are non-white, impoverished, disabled, and for whom English is not the first language. In addition, gender,

religion, age, and other factors such as gifted placement are considered. Item reviewers are taught about bias and sensitivity research and issues related to test administration and scoring.

Math is commonly (and incorrectly) perceived as objective, or somehow value neutral (Moses & Cobb, 2001), so when we study math test performance, especially when it comes to measurement bias, we are interested in how much of the performance can be explained by the student's true ability and how much can be explained by something else: student identity, a feature of the question text, an inconsistency in administration, guessing, etc. (Helms et al., 2009; Kunnan, 2007; Sacket, Borneman, & Connelly, 2009). At the core, bias and fairness issues represent a threat to the validity of the use of a given test for its intended purposes (Caines, Bridglall, & Chatterji, 2014; Chatterji, 2013; Kane, 2010). Here, however, I'd like us to consider the idea that in order to truly address the problematic item contexts we encounter in math test materials we interact with as educators and researchers, we may need to call the purposes of assessment into question.

4.4 Teacher Reflections on Math Testing and Representation

Through online social networks for math educators, I was able to get anonymous survey feedback from a small group of teachers regarding test preparation, test question themes, and, most importantly, representations of identities. One taught basic math (arithmetic, fractions, and number systems) at the primary (P-5) level. Another taught Algebra, Geometry, Pre-Calculus, Statistics, and Business Math at the middle and high School levels. The third, a "veteran" teacher with more than 20 years of experience, taught Algebra, Geometry, Advanced Algebra, Trigonometry, and Pre-Calculus at the high school level. Two identified as white and one identified as white and Hispanic. All three were women. One indicated being a parent and another identified as Jewish.

I asked them if they had received any pre-service instruction on assessment, generally, or math-specific, in terms of writing, revising, or engaging test content on a deeper level. None of them had. One indicated that math was "more or less an *afterthought* in" her program (emphasis added). When asked what they would have liked to have seen, one teacher reported a distinct lack of assessment instruction and stated that she would have like to learn "anything at all" and that she felt like she was "flying blind." Another, who went through an alternative certification process and didn't attend a formal teacher preparation program reported that some content on what makes a good question "would have been a good topic of debate." She wanted to take "critical looks at various problems" and "presentation styles" to have a discussion about what a good question can and should look like.

In addition, they were asked about their main critiques of standardized, high-stakes testing and the challenges they face in making math accessible to students. There were three main themes in their responses: content-irrelevant factors such as the timed component of the test, making it accessible and relevant to a range of student abilities and identities, and that it assesses "test-taking ability" (or other non-math skills like reading comprehension) in addition to math ability. These are consistent with the challenges identified by other researchers (Abedi & Levine, 2013; Au, 2007). One teacher made an important and unfamiliar observation that she often finds that "students who make insightful comments in class and clearly demonstrate a solid or even advanced understanding of the concepts do not perform well on assessments." This was supported by another comment made about the type of skill assessed in typical tests ("content vs. ability to critique, find patterns, verbalize observations, etc." These teachers' responses demonstrate the need for both more guidance on making assessment meaningful and

accessible to students, as well as the fact that they are well-attuned to the conflicting expectations of math tests and students' best interest in terms of learning critical math content.

The teachers were next asked how they felt their own identities informed their teaching. The responses to these two questions really highlighted the disconnect between what teachers want to offer their students and what they have available to them. Two teachers described themselves as "good at math" and described the ways that impacts their teaching. The veteran high school teachers h described her "high expectations for students" and admitted that she "sometimes find[s] it hard to understand where a student isn't making connections." However, in describing her role as a parent, she also indicated her skepticism of "drill-and-kill" homework focused on quantity over depth. This has had a direct impact on her instructional approach. She reported giving increasingly less homework over the years, sticking to the "oldies but goodies" and using "deeper exploratory exercises" to compensate for narrower curriculum in other math classes. Further, this teacher has experience teaching outside of the United States and expressed that she is "critical of the math we teach in the U.S." and "dumb-down" concepts and make assumptions about what should be presented and when.

Regarding colleagues, a different teacher reported feeling alienated and "irritated" that her women coworkers "complain about hating/not being good at math." This suggests a distinct lack of tangible professional supports for engaging math education in a collaborative way across subjects. In terms of the salience student identities, one teacher's focus was clearly on gender: "I think it's important to encourage girls to speak up, to argue their perspective, to disagree with classmates." The veteran reported being in an extremely affluent white district with 99% high school graduation and 96% acceptance to a 2- or 4-year college. Testing plays a significant role in her instructional approach ("I sometimes gear instruction toward the standardized tests

students will have to take, and definitely emphasize concepts I know are tested on the SAT or ACT") and the population means she can make certain assumptions like that "students can afford school supplies" for a math class (ex. a graphing calculator) and that "nearly 100% do [their homework] on any given day." These are important assumptions that allow for a certain approach to teaching and test prep that this teacher acknowledges is not the norm. Interestingly, one of the respondents elected not to respond to the two questions asked about identity salience, indicating instead that she didn't understand the questions. It is unclear which part was unclear for her, but the specific omission raises interesting question about how teachers are(n't) supported in thinking through their own and students' identities and the ways it informs their teaching practice.

Finally, I asked them what recommendations they would make to test developers to make math testing more accessible and what their favorite part of teaching math is. For test developers, the teachers asked for more student agency ("giving choices of problems to do so students could choose problems that they can relate to more") and more clarity in presentation ("I hate it when I feel that a test question is designed to trip up students instead of finding out what they understand"), as well wanting more relevance in contextualization ("Make them applicable in real-life. Do not talk about pies and cookies.") These themes demonstrate the need for a new approach to test development and preparation that captures the nuance of student ability and experience.

The things that made teaching math fun for these educators deserve a place in assessment: "having students see that math is really around the, and in places they would not have ordinarily considered to have math," "seeing the excitement of students when they have struggled hard and finally figure out something difficult," and "teaching the whole child." These

things aren't incompatible with what assessment *could* be, but they are rarely addressed directly in test development and in assessment research assessment (Kan & Bulut, 2014; Lan & Li, 2014; Li, Cohen, & Ibarra, 2004; Mendes-Barnett & Ercikan, 2006; Ryan & Fan, 2996; Taylor & Lee, 2012). While the landscape of testing may be slow to adapt to the evolving needs of students and teachers, there are still sites for intervention. In the next section, I begin to lay out a framework for deconstructing mass assessment context and provide some procedures for working with test questions that address many of the concerns expressed by these teachers and supported by research on test equity (Abedi & Levine, 2013; Abedi, Hoffstetter, & Baker, 2001; Au, 2007, 2011).

4.5 WHAT DO I MEAN BY 'PROBLEMATIC ITEM CONTEXTS?'

There are a number of ways a math word problem context could be considered problematic from an equity perspective. In formal bias and sensitivity procedures, reviewers are typically looking for things like specialized language that might only be accessible to one group, figurative language and idioms, social group stereotypes, or any other facets of items which might privilege, disadvantage, "offend, upset, or otherwise distract" a subgroup of test takers (SBAC, 2012). Considering the fact that including culturally relevant material (one main goal of equity standards) for a particular student or group of students may simultaneously represent a sensitive topic for another, the process of eliminating all "problematic item contexts" can be difficult.

Topics including "couples social dancing," "upsetting aspects of slavery," "pregnancy of human beings," "climate change caused by human behavior," and "ski trips" have been excluded from the Smarter Balanced test (Smarter Balanced, 2012; Strauss, 2015) and the list of topics to avoid across different assessment guidelines is extensive. While the reasons for these exclusions are often somewhat apparent, the cost of excluding these topics is also concerning. At what point

does trying to sanitize the questions start to compromise the realism of the material included? To what extent should personal politics (i.e., in the case of climate change) dictate which perspectives are "protected" and considered in the item review process? Do controversial issues have any place in assessment? Who is served by these decisions in the short term? Long term?

Beyond taking a first pass at items to identify particularly glaring themes, contexts, or representations, these more nuanced questions can help frame the goals of critical item review. From my perspective, the aim of this particular exercise is less to completely "neutralize" a question and is more about reflecting on what is gained and what is lost through the addition or subtraction of themes and representations in the test content. While creating items that are accessible to a range of students is a primary goal of industry test developers (Abedi, Hoffstetter, & Baker, 2001; NAGB, 2010, 2014; NCES, 2017), classroom teachers have the advantage of getting to create tests and engage in assessment critique without the burden of rigorous item vetting. Whether test preparation is a large or small component of a given curriculum, there are opportunities to strike a balance between these two objectives.

