EXPLORING THE IMPACT OF CULTURALLY-RELEVANT ASSESSMENT MATERIAL ON BLACK STUDENTS' READING COMPREHENSION

by

ARMANI M. MORRIS

(Under the Direction of Matthew J. Madison & George Engelhard, Jr.)

ABSTRACT

Culturally-relevant assessment materials include content and strategies that prioritize the ways that students of color demonstrate competence. These materials have been shown to improve academic and assessment outcomes for culturally and linguistically diverse students. This study applies a diagnostic classification model (DCM) to investigate the impact of culturally-relevant assessment materials on Black students' reading comprehension scores. Firstly, I conducted a simulation study examining three methods of incorporating a covariate into a DCM. Results indicate that the simultaneous inclusion of covariates in DCM estimation yields higher accuracy and reliability than post-hoc methods. Secondly, I conducted an empirical analysis to explore the impact of Black examinees' prior reading level and cultural knowledge on their mastery classifications for reading. Results of the empirical analysis suggest that average-level readers with high prior knowledge of African-American culture have a higher chance of being masters on African-American texts than they do on other cultural texts.

INDEX WORDS: Black/African-American students, culturally-relevant assessment, diagnostic classification models, reading comprehension, assessment design

EXPLORING THE IMPACT OF CULTURALLY-RELEVANT ASSESSMENT MATERIAL ON BLACK STUDENTS' READING COMPREHENSION

By

ARMANI M. MORRIS

Master of Arts, The University of Georgia, 2024

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

MASTER OF ARTS

ATHENS, GEORGIA

2024

EXPLORING THE IMPACT OF CULTURALLY-RELEVANT ASSESSMENT MATERIAL

ON BLACK STUDENTS' READING COMPREHENSION

by

ARMANI M. MORRIS

Major Professor:	Matthew J. Madison
	George Engelhard, Jr.
Committee:	Ruanda Garth-McCullough

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2024

DEDICATION

The support that I've received while writing this thesis has been incredible. Without God, I wouldn't have seen this chapter of my life come to fruition. And to my family, friends, and academic colleagues, I appreciate your support, guidance, and wisdom more than you could understand. Thank you for helping me recognize that the work that I am pursuing is much bigger than myself. I couldn't have understood my potential impact without you all.

TABLE OF CONTENTS

GLOSSARY

Term	Definition
Attribute	In the DCM context, the categorical latent skill or concept that examinees demonstrate proficiency on in an exam.
Attribute profiles	In the DCM context, these profiles represent the various master and non-master classifications that can exist per attribute in an item. The number of profiles is given by 2^A, where A is the number of attributes in an item.
Base rate	In the DCM context, the proportion of masters and non-masters on an assessment. Normally represented as a decimal.
Criterion-referenced	An inference made about an examinee by comparing their score on assessment to a predetermined score set to represent student proficiency.
Cultural capital	The skills, knowledge, and assets that originate and are circulated in cultural groups.
Culturally-relevant pedagogy (CRP)	Coined by Gloria Ladson-Billings in 1995, CRP is a framework that encourages the centering of Black culture and experiences in mainstream teaching.
Culturally-sustaining practices	Practices designed to maintain the ways of knowing from students of color.
Diagnostic classification model (DCM)	An assessment modeling framework in which examinee proficiency on various attributes is measured as a categorical latent trait.
Educational stakeholder	For this study, an educational stakeholder is any person that is directly or indirectly involved in teaching students, facilitating classroom experiences and materials, and conducting research on schools and students.
Formative assessment	An assessment used to evaluate students' understanding of a unit while students are still enrolled in the class.
Item characteristic curve	A line graph that represents examinee probability of answering an item correctly based on their ability level. Used in IRT contexts.
Item characteristic bar chart	A visualization of item parameters when using a DCM model.
Item discrimination	A measure of how well an item differentiates between examinees of varying ability levels.