Perhaps the greatest limitation of the formal bias and sensitivity review is the reviewer pool. While bias and sensitivity review guidelines do require that a variety of people write and evaluate items, one critical group is missing: students. The fact that students are not invited to participate in the item evaluation process represents a missed opportunity to provide them with the sort of agency and authority to define and determine their own needs when it comes to test equity. While I fully advocate for a more student-driven and -centered approach to item/test design and review from within the test development industry, the classroom is a critical site for accelerating students' understanding, interest, and engagement with the tests they take. Students' perspectives on the structure and content of tests provide invaluable feedback for producing

assessments that are more culturally relevant, while still remaining accessible to a broad population of test-takers (Abedi & Levine, 2013; Brown, 2011, Hendy et al., 2014).

4.6 WHAT TO LOOK FOR…

Below are some of the questions my collaborators and I asked when doing preliminary coding of items for gender representation in our earlier work. They can be used as they are, or easily modified to extend to other contexts, features of identity, and grade levels.

- How many people are featured in the item?
- Is gender explicit (use of gender pronouns she/her/hers, he/him/his) or implied (name, activity)?
- What is the character in the item doing?
    - Ex. preparing food, caring for a pet, doing a science experiment, building a shed, running in a race, solving a math problem
    - Start to make note of patterns in association between gender and activity
- What kinds of words are used to describe these activities?
    - Are the contexts presented as positive or negative?
    - Are characters competing?
    - How much detail is included? How much is relevant to the math skill assessed?
- Does the context feel believable? Does it seem like something a real person would do or does it feel artificial?
- What messages about school, gender roles, behavior, or other aspects of society or identity are sent by the item?

These questions are meant to serve as a "first cut" for items and are by no means an exhaustive list of ways any given social identity might be represented in a test. There may be many items flagged by this procedure or relatively few. This is ok. Pay particular attention to the "I'm not sure" questions. If you have some uncertainty or uneasiness reading a question, there is a good chance another person would, too. The goal is not to become "perfect" at identifying every possible source of bias in an item. Every person brings their unique experience to the task of item review and so there will be things that are more salient to some and not to others. This is

also ok. What we want is to start developing a new lens through which we read and understand what contextual baggage test questions bring with them.

The goal of this exercise is to start to think about items in a new way. As patterns emerge, more complex questions can be raised:

1. Does the context of the item reinforce gender stereotypes?
2. Does the context of the item assume that gender identity predicts interest or behavior?
3. Does the context of the item favor one family structure over others?
4. Does the context of the item assume a common family structure for all people?
5. Does the context of the item favor one sexual/romantic orientation over others?
6. Does the context of the item assume a common sexual/romantic orientation for everyone?

The goal is *not* to become "perfect" at identifying every possible source of bias in an item. Every person brings their unique experience to the task of item review and so there will be things that are more salient to some and not to others. This is also ok. What we want is to start developing a new lens through which we read and understand what contextual baggage test questions bring with them.

4.7 …AND HOW TO ADDRESS IT

Whether you are wanting to apply these principles to test materials you've created or to content you did not create, my primary recommendation is to consider the process as having two components: 1) the individual item, and 2) the test as a whole (as applicable). We want to address the item as a stand-alone unit that carries contextual weight and which has the potential to have a lasting impact (positive or negative), while also acknowledging the ways items work together to reinforce certain ideas and the patterns of representation that might emerge only when all items included in an assessment are considered together. While it may not be feasible or even

necessary to apply a rigorous evaluatory process to all test items on all assessments, these principles are meant to form the basis of a reflexive process that can be engaged explicitly or more passively in a variety of academic contexts.

4.7.1 *Creating Your Own Test Materials*

If you want to create high-quality, accessible, culturally-relevant classroom tests, one of the first steps is to challenge yourself to think about test questions differently. On the one hand, test questions are a means to an end. Their purpose is to capture student learning, ability, skills, knowledge, etc. On the other hand, they exist as pieces of historically and politically contextualized written language. While the first premise characterizes test questions more as "objective" and "utilitarian," the second framing leaves room for the "tension 'between what [the item text] manifestly means to say and what it is nonetheless constrained to mean'" (Norris, 1987, p. 19, as cited in Rolfe, 2004, p. 274). Within an educational context, these tensions similarly play out between the intended purpose of an assessment item/tool and the impact it has on the test-taker, nonetheless on the broader collective consciousness surrounding academic accountability systems.

Figure 2. Guiding questions for writing original test items or refining existing items.

- ***Questions to ask when writing new or modifying existing test items***

1. What skill/knowledge is being assessed?
    a. Does this limit the structure of my item(s)?
2. What identities do I hold?
    a. How do those identities impact the topics and ideas I write into my test?
3. Does the context of the item(s) I am writing/reviewing feel realistic?
    a. Do they feature an activity that I (or someone I know) would actually do?

b. Will *all* my students know what I am talking about?
    i. If not, which students may or may not? If it seems like your items are not consistently accessible to your students' range of identities/experiences, you might consider applying step 6 earlier/more frequently in this process.

4. Do the items present a range of ways of being (ex. demographic identities, settings, behaviors, epistemologies/theories of learning)?

5. Taken as a whole, do the items collectively tell a certain "story"? That is, do patterns emerge between activity categories (ex. sports, food, or work) and the embedded characterizations (ex. identity, question type, assessed skill)?

6. Have you gotten peer feedback on any of the items? This could be from colleagues, friends, family, or partners.
    a. Are those people demographically similar to you?
        i. If so, I strongly encourage sharing at least a sample of your items with a trusted person who is different from you in at least one personally-salient way. This may be on the basis of race, gender, socioeconomic class, or some other life experience.

           **This is especially important if you hold one or more privileged identities such as Whiteness or affluence. It can be really difficult to recognize our own gaps in perception and awareness, both of ourselves and others, and this feedback will be pivotal in helping you develop these test equity "muscles," especially when you are first starting out.

Holding this two-pronged orientation, these seemingly-incompatible paradigms can, and *should*, be used in tandem to inform the test development process. Applied together, test questions become sites for both 'educational measurement' and historized and politicized inquiry. Rather than simply reproducing existing norms and protocols, we can use test development as an opportunity to critically reflect on the biases and assumptions we are bringing to the table. The equity-driven item-writing steps provided below offer a starting point for developing and/or strengthening a reflective test development/evaluation practice.

4.7.2 *Refining Test Materials You Did Not Create*

Modifying existing items is another way of confronting implicit bias and problematic representations in test materials. Depending on the source, this may be a labor-intensive undertaking. While already-vetted items like those publically available from NAEP or another testing agency may still have problematic patterns of representation (as my earlier work demonstrated), overall, it is unlikely that these items will contain egregious examples of stereotypes and concerning generalizations. However, other academic materials that are not put through as rigorous an evaluation process can and do often include contexts and representations that are heavily culturally loaded at best and outright offensive at worst (Good et al., 2010; Hamdan, 2010; Osaďan, Belešová, & Szentesiová, 2018; Woyshner, 2006).

The above steps apply to both creating and modifying test content to make it align more closely with an "equity and justice" model of assessment. That said, when modifying existing items, there are a couple additional factors to take into consideration:

- ✓ If you are adapting content from an official testing agency such as College Board, ETS, or Smarter Balanced, will the modifications fundamentally change what the question is asking/measuring?
    - **Ex.** If I change "John spent 30 minutes lifting weights on Monday, Wednesday, and Friday..." to "Akedi spent 30 minutes practicing ballet on Monday, Wednesday, and Friday...," the structure and skills assessed are mostly unchanging. If, however, I were to rewrite the question to "Akedi and his husband spent 30 minutes doing a puzzle on Monday, Wednesday, and Friday…," aside from the relative culturally weight of the non-Eurocentric name and the homonormative representation of married gay men, the item has now increased in

word length and potential conceptual complexity.

This is not 'bad,' per se, but what is gained by adding a real-world representation of a denormalized identity should be considered carefully alongside what the goals of the activity is, specifically, as well as of the assessment, more broadly.

4.7.3 *Engaging Students in Guided and Independent Test Critique*

Perhaps the most important recommendation I have for effectively mitigating the effects of problematic test content is to *talk with students* (Amit & Fried, 2005; Gunderson et al., 2012). It may seem obvious, but students are given relatively few opportunities to learn about (nonetheless participate in) the test development process. High school and even undergraduate students I work with are often surprised to learn about why we test, how tests are made, and what scores mean. Many have never thought about this before at all. But this specific knowledge about test creation is a critical component of holistic "assessment literacy" and overlooking its importance represents a failure to fully equip students with the intellectual tools they need to attain mastery of the test-taking process.

One of the easiest ways to engage students about test items is to embed this inquiry about representation into a lesson or assignment centered on the structure of test questions and strategies for answering different types of questions. Through discussing the format of an item, teachers can introduce and model the item development process and get students thinking about what kind of planning goes into it and what factors a test writer might consider. While the activity modeled below is created around reviewing existing items, asking students to write their own items can be a great way to better understand a) how students internalize (or don't) existing test structures and are influenced by what they are used to seeing, b) what contexts, themes, and activities are most salient to them, and c) the potential sites for richer discussion of themes like

representation, equity, high-stakes testing, and what we mean when we talk about "achievement." In fact, the following procedures can easily be modified for discussing items students have written themselves, such as through a small-group peer review.

Figure 3. Student Item Review Guided Activity

- ✓ Here are some guiding prompts for looking at test questions with students one-on-one, in small groups, or as part of a whole-class discussion. Steps 1-3 can be repeated with any number of items and step 4 can easily be asked as a stand-alone question.

1. *First, I'm going to show you a math word problem. Read through it at your own pace and let me know when you are finished. There is no rush and you do not need to solve the problem, but you can try it if you want to. Either way, I'd like you to feel free to write, circle, underline, or make any other marks on the page.*

2. *Now, can you tell me anything you noticed about the person or "character"(s) in the question? If you need to read it through again, please do.*
    a. If student is reticent to respond, prompt using additional leads, such as
        i. "*What can you tell me about…* (insert item-relevant feature)?"
            - the person's gender?
            - the person's name?
            - what the person was doing?

        ii. "*What else did you notice about…* (insert item-relevant feature)?"
            - the kind of words used to describe the person? (i.e., positive/negative, supportive/critical)
            - the details included in the question? (i.e., relevant vs irrelevant to the embedded math problem, making the person seem more/less relatable)

3. *Is the person in the question someone you feel like you can relate to? Are they similar to you? Different?*
    a. *Yes? What about them?*
    b. *No? Why not?*

4. *Now think about other math tests you've taken. Do you feel like you can usually relate to the characters in math story problems?*
    a. *How do you feel about that? Is it important? Do you feel like it helps you solve the problem?*

4.8 Reimagining Representation

In 1985, an independent educational organization committee formed to address issues of fairness in K-12 testing. In 1988, this group, the National Center for Fair and Open Testing (aka FairTest), published a report in which they laid out the (negative) effects of high-stakes testing-driven reform, from the number of tests administered, to the flaws within tests, to the short and long term impacts of high-stakes tests (Medina & Neill, 1988). Medina and Neill's findings from more than three decades ago are actually rather un-shocking by contemporary testing standards. Millions of standardized tests were administered the year their data was collected. Big school districts were testing a lot and certain regions seemed to test more than others. They observed concerning flaws in test development that produced unreliable and biased tests, and an uncomfortable over-reliance on scores as singular measures of academic improvement and accountability. More unsettling still, the content and structure of standardized tests heavily influenced curriculum and how school time school was spent.

Medina and Neill (1988) rejected the idea that tests are objective and essential. Far from being objective, the authors claimed that tests routinely produced inaccurate or biased results that disadvantaged historically marginalized groups including women and girls, Black and brown people, and those living in poverty. They were concerned about how much power test developers and even the tests themselves had. Simply put, tests were creating more problems than they were solving. In terms of the specific problems Medina and Neill outlined regarding tests and test use, they articulate several specific issues with test construction, content, administration, and scoring: biased items, inadequate validation procedures, and failure to create a truly standardized experience. While demand for standardized test protocols has grown significantly over the last

30 years, the extent to which test developers, evaluators, and users have centered issues of equity and justice is more ambiguous and plenty of work is left to be done. Then and now, standardized tests are overwhelmingly written by and for the middle to upper-class white population (Dee, 2005; Gorski, 2012; Harris, Ravert, & Sullivan, 2017; Howard, 2010).

Fortunately, a number of the issues articulated by Medina and Neill (1988) have been addressed by test developers and measurement researchers, such as identifying ways of addressing multidimensionality and improved norm referencing procedures (Hidalgo-Montesinos &Gómez-Benito, 2003; Zwick, 2012). However, the more philosophical challenges remain: the ethics of using test scores to make student-, school-, or district-level decisions, the inherent biases that "experts" bring to the test construction and validation process, persistent gaps in group-level access to high-quality educational experiences. This work isn't finished and "business-as-usual" isn't getting us there.

In 1984, in an address to the NYU Institute for the Humanities Conference, Black feminist Audre Lorde said, "The master's tools will never dismantle the master's house." What she is saying here is that oppressive ideologies have no place within equity and justice work. Essentially, it is antithetical to maintain oppressive ideologies *within* the study of systemic gender bias and the negative social, emotional, and academic consequences thereof. Richer integration of feminist frameworks into educational frameworks offers an opportunity to resist the conflation of "equality of opportunity" and "equality of outcome" that continues to plague education and educational research. This has specific implications for measurement in education, which is almost entirely outcome-driven. One major shortcoming of the literature on gender equity in educational testing is that it tends to be descriptive rather than transformative. It

exposes the *impact* of gender bias, stereotypes, and discrimination on children's academic and social experiences, acting as a mirror of what *is* rather than a window to what *could be.*

What Lorde asks us to do is not easy. It requires a drastic shift away from that which we are all most accustomed, a complete overhaul and rethinking of the systems we rely on to keep the machine of society running. To adopt a truly revolutionary theory of educational measurement requires a deep interrogation of the ways that current standards and testing models are complicit in cultural reproduction. High-stakes testing remains a key mechanism of schooling to perpetuate class, race, and other social divisions within the academic and professional spaces. As Acker (1987) puts it, "the tension between education as reproductive and education as liberating is encountered daily" (p. 432). In the US and abroad, the "continued ideological hostility of central government" in regards to equity issues poses a real challenge for feminist education researchers and practitioners who face resistance to radical reframings, especially given the way high-stakes testing movements have drastically narrowed curriculum foci (Weiner, 1994). This can leave us feeling disempowered to enact meaningful change in education policy and practice at multiple levels of the academic knowledge production machine and further limits who is given voice and representation from the classroom to the boardroom.

However, we should not let this disenfranchisement hold us back from engaging in education as a moral endeavor. In math learning and assessment specifically, radical reframing can be enacted in a number of ways. It might include representing a range of international math and number systems in curriculum and in test content, allowing young people to define their educational goals and ways of demonstrating mastery, decentering oppressive and overrepresented histories and methods, and a readiness to reject some of the deeply ingrained assumptions that drive our research practices. This requires rethinking the very purposes of

assessment and standards-setting and destabilizing existing systems of accountability that are inextricable from capitalist projects driven by privatization, monetization, and standardization. This is hard work, but we can do it. Together.

4.9 REFERENCES

Abedi, J., & Levine, H. G. (2013). Fairness in assessment of English learners. *Leadership, 42*(3), 26-38.

Abedi, J., Hofstetter, C., & Baker, E. (2001). NAEP math performance and test accommodations: Interactions with student language background. CSE Technical Report 536. CRESST and NCES.

Amit, Miriam, & Fried, Michael N. (2005). Authority and authority relations in mathematics education: A view from an 8th grade classroom. *Educational Studies in Mathematics, 58*(2), 145-168.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258-267.

Au, W. (2011). Teaching under the new Taylorism: High-stakes testing and the standardization of the 21st century curriculum. *Journal of Curriculum Studies, 43*(1), 25-45.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Measurement methods for the social sciences series; 4). Thousand Oaks [Calif.]: Sage Publications.

Caines, J.L., Bridglall, B., & Chatterji, M. (2014). Understanding validity and fairness issues in high-stakes individual testing situations. *Quality Assurance in Education, 22*(1), 5-18.

Chatterji, M. (Ed.). (2013). *Validity and test use: An international dialogue on educational assessment, accountability and equity* (First edition. ed.). Bingley United Kingdom: Emerald.

Dee, T. (2005). A teacher like me: Does race, ethnicity, or gender matter? *The American Economic Review, 95*(2), 158-165.

Good, J., Woodzicka, J., & Wingfield, L. (2010). The effects of gender stereotypic and counter-stereotypic textbook images on science performance. *The Journal of Social Psychology, 150*(2), 132-47.

Green, R., & Griffore, R. (1980). The impact of standardized testing on minority students. *The Journal of Negro Education, 49*(3), 238-252.

Gunderson, E., Ramirez, A., Levine, G., & Beilock, S. (2012). The role of parents and teachers in the development of gender-related math attitudes. *Sex Roles, 66*(3), 153-166.