Item response function (IRF)	The graphical representation of examinee responses probabilities based on the characteristics of the item.
Item response pattern	An examinee's pattern of correct and incorrect responses on an assessment.
Item response probability	The likelihood of a correct response based on the examinee's attribute pattern on an item.
Item response theory (IRT)	An assessment modeling framework that evaluates students along a continuum of scores based on their assessment scores.
Latent trait	A characteristic or behavior that cannot be measured by visual observation alone. These traits can include emotions, cognitive functions, attitudes, and thinking patterns.
Mastery classification	A category that represents examinee proficiency on an assessment, typically "master" and "non-master" in DCM contexts.
Meritocracy	In testing, the idea that an individual's score is solely based on how much effort they put into an exam.
Norm-referenced	An inference made about an examinee by comparing their score on an assessment to other examinees (or groups of examinees) that took the same assessment.
Posterior probability	In the DCM context, it is the probability of an examinee being classified as proficient (or a master) on an attribute based on their response pattern for items measuring that attribute.
Summative assessment	An assessment used to evaluate total students' understanding of skills and concepts at the end of a course.

CHAPTER 1

INTRODUCTION

"When I originally began searching for research on successfully educating African-American students, I found nothing. …It was clear that there was no language of academic excellence associated with African-American students." – (Ladson-Billings, 2014, p. 76)

What images or words come to mind when you think of a Black student? The above quote from Ladson-Billings brings me back to the beginning of my pursuit of an academic research career at Spelman College. With my undergraduate thesis centered around middle-school-aged African-American students' classroom experiences, I often combed through decades of literature on student behavior and academic trajectories. When I would type in "Black students and" into the database search box, words such as "dropout rates," "underachievement," "disciplinary referrals," and more would be suggested. These buzzwords in the literature project an image that Black students' academic underperformance and maladjustment are constantly observed.

I found many articles of researchers observing Black students from a deficit rather than a growth perspective. One must consider how these reported experiences contribute to academic stakeholders' overall perception of Black students' ability to learn, grow, and thrive in their learning environments. How are their trajectories measured and interpreted compared to students of other racial groups? Specifically, how do these perceptions affect how Black students take large-scale assessments, and how do educational stakeholders interpret these scores?

For decades, educational researchers and testing companies have reported that Black students underperform academically compared to their White counterparts (Stewart & Haynes, 2016; Au, 2016; Sireci & Randall, 2021; Randall, Poe, and Slomp, 2021; Perez, 2002). Despite

educational innovations and increased diversity in student populations in urban classrooms in the last five decades, this trend in assessment results is still constantly observed. Considering the historical implications of racism in education, researchers must consider how the United States educational system and mainstream curriculum often downplay Black students' learning abilities. Further, we must investigate how the design of the most widely used academic assessments may not allow Black students to demonstrate the best of their ability. Therefore, standardized testing practices must be examined through a racialized lens to address these discrepancies in scoring and academic outcomes for African-American students.

Ladson-Billings (1995, 2014) first introduced the Culturally-Relevant Pedagogy (CRP) framework as a method for teachers (especially White teachers) in urban schools with predominantly Black populations, to engage their students in the learning environment by contributing aspects of their languages, experiences, and ways of thinking into the lessons and activities. In a recent publication, she clarified that culture is dynamic and changes through time, location, and social context. Therefore, teachers should navigate away from static and often outdated understandings of how Black students engage with the world; teachers' lessons should be as dynamic as their students are. Therefore, CRP centers on how students of color understand the learning material and prioritizes their ways of demonstrating their academic excellence (Randall, 2021). Given these lessons from Ladson-Billings, educational stakeholders should consider how this concept can be extended into the academic testing space.

There is an extensive history of standardized intelligence tests that place students of color and low-income students in special education and lower academic tracks than their White classmates (Sireci & Randall, 2021; Perez, 2002). For most of the United States' history, people of color have been physically, legally, and intellectually separated from majority-White spaces.