Hamdan, S. (2010). English-language textbooks reflect gender bias: A case study in Jordan. *Advances in Gender and Education*, 2, 22-26.

Helms, J., Sackett, P., Borneman, M., & Connelly, B. (2009). Defense of tests prevents objective consideration of validity and fairness/responses to issues raised about validity, bias, and fairness in high-stakes testing. *The American Psychologist, 64*(4), 283.

Harris, B., Ravert, R. D., & Sullivan, A. L. (2017). Adolescent racial identity: self-identification of multiple and "other" race/ethnicities. *Urban Education, 52*(6), 775-794.

Hidalgo-Montesinos, M. D., & Gómez-Benito, J. (2003). Test purification and the evaluation of differential item functioning with multinomial logistic regression. *European Journal of Psychological Assessment, 19*(1), 1–11.

Howard, T. (2010). *Why race and culture matter in schools: Closing the achievement gap in America's classrooms* (Multicultural education series (New York, N.Y.)). New York, N.Y.: Teachers College Press.

Kan, A., & Bulut, O. (2014). Examining the relationship between gender DIF and language complexity in mathematics assessments. *International Journal of Testing, 14*(3), 245-264.

Kane, Michael. (2010). Validity and fairness. *Language Testing*, 27(2), 177-182.

Kunnan, A. (2007). Test fairness and DIF. *Language Assessment Quarterly*, 4(2), 109-112.

Lan, M., & Li, Min. (2014). Exploring gender differential item functioning (DIF) on eighth grade mathematics items for the United States and Taiwan. Seattle: University of Washington.

Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of Mathematics Items Associated with Gender DIF. *International Journal of Testing, 4*(2), 115-136.

McGraw, R., Lubienski, S. T., & Strutchens, M. E. (2006). A closer look at gender in NAEP mathematics achievement and affect data: Intersections with Achievement, race/ethnicity, and socioeconomic status. *Journal for Research in Mathematics Education, 37*(2), 129-150.

Mendes-Barnett, S., & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education, 19*(4), 289-304.

Moses, R.P. and Cobb, C.E. (2001). *Radical equations: Math literacy and civil rights*. Boston, MA: Beacon Press.

National Center for Educational Statistics (NCES). (2017). About NAEP: 'Inclusion of students with disabilities and English language learners.'

National Assessment Governing Board (NAGB). (2010, 2014). NAEP testing and reporting on students with disabilities and English language learners: Policy statement.

Osaďan, R., Belešová, M., & Szentesiová, L. (2018). Sissies, sportsmen and moms standing over stoves. Gender aspect of readers and mathematics textbooks for primary education in Slovakia. *Foro de Educación, 16*(25), 243-261.

Office of the Superintendent of Public Instruction (OSPI). (2011). Bias and sensitivity review of the common core state standards in English language arts and mathematics (implementation recommendations report). Prepared by Relevant Strategies.

Ryan, K. E., & Fan, M. (1996). Examining gender DIF on a multiple-choice test of mathematics: A confirmatory approach. *Educational Measurement: Issues and Practice, 15*(4), 15-20, 38.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2009). Responses to issues raised about validity, bias, and fairness in high-stakes testing. *American Psychologist, 64*(4), 285-287.

Smarter Balanced Assessment Consortium (2012). Bias and sensitivity guidelines. UC-Santa, Cruz, CA: Smarter Balanced Assessment Consortium.

Sullivan, A. L. (2013). School-based autism identification: Prevalence, racial disparities, and systemic correlates. *School Psychology Review, 42*(3), 298-316.

Taylor, C., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education, 25*(3), 246-280.

Woyshner, C. (2006). Picturing women: Gender, images, and representation in social studies. *Social Education, 70*(6), 358-362.

Zumbo, B. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. Language Assessment Quarterly, *4*(2), 223-233.

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. ETS Research Report Series, *2012*(1), 1-30.

Chapter 5. CONVERGENCE

In order to do ground-breaking assessment work, an interdisciplinary epistemological approach must be taken. The major critiques of conventional assessments relate to their use. However, in order to address these misuses, it is first necessary to revise the foundation upon which many of the problematic assumptions are made. Essentially, the flaws of assessment are inherent in the very structure out of which the assessment movement developed. The goal, then, becomes to identify strategies for assessment that intertwine traditional methods with novel ones, reframing the ways assessments are conceptualized, carried out, and interpreted. This is no easy task, and requires deploying a wide methodological net. Within the educational field, methodologies such as those based on critical race theory begin to do the hard work of challenging a very structurally oppressive and demographically enmeshed political system.

However, in practice, these approaches only go so far and focus too narrowly on specific identity oppression (i.e., race, even when used heuristically). They are not as fully encompassing of difference as they must be to radically transform the reductive and categorizing field of educational assessment as it stands currently. As a collection, my three papers serve as a foundation for developing a *new* understanding of how gendered ideas get encoded into high-stakes testing materials, how we can discover those hidden messages, and what we as educators and measurement professionals can do to address those issues with students, families, policy-makers, and other major testing stakeholders.

By describing these patterns within and across NAEP grade 8 math items in the first study, I was able to present strong evidence that this is not only a ubiquitous problem, but a complex one as well. There is no simple solution, such as removing a few problematic items here and there, as much of the literature would suggest is all that is necessary. In reality, what is

needed is a perspective shift, both in terms of what we mean when we describe an item as "fair" and what our goals of assessment are from the outset.

In the second study, we explored the potential for applying Machine Learning (ML) and Natural Language Processing (NLP) methodologies to the process of math test item bias and sensitivity review. In Study 1, while the method employed certainly revealed notable patterns in gendered representations in the contextualized math problems examined, it was also somewhat unclear how that new knowledge could be leveraged in a test development setting. Through a process of parsing a large corpus of math story problems and developing a text classifier that would automate the less-nuanced, but still crucial and time-intensive task of manual rating each item, we hope to have demonstrated one possible way to answer this procedural question.

Finally, in the last paper, I unpack both the "why" and the "how" of engaging with test items to reveal or resist biased representations of identity and culture. Math teacher feedback echoed the problems and needs addressed by my and other research on assessment inequity: more realistic item settings, less measurement of non-skill content. From a test development standpoint, these two goals can seem to be in opposition, but I offer guidelines and resources models for teachers or researchers to apply sound equitable item creation principles to their own work. I emphasize the importance of continued interrogation of our assessment systems and ongoing conversations with the actual end users of these *products*. Further, this 3rd paper sets the groundwork for the future development of a handbook for evaluating test materials for social/identity-based bias in language and structure.

Certain major themes arise when all three of the explorations are considered together. For one, cultural values, in this case those corresponding to gendered behavior are inscribed in our methods, our application of them, and our interpretation of data. Previous research frames equity

as a “validity problem” (a la Haertel, 2013 and Kane, 2013), and prescribes the use of outdated, partial, and/or purely statistical methods for addressing inequity. The hope is that through the new methodological lenses offered in the first and second papers, those limitations will not only be highlighted, but that a new framework for test development and item review might emerge. Test development and item review are presented in this work as inherently political. This means that rather than being a “neutralizing” process, item review nonetheless reproduces cultural values, at the item, reviewer, and interpretation level. Not only are the items we *reject* political, so too are the ones we keep.

This work raises challenging questions about representation, visibility, and invisibility in educational settings. More diverse and dynamic item contextualization poses obstacles to the parallel goal of making test *more* accessible to all students. How do we reconcile the need for wider representation in item contexts with the fact that this means not all items are “equally accessible” to all students? How would that impact reporting systems and data analysis? Is there to expand the range of representation without compromising our evaluative goals? These are not easy questions to answer, but my aim is to expand the conversation about what may be possible.

Future research may be able to address this through an exploration of what contexts and details contained in items are or aren’t legible to students belonging to specific cultural/identity groups. This is not a new idea (Abedi & Levine, 2013; Brown, 2011, Hendy et al., 2014), but it has yet to be explored in the context of the ways marginalized groups internalize dominant cultural messages and ideals. It would be quite interesting to explore how accessible questions where dominant cultural values (of gender or other identities) are emphasized impact students in the margins compared to those who are centered. If women were asked questions with stereotypically masculine-associated themes and men were asked questions with stereotypically

feminine-associated themes, who would perform better? Further, if items were written to appeal to and resonate with the experience of a very specific cultural group (ex. inner-city Puerto Rican girls, small-town working-class black boys, etc.), would the salience of the cultural references lead to more correct answering compared to items without those features included? While the focus of this work was not on test performance, these unanswered questions further the conversation about the purposes and approaches to thinking through assessment equity.

One critical question remains unanswered: Why does tweaking test items matter? It is an undeniably small aspect of the student experience, so why spend so much time exploring subliminal messages? The assertion in this work is *not* that test questions, themselves, pose such a severe, immediate threat. Rather, bias in test items represents the inescapability of cultural values and messages. Rigorous procedures exist for item review, specifically for identifying bias and stereotypical representation of identity. However, as evidenced from the first and second papers, the reviewers themselves, even those considered "experts" attuned to the content and contextual implications are operating within the same societal frameworks. From this perspective, it makes sense that questions including deeply ingrained ideas about gender and behavior would make their way to students and that students would answer these questions correctly. In a sense, assessment is doing exactly what it is supposed to: reproducing the status quo of hierarchical knowledge and test-performance-based displays of intelligence and competency, and this is no easy "fight" to engage in.

On the one hand, the broad contribution I aim to make with this research is to help demystify tests and testing, which this and previous research demonstrate is a distinct need for both teachers and students. Through a deep exploration of the test development and review process, I offer insight into a process that would otherwise be invisible to most test users. This

transparency is a key step in empowering students and educators to approach the tests with confidence and understanding. Throughout the work presented, however, there is also an inherent tension between the approaches I take to understanding test equity and the routine statistical procedures applied to equity and validity work. This raises many questions about how to reconcile a more expansive perspective on equity with the parameters and limitations of our current methodologies. I am left wondering how to envision "universal design" in testing that doesn't foreclose on a full representation of identity and that serves a more broadly equitable educational world. Student perspective was not made a central part of this research, but it seems clear from the literature, teacher feedback, and my own scholastic experiences, both as a student and as an educator, that their participation in and feedback on these topics would reveal critical sites for exploration and reframing of goals, methods, and outcomes, and help usher us into a new era of test literacy and equity in assessment.

If you have made it this far, I applaud and thank you! And if you have taken away even a small lesson on representation, identity, gender, or equity in math education and assessment, I have done my job and I can lay down my work satisfied.

Chapter 6. EPILOGUE

What comes before represents the culmination of 25 years of an education. More than 80% of my life thus far lived. Graduate school, alone, 8 years. My 20s. My "youth." To simply say that educational systems have shaped the course of my life would be a gross understatement. They are all I know. To be a "student" is to "be." My most complete and also partial identity. "I think, therefore I am" institutionalized. And still, to be at this point is surprising and surreal and dissociating and strange. Who is this person who "went all the way" and came out the other side a scholar and activist deeply committed to educational equity? How did they get here?

> 'ACT VII: INSTITUTIONS AS LIVING ORGANISMS
>
> *"Getting a Ph.D. is like being in an abusive relationship with someone you haven't actually met, yet."*
>
> I have been thinking a lot about institutions as beings, living things, that breathe and consume. If we frame institutions in this way, then we must also change how we frame and understand our relationship with them. In this sense, our relationships with institutions are not so different from our relationships with people. It is out of this thought project that the opening statement was borne. While harsh, it is undeniably true. We make a choice to be here, but we also must remember our power. The relationship has potential. All of the relationships we have with others and with individuals have potential. This is coalition-building. Cultivating these relationships is the meaning of life. *Gabbs Gorsky, 2015*'

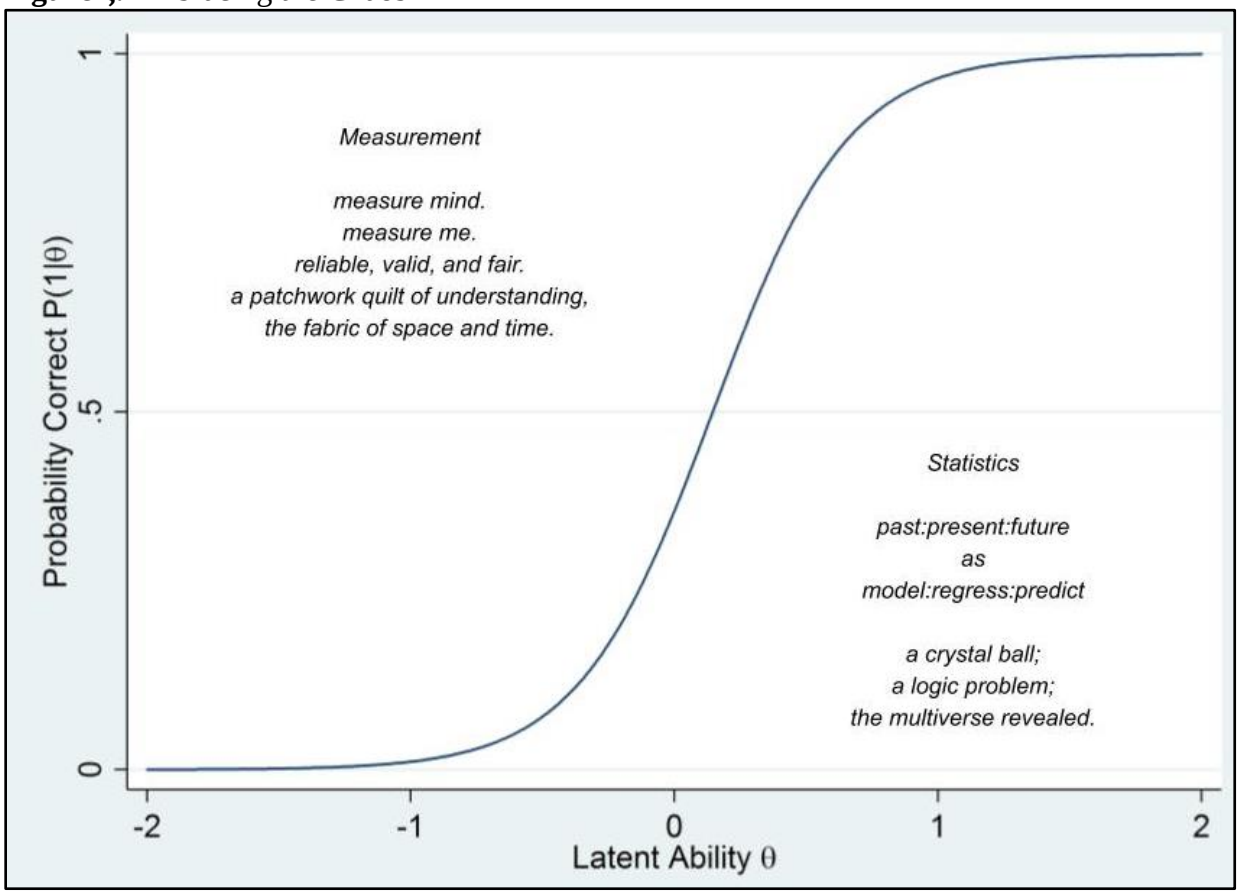
Several years back, in a deeply personally impactful class on the experience of women of color in academia, I wrote the above short piece about how my relationship with the academy was almost like a relationship with a person. With a romantic partner, even. A fraught relationship full of heartache and loss and sacrifice and compromise and empty faith and disappointment and exploitation and erasure and silencing. Also joy and empowerment and courage and gratitude and revelation and hope and care and potential and transformation and positive impact. Truly a labor of love. Of survival. I'm in awe that the 4-years-ago version of

myself had so much budding insight into what this whole thing really is about, for asking the hard questions before I even really understood their emotional and epistemological depth.

On the morning before I wrote this final piece, a sort of self-send-off, my therapist reminded me of a very important thing. She reminded me that the sum of my work is *not* the papers I've written nor the methods I've learned. It is my *impact* (given and received). The role I've played in reframing the problems that continue to plague educational systems through my unique lens. The challenging ideas wrestled with on no particular deadline aside from the ever-pressing urgency of social change. Most importantly, the relationships I've formed with others in so doing. The community I've forged. The bonds, both broken and built with people, places, and ideas. The highs and lows of self-discovery via the distorted reflection of institutional funhouse mirrors.

It is *these* educational moments that guided my work, that sustained me, that broke me and restored me again. In the introduction to the very recently released special issue of *The ANNALS of the American Academy of Political and Social Science* (2019), the editors ask: "What Use is Educational Assessment?" *What use?*, indeed. Very often when I share that I do research on [high-stakes] tests, I am met with some version of "So *you're* going to fix them, right?" While this is obviously a non-sequitur (of course I, as one individual cannot do that), the solution is actually in the question, itself. A return inward *is* the answer. The only answer. How can we claim to separate our systems from our souls? How to not care deeply about educational measurement-evaluation-assessment when you have lived it almost every day of your life? How to distance oneself from defining identities? From defining experiences? From gratitude?