Therefore, it is unfair that the information (cultural and intellectual) that stemmed from and circulated within the majority-White spaces is the standard learning material used to test students outside of this context. Though Jim Crow laws allowed these racist designs and interpretations of intelligence tests to go unchecked for decades, we must consider how the legacy of these tests continue to misidentify African-American students into lower academic tracks and limit their opportunities for academic excellence compared to White students.

Understanding the racialized histories of standardized testing, we should consider the implications of these tests on students' futures. After all, the data that testing companies, teachers, and researchers report represent real students whose academic trajectory and future can be affected by a single test score. In that case, students and educators should understand how to identify and address the specific needs of their students to optimize their academic journeys. Diagnostic classification models (DCMs) are a promising modeling tool that will allow examinees and other users of testing data to understand which skills and concepts students demonstrate understanding. Many traditional assessments use single scores of students' overall performance on an exam or on subscales to quantify student ability. On the other hand, DCMs are used to quantify students' performance on the specific skills and concepts that the exams measure. This DCM feedback can give stakeholders are more actionable image of students' academic performance.

Purpose of the Study

Entering my graduate studies in education after the peak of the COVID-19 pandemic, I became increasingly interested in how the educational landscape was being shaped as students returned to in-person schooling. During this time, many educators around the world received a more intimate view of their students' home contexts through online teaching. Teachers often had

to adapt to home factors (e.g., students' home habits, parental cooperation and expectations, and environmental distractions) that became amplified as the classroom and home environment merged. When attending educational conferences and workshops with researchers and educators, everyone's main question was: Where do we go from here? Therefore, I became interested in how educators would continue implementing the lessons they learned about student engagement and success during the pandemic in a post-pandemic environment. Most importantly, how will aspects of students' personal needs, preferences, and interests be highlighted or prioritized in their learning environments?

This question is critical to me when considering young African-American students whose culture, needs, and voices are often not reflected in mainstream education. These students strive to succeed in a system that was, at its core, designed for their failure. This educational renaissance is a perfect opportunity to shed light on the complex and deeply-rooted educational inequalities that are often hidden under more shallow efforts to increase diversity and equity in schools. These investigations are critical in the testing fields as well.

Academic assessments must decenter White culture and intelligence to allow students of color to demonstrate their knowledge fairly. Therefore, culturally-relevant assessments need to be considered. Since African-American students' academic trajectory is often reported compared to students of other races, I believe that using DCMs in conjunction with traditional assessment modeling encourages researchers to examine Black students from a growth perspective. This statement does not negate that differences between groups may exist. Instead, DCMs provide feedback on how students perform based on each skill that an assessment measures, allowing educators to provide more specific interventions for Black students to give them what they need to succeed and grow.

Overall, this study explores the effects of culturally-relevant assessment material on examinee reading comprehension scores in a DCM context. In this study, I will conduct a secondary analysis using Garth-McCullough's (2008) data to explore the effects of culturally-relevant assessment material when including students' prior cultural knowledge or reading ability.

In sum, the purpose of the study is to:

1. Conduct a simulation study testing three methods of including a covariate in a DCM analysis.
2. Conduct an empirical analysis of reading assessment data to evaluate the impact of culturally-relevant assessment material on Black students' reading comprehension scores using the method chosen from the simulation study.
3. Use the findings from the empirical analysis to discuss the implications of culturally-relevant material.

CHAPTER 2

LITERATURE REVIEW

This chapter introduces the foundational concepts for culturally-relevant assessment and diagnostic classification models (DCMs). I will provide background on each concept and its connection to the present study. This literature review has five main goals:

1. To introduce what standardized testing is and how it is used in K-12 education,
2. To discuss the pattern of Black students underperforming on standardized tests compared to White students and the role that unfair testing design and content play in this pattern,
3. To discuss the effects of unfair testing practices on Black students' testing behavior and academic outcomes,
4. To introduce elements of culturally-relevant assessment and how it could be used to address the racial discrepancies in testing scores, and finally,
5. To introduce diagnostic classification models (DCMs) and discuss how they can be used as assets for culturally-relevant assessments.