Figure ζ. Embracing the Chaos



Note: Original poetry by Doctor Slacker, 2019.

Figure ζ. Logistic regression curve representing the growth of my self-understanding as a function of my lived experience.

FULL REFERENCES

Abedi, J., Hofstetter, C., & Baker, E. (2001). NAEP math performance and test accommodations: Interactions with student language background. CSE Technical Report 536. CRESST and NCES.

Abedi, J., & Levine, H. G. (2013). Fairness in assessment of English learners. *Leadership, 42*(3), 26-38.

Ambady, N., Shih, M., Kim, A., & Pittinsky, T. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science, 12*(5), 385-390.

Amit, M., & Fried, M. N. (2005). Authority and authority relations in mathematics education: A view from an 8th grade classroom. *Educational Studies in Mathematics, 58*(2), 145-168.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*(11), 1-16.

Apple, H. (2015). The $10,000 woman: Trans artifacts in the Pittsburgh queer history project archive. *TSQ: Transgender Studies Quarterly, 2*(4), 553-564.

Arendasy, M. E., & Sommer, M. (2012). Using automatic item generation to meet the increasing item demands of high-stakes educational and occupational assessment. *Learning and Individual Differences, 22*(1), 112-117.

Attali, Y. (2018). Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10947, 17-29.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment, 4*(3).

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258-267.

Au, W. (2011). Teaching under the new Taylorism: High-stakes testing and the standardization of the 21st century curriculum. *Journal of Curriculum Studies, 43*(1), 25-45.

Bhabha, H. K. (2004). *The location of culture.* Routledge.

Brown, Gavin T. L. (2011). Self-regulation of assessment beliefs and attitudes: A review of the students' conceptions of assessment inventory. *Educational Psychology*, 31(6), 731-748.

Bolte, E., Brown, D., Segrist, D., & Dudley, M. (2015). Investigating the moderating role of gender collective self-esteem on the relationship between the experience of gender microaggressions and psychological distress in women, ProQuest Dissertations and Theses.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems* (pp. 4349-4357).

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071.

Boyd, N., & Ramírez, H. N. R. (2012). *Bodies of evidence: The practice of queer oral history* (Oxford oral history series). New York, NY: Oxford University Press.

Bronfenbrenner, U. (1994). Ecological models of human development. In T. Husten & T. N. Postlethwaite (Eds), *International Encyclopedia of Education* (2nd ed., *Vol. 3*, pp. 1643–1647). New York: Elsevier Science.

Caines, J.L., Bridglall, B., & Chatterji, M. (2014). Understanding validity and fairness issues in high-stakes individual testing situations. *Quality Assurance in Education, 22*(1), 5-18.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Measurement methods for the social sciences series; 4). Thousand Oaks [Calif.]: Sage Publications.

Casad, B., Hale, P., & Wachs, F. (2017). Stereotype threat among girls. *Psychology of Women Quarterly, 41*(4), 513-529.

Chatterji, M. (Ed.). (2013). *Validity and test use: An international dialogue on educational assessment, accountability and equity* (First edition. ed.). Bingley United Kingdom: Emerald.

Cheryan, S., Plaut, V. C., Davies, P. G., & Steele, C. M. (2009). Ambient belonging: How stereotypical cues impact gender participation in computer science. *Journal of Personality and Social Psychology, 97*, 1045–1060.

Cimpian, J. R., Lubienski, S. T., Timmer, J. D., Makowski, M. B., & Miller, E. K. (2016). Have gender gaps in math closed? Achievement, teacher perceptions, and learning behaviors across two ecls-k cohorts. *Grantee Submission, 2*, En.

Condie, S., Lefgren, L., & Sims, D. (2014). Teacher heterogeneity, value-added and education policy. *Economics of Education Review, 40*(C), 76-92.

Cooley, E., Payne, B., & Phillips, K. (2014). Implicit bias and the illusion of conscious ill will. *Social Psychological and Personality Science, 5*(4), 500-507.

Cundiff, J. L., Vescio, T. K., Loken, E., & Lo, L. (2013). Do gender–science stereotypes predict science identification and science career aspirations among undergraduate science majors? *Social Psychology of Education, 16*(4), 541–554.

Dee, T. (2005). A teacher like me: Does race, ethnicity, or gender matter? *The American Economic Review*, *95*(2), 158-165.

Deluca, C., & Bellara, A. (2013). The current state of assessment education: aligning policy, standards, and teacher education curriculum. *Journal of Teacher Education, 64*(4), 356-372.

Derrida, J. (1976). *Of grammatology* (1st American ed.). Baltimore: Johns Hopkins University Press.

Driskill, Q. *Asegi stories: Cherokee queer and two-spirit memory*. Tucson: University of Arizona Press, 2016.

Eccles, I., Midgley, C., & Adler, T.F. (1984). Grade-related changes in the school environment: Effects on achievement motivation. In G. Nicholls (Ed.), *Advances in Motivation and Achievement*, 3, 283-331. Greenwich, CT: JAI Press.

Educational Testing Service. (2014). ETS standards for quality and fairness. Princeton, N.J.: Educational Testing Service.

Eisenberg, D., Golberstein, E., & Hunt, J. B. (2009). Mental Health and Academic Success in College. *The B.E. Journal of Economic Analysis & Policy*, *9*(1) (Contributions), 40.

Ellis, S., & Smith, V. (2017). Assessment, teacher education and the emergence of professional expertise. *Literacy*, 51(2), 84-93.

Emde, R. (2006). Culture, diagnostic assessment, and identity: Defining concepts. *Infant Mental Health Journal,* 27(6), 606-611.

Galdi, S., Cadinu, M., & Tomasetto, C. (2014). The roots of stereotype threat: When automatic associations disrupt girls' math performance. *Child Development, 85*(1), 250-263.

Gallagher, A. M. (1998). Gender and antecedents of performance in mathematics testing. *Teachers College Record*, 100, 297-314.

Gallagher, A. M., Morley, M. E., & Levin, J. (1999). Cognitive patterns of gender differences on mathematics admissions tests. In *New Directions in Assessment for Higher Education* (GRE FAME Rep. 3; pp. 4-11). Princeton, NJ: Educational Testing Service.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences, *115*(16), E3635-E3644.

Gayles, J. (2011). Attracting and retaining women in STEM (New directions for institutional research; no. 152). San Francisco: Jossey-Bass.

Gierl, M. J., & Lai, H. (2012). The role of item models in automatic item generation. *International Journal of Testing, 12*(3), 273-298.

Gierl, M., Lai, H., & Chen, Y. (2018). Using automatic item generation to create solutions and rationales for computerized formative testing. *Applied Psychological Measurement, 42*(1), 42-57.

Gillock, K. L., & Reyes, O. (1996). High school transition-related changes in urban minority students' academic performance and perceptions of self and school environment. *Journal of Community Psychology, 24*(3), 245-261.

Glick, P., & Fiske, S. T. (2001). An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. American Psychologist, 56(2), 109–118.

Good, J., Woodzicka, J., & Wingfield, L. (2010). The effects of gender stereotypic and counter-stereotypic textbook images on science performance. *The Journal of Social Psychology, 150*(2), 132-47.

Gorski, Paul C. (2008). Good intentions are not enough: A decolonizing intercultural education. *Intercultural Education, 19*(6), 515-525.

Gorski, Paul C. (2012). Perceiving the problem of poverty and schooling: Deconstructing the class stereotypes that misshape education practice and policy. *Equity & Excellence in Education*, 45(2), 302-319.

Gorsky, G. S., Wang, N., & Phelps, D. Rethinking consequential validity: Linguistic deconstruction of 'gold-standard' math story problems. In review.

Green, R., & Griffore, R. (1980). The impact of standardized testing on minority students. *The Journal of Negro Education, 49*(3), 238-252.

Greenwald, A. G.; Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*. 102, 4–27.

Greenwald, A. G.; McGhee, D. E.; Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*. 74, 1464–1480.

Gunderson, E., Ramirez, A., Levine, G., & Beilock, S. (2012). The role of parents and teachers in the development of gender-related math attitudes. *Sex Roles, 66*(3), 153-166.

Haertel, E. (2013). Getting the help we need. *Journal of Educational Measurement*, *50*(1), 84-90.

Hamdan, S. (2010). English-language textbooks reflect gender bias: A case study in Jordan. *Advances in Gender and Education*, 2, 22-26.

Harris, B., Ravert, R. D., & Sullivan, A. L. (2017). Adolescent racial identity: Self-identification of multiple and "other" race/ethnicities. *Urban Education, 52*(6), 775-794.

Harris, D. (2011). Value-added measures and the future of educational accountability. *Education Science, 333*(6044), 826-827.

Helms, J., Sackett, P., Borneman, M., & Connelly, B. (2009). Defense of tests prevents objective consideration of validity and fairness/responses to issues raised about validity, bias, and fairness in high-stakes testing. *The American Psychologist, 64*(4), 283.

Hendrix, K., Trotter, S., Counts, M., Penick, S., & Popkin, J. (2009). The impact of gender on standardized testing a meta-analysis of existing studies,1987-2007, ProQuest Dissertations and Theses.

Hendy, H., Schorschinsky, N., & Wade, B. (2014). Measurement of math beliefs and their associations with math behaviors in college students. *Psychological Assessment, 26*(4), 1225-1234.

Hernandez, R., Conoley, C. W., Israel, T., & Consoli, M. M. (2014). Responding to perceived racial microaggressions: Impacts on the mental health and academic persistence attitudes of Latina/o college students, ProQuest Dissertations and Theses.

Hidalgo-Montesinos, M. D., & Gómez-Benito, J. (2003). Test purification and the evaluation of differential item functioning with multinomial logistic regression. *European Journal of Psychological Assessment, 19*(1), 1–11.

Hollands, A., Temple, J. B., Henry, S. L., Yves, R. C., & Thao, Y. (2012). Fostering hope and closing the academic gap: An examination of college retention for African-American and Latino students who participate in the Louis Stokes Alliance minority participation program (learning community) while enrolled in a predominately white institution, ProQuest Dissertations and Theses.

Howard, T. (2010). *Why race and culture matter in schools: Closing the achievement gap in America's classrooms* (Multicultural education series (New York, N.Y.)). New York, N.Y.: Teachers College Press.

Hyde, J., & Linn, M. (2006). Diversity. Gender similarities in mathematics and science. *Science, 314*(5799), 599-600.

Ibarra, R. A. (2001). *Beyond affirmative action*. Madison, WI: University of Wisconsin Press.

Jamieson, J., & Harkins, S. (2012). Distinguishing between the effects of stereotype priming and stereotype threat on math performance. *Group Processes & Intergroup Relations, 15*(3), 291-304.

Kan, A., & Bulut, O. (2014). Examining the relationship between gender DIF and language complexity in mathematics assessments. *International Journal of Testing, 14*(3), 245-264.

Kane, Michael. (2010). Validity and fairness. *Language Testing*, 27(2), 177-182.

Kane, M. (2013). Validation. *National Conference of Bar Examiners*, 17-64.

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of Advertising, 46*(1), 141-155.

Keith, J. (2016). The impacts of microaggressions on the performance of multiracial and monoracial college students. McNair Scholars Online Journal, 10(1), *McNair Scholars Online Journal, 10*(1).

Kerr, B., & Multon, K. (2015). The development of gender identity, gender roles, and gender relations in gifted students. *Journal of Counseling & Development, 93*(2), 183-191.

Kunnan, A. (2007). Test fairness and DIF. *Language Assessment Quarterly*, 4(2), 109-112.

Lai, A., & Tetreault, J. (2018). Discourse coherence in the wild: A dataset, evaluation and methods. arXiv preprint arXiv:1805.04993.

Lan, M., & Li, Min. (2014). Exploring gender differential item functioning (DIF) on eighth grade mathematics items for the United States and Taiwan. Seattle: University of Washington.

Lester, J., Yamanaka, A., & Struthers, B. (2016). Gender microaggressions and learning environments: The role of physical space in teaching pedagogy and communication. *Community College Journal of Research and Practice,* 1-18.

Levine, M., & Levine, A. (2013). Holding accountability accountable: A cost-benefit analysis of achievement test scores. *American Journal of Orthopsychiatry, 83*(1), 17-26.

Levy, R. (2008). 'Third spaces' are interesting places: Applying 'third space theory' to nursery-aged children's constructions of themselves as readers. *Journal of Early Childhood Literacy*. *8*(1), 43–66.

Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of Mathematics Items Associated with Gender DIF. *International Journal of Testing, 4*(2), 115-136.

Ling, W., Yogatama, D., Dyer, C., & Blunsom, P. (2017). Program induction by rationale generation: Learning to solve and explain algebraic word problems. In Proc. ACL.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Lomax, R., West, M., Harmon, M., Viator, K., & Madaus, G. (1995). The Impact of Mandated Standardized Testing on Minority Students. *The Journal of Negro Education, 64*(2), 171-185.

Los Angeles Community College District. (2012). Investigations for bias, offensiveness and sensitivity. Prepared for the California Community Colleges Chancellor's Office Assessment Validation Training. https://www.laccd.edu/Departments/EPIE/Documents/TestBias.pdf

Luckin, R. (2018). *Machine learning and human intelligence: The future of education for the 21st century*. London: UCL Institute of Education Press.

Ma, A. (2017). The Application of Machine Learning to Education. *Journal of Student Science and Technology, 10*(1).

Madaus, G. F., & Clarke, M. (2001). *The adverse impact of high stakes testing on minority students: Evidence from 100 years of test data.*

Maniotes, L. K. (2005). The transformative power of literary third space. Doctoral dissertation, School of Education, University of Colorado, Boulder.

Martin, C. & Ruble, D. (2004). Children's search for gender cues. Cognitive Perspectives on gender development. *Current Directions in Psychological Science, 13*(2), 67-70.

Mattarella-Micke, A., & Beilock, S. (2010). Situating math word problems: The story matters. *Psychonomic Bulletin & Review, 17*(1), 106-111.

Mayes, C. (2007). *Understanding the whole student: Holistic multicultural education*. Lanham: Rowman & Littlefield Education.

McGraw, R., Lubienski, S. T., & Strutchens, M. E. (2006). A closer look at gender in NAEP mathematics achievement and affect data: Intersections with Achievement, race/ethnicity, and socioeconomic status. *Journal for Research in Mathematics Education, 37*(2), 129-150.

McGuire, T., & Miranda, J. (2008). New evidence regarding racial and ethnic disparities in mental health: Policy implications. *Health Affairs (Project Hope), 27*(2), 393-403.

McKellar, S., Marchand, A., Diemer, M., Malanchuk, O., & Eccles, J. (2019). Threats and supports to female students' math beliefs and achievement. *Journal of Research on Adolescence, 29*(2), 449-465

Medina, N., Neill, D. Monty, & National Center for Fair & Open Testing. (1990). *Fallout from the testing explosion: How 100 million standardized exams undermine equity and excellence in America's public schools* (Rev. 3rd ed.). Cambridge, MA: National Center for Fair & Open Testing (FairTest).

Mendes-Barnett, S., & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education, 19*(4), 289-304.

Messick, S. (1987). Validity. *ETS Research Report Series*, (2), 13-208.

Millsap, R. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research*, 30(4), 577-605.

Moses, R.P. and Cobb, C.E. (2001). *Radical equations: Math literacy and civil rights*. Boston, MA: Beacon Press.

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

Mizelle, N. B. & Irvin, J. L. (2000). Transition from middle school into high school, *Middle School Journal*, (*31*)5, 57-61.

Nadal, K. (2013). *That's so gay! : Microaggressions and the lesbian, gay, bisexual, and transgender community* (1st ed., Perspectives on sexual orientation and gender diversity series). Washington, D.C.: American Psychological Association.

Nadeem, F., & Ostendorf, M. (2017). Language based mapping of science assessment items to skills. In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications (pp. 319-326).

Nadeem, F., & Ostendorf, M. (2018). Estimating linguistic complexity for science texts. In Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 45-55).

Napolitano, D., Sheehan, K., & Mundkowsky, R. (2015). Online readability and text complexity analysis with TextEvaluator. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (pp. 96-100).

National Center for Educational Statistics (NCES). (2017). About NAEP: 'Inclusion of students with disabilities and English language learners.'

National Assessment Governing Board (NAGB). (2010, 2014). NAEP testing and reporting on students with disabilities and English language learners: Policy statement.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math = male, me = female, therefore math ≠ me. *Journal of Personality and Social Psychology, 83*(1), 44–59.

Osaďan, R., Belešová, M., & Szentesiová, L. (2018). Sissies, sportsmen and moms standing over stoves. Gender aspect of readers and mathematics textbooks for primary education in Slovakia. *Foro de Educación, 16*(25), 243-261.

Parrott, L., Spatig, L., Kusimo, P. S., Carter, C. C., & Keyes, M. (2000). Troubled waters: Where multiple streams of inequality converge in the math and science experiences of nonprivileged girls. *Journal of Women and Minorities in Science and Engineering*, *6*(1).

Pérez, E. (1999). *The decolonial imaginary: Writing Chicanas into history* (Theories of representation and difference). Bloomington: Indiana University Press.

Picho, K., & Schmader, T. (2017). When do gender stereotypes impair math performance? A study of stereotype threat among Ugandan adolescents. *Sex Roles*.

Pierce, C. (1970). Offensive mechanisms. In F. Barbour (Ed.), *The Black Seventies* (pp. 265-282). Boston, MA: Porter Sargent.

Prather, C., & Ruiz, John. (2015). *Nice dissertation, for a girl: Cardiovascular and emotional reactivity to gender microaggressions*, ProQuest Dissertations and Theses.

Prentice, D. A., & Carranza, E. (2002). What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of Women Quarterly, 26*(4), 269–281.

Powers, D., Burstein, J., Chodorow, M., Fowles, M., & Kukich, K. (2002). Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research, 26*(4), 407-425.

Rawson, KJ. (2015). Introduction: 'An inevitably political craft.'[Special Issue on Archives and Archiving] *TSQ: Transgender Studies Quarterly 2*(4), 544-552.

Rolfe, G. (2004). Deconstruction in a nutshell. *Nursing Philosophy, 5*(3), 274-276.

Rosenthal, B., & Wilson, W. (2012). Race/ethnicity and mental health in the first decade of the 21st century. *Psychological Reports, 110*(2), 645-662.

Roth, W. (1996). Where IS the context in contextual word problem?: Mathematical practices and products in grade 8 students' answers to story problems. *Cognition and Instruction, 14*(4), 487-527.

Ryan, K. E., & Fan, M. (1996). Examining gender DIF on a multiple-choice test of mathematics: A confirmatory approach. *Educational Measurement: Issues and Practice, 15*(4), 15-20, 38.

Ryan, K. E., & Ryan, A. M. (2005). Psychological processes underlying stereotype threat and standardized math test performance. *Educational Psychologist, 40*(1), 53-63.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2009). Responses to issues raised about validity, bias, and fairness in high-stakes testing. *American Psychologist, 64*(4), 285-287.

Sap, M., Prasettio, M. C., Holtzman, A., Rashkin, H., & Choi, Y. (2017, September). Connotation frames of power and agency in modern films. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 2329-2334).

Sass, T. R., Semykina, A., & Harris, D. N. (2014). Value-added models and the measurement of teacher productivity. *Economics of Education Review, 38*(C), 9-23.

Schley, D. R., & Fujita, K. (2014). Seeing the math in the story: On how abstraction promotes performance on mathematical word problems. *Social Psychological and Personality Science, 5*(8), 953-961.

Schmader, T. (2002). Gender Identification Moderates Stereotype Threat Effects on Women's Math Performance. *Journal of Experimental Social Psychology*, 38(2), 194–201.

Schreier, M. (2014). Qualitative content analysis, In Flick, U., Metzler, Katie, & Scott, Wendy (Eds.), *The SAGE Handbook of Qualitative Data Analysis* (pp. 170-183). Los Angeles; London [England]: SAGE.

Sekaquaptewa, D., & Thompson, M. (2002). Solo status, stereotype threat, and performance expectancies: Their effects on women's performance. *Journal of Experimental Social Psychology*, 39, 68–74.

Settles, I., O'Connor, R., & Yap, S. (2016). Climate perceptions and identity interference among undergraduate women in stem: The protective role of gender identity. *Psychology of Women Quarterly, 40*(4), 488-503.

Sexton, D., Bartlett, L., Stoddart, T., & Warren L. J. (2018). Becoming a math teacher: Recruitment, preparation, and practice, ProQuest Dissertations and Theses.

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

Sheehan, K. M., Flor, M., & Napolitano, D. (2013). A two-stage approach for generating unbiased estimates of text complexity. In Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility, 49-58.

Singley, M., & Bennett, R. E. (1998). Validation and extension of the mathematical expression response type: Applications of schema theory to automatic scoring and item generation in mathematics (GRE Board professional report; no. 93-24P). Princeton, NJ: Educational Testing Service.

Skerrett, A. (2010). Lolita, Facebook, and the third space of literacy teacher education. *Education Studies*. *46*(1), 67–84.

Smarter Balanced Assessment Consortium (2012). Bias and sensitivity guidelines. UC-Santa, Cruz, CA: Smarter Balanced Assessment Consortium.

Smeding, A. (2012). Women in science, technology, engineering, and mathematics (STEM): An investigation of their implicit gender stereotypes and stereotypes' connectedness to math performance. *Sex Roles, 67*(11), 617-629.

Solano-Flores, G., & Nelson & Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 38(5), 553-573.

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.

Spikol, D., Avramides, K., Cukurova, M., Vogel, B., Luckin, R., Ruffaldi, E., & Mavrikis, M. (2016). Exploring the interplay between human and machine annotated multimodal learning analytics in hands-on STEM activities. Proceedings of the Sixth International Conference on Learning Analytics & Amp Knowledge; 522-523.

Srihari, S., Collins, J., Srihari, R., Srinivasan, H., Shetty, S., & Brutt-Griffler, J. (2008). Automatic scoring of short handwritten essays in reading comprehension tests. *Artificial Intelligence, 172*(2), 300-324.

Steele, C. (2010). *Whistling Vivaldi: And other clues to how stereotypes affect us* (1st ed., Issues of our time (W.W. Norton & Company)). New York: W.W. Norton & Company.

Steele, C., Aronson, J., & Kruglanski, Arie W. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797-811.

Steffens, M., Jelenec, P., & Noack, P. (2010). On the leaky math pipeline: Comparing implicit math-gender stereotypes and math withdrawal in female and male children and adolescents. *Journal of Educational Psychology, 102*(4), 947-963.

Sterzing, P., Gartner, R., Woodford, M., & Fisher, C. (2017). Sexual orientation, gender, and gender identity microaggressions: Toward an intersectional framework for social work research. *Journal of Ethnic & Cultural Diversity in Social Work, 26*(1-2), 81-94.

Sue, D. (2010). *Microaggressions in everyday life: Race, gender, and sexual orientation*. Hoboken, N.J.: Wiley.

Sue, D. (2010). *Microaggressions and marginality: Manifestation, dynamics, and impact.* Hoboken, N.J.: Wiley.

Sullivan, A. L. (2013). School-based autism identification: Prevalence, racial disparities, and systemic correlates. *School Psychology Review, 42*(3), 298-316.

Taie, S., & Goldring, R. (2017). Characteristics of public elementary and secondary school teachers in the United States: Results from the 2015-16 national teacher and principal survey. First look. National Center for Education Statistics, 2017.

Taylor, C., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education, 25*(3), 246-280.

Thomas, K. & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior, 77,* 184-197.

Tomasetto, C., Alparone, F. R., & Cadinu, M. (2011). Girls' math performance under stereotype threat: The moderating role of mothers' gender stereotypes. *Developmental Psychology, 47*(4), 943-949.

Voyer, D., Voyer, S. D., & Hinshaw, S. P. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin, 140*(4), 1174-1204.

Warne, R., Yoon, M., Price, C., Zárate, M. A., & Lee, R. M. (2014). Exploring the various interpretations of "test bias". *Cultural Diversity and Ethnic Minority Psychology, 20*(4), 570-582.

Wei, M., Ku, T., Liao, K., & Zárate, Michael A. (2011). Minority stress and college persistence attitudes among African American, Asian American, and Latino students: Perception of university environment as a mediator. *Cultural Diversity and Ethnic Minority Psychology, 17*(2), 195-203.

Williams, P. (1991). *The alchemy of race and rights*. Cambridge, Mass.: Harvard University Press.

Woyshner, C. (2006). Picturing women: Gender, images, and representation in social studies. *Social Education, 70*(6), 358-362.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1480-1489).

Zumbo, B. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. Language Assessment Quarterly, *4*(2), 223-233.

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. ETS Research Report Series, *2012*(1), 1-30.

VITA

Gabriella Silva Gorsky (they/them/theirs) was born in Chicago, IL. They graduated from DePaul University in 2011 with a BA in Psychology with a minor in Women's and Gender Studies. In 2013, Gabriella completed their MA in Educational Research Methodology from Loyola University Chicago. In 2013, they moved to Seattle, WA to pursue their PhD in Educational Measurement and Statistics at the University of Washington Seattle. They currently live in Seattle with their spouse, Dash, and their little dog, Pip. Gabriella will be staying in Seattle to launch their professional career after graduation and is excited for this next stage of life.