What Is a Standardized Assessment?

In this section, I discuss how standardized educational assessments are developed for K-12 education and stakeholders' decisions using these tests.

Standardized assessments have become integral to the general United States education system. Students are administered exams from various test design entities, including independent testing companies and state governments. These assessments measure students' understanding of school subjects and evaluate their ability to demonstrate specific skills. Educational stakeholders such as school administrators, school boards, government agencies, and others use student scores

on these exams to track student learning, evaluate school and teacher quality, and make funding decisions for schools and academic programs (Ladd, 2017; No Child Left Behind, 2001).

In a large-scale effort to keep school districts accountable for improving students learning outcomes, meeting educational standards, and providing the necessary programs and interventions to support students, the No Child Left Behind (NCLB) Act was implemented in 2001 (No Child Left Behind Act, 2001). This act required all state governments to administer standardized assessments to students from third to eighth grade, and at least once in high school, to encourage schools to help keep all students on track towards proficiency in core subjects such as math, reading, and science (Ladd, 2017). With its implementation, standardized tests became the norm for the K-12 education system and curriculum. State governments, the national government, and independent testing companies such as College Board have developed tests that reflect curriculum standards or goals that students are expected to master in their grade levels, courses, or college-level courses. Though these standards vary based on the testing entity that created them, they are designed to reflect grade-appropriate learning goals.

Students take assessments throughout and at the end of the school year (or often at the end of a semester for shorter courses). Formative assessments measure student learning while the students are still enrolled in the class. Summative assessments measure what students know by the end of the class. Formative assessments allow teachers and school administrators to identify the concepts or skills students need additional support on and what they are mastering. Therefore, these educators can promptly implement interventions, such as new materials, lessons, or programs, to help their current students meet their goals. On the other hand, summative assessments can help teachers understand how prepared their students are for the next

milestone (e.g., grade, course, post-high-school journey) and allow educators to adjust expectations and lesson plans for the incoming group of students.

Who Represents the "Standard" in Standardized Testing?

Despite the idea that standardized tests are meant to measure student ability on an equal and fair level, there is a consistent trend of Black students scoring lower on assessments than White students despite time and assessment. I explore previous research on this "Black-White gap" and provide a brief historical background on how and why these disparities continue to be found.

Researchers and educators evaluate student learning progress and gauge school and teacher quality by comparing various groups of students (e.g., grade level, race, gender, region) based on their assessment results. These types of comparisons allow stakeholders to gain more information about what subjects, concepts, or skills students excel in but also allow differentiation for academic intervention to address students' growth areas. However, these comparisons often reveal significant underperformance of students of color, students with disabilities, and multilingual students, especially nationally (Robinson, 2010).

Black students have been reported to underperform compared to White students for decades. In Hedges and Nowell's (1999) comprehensive longitudinal study of the Black-White achievement gap on six standardized exams, the researchers revealed that between 1965 and 1992, White students' consistently had higher average test scores than Black students on all six assessments. However, the gap seemed to be shrinking as time progressed. Though the study may have indicated shrinking**,** this disparity is still ever-present in standardized testing reports. In a study conducted by Robinson (2010) using scores from three nationally representative exams,

the researcher found that the disparity in math reading test scores between Black and White students is most significant when students are in the fourth grade. The most recent report from the National Assessment of Educational Progress (or the Nation's Report Card) outlines this finding. In 2022, Black fourth-grade students had an average reading score of 199, while White fourth grade students had an average score of 227. Black students had the second-lowest average reading scores of all racial groups tested.

This racial disparity in test scores could be attributed to biased content, scoring, and interpretation of standardized assessments. Much of the material in standardized tests is based on White, middle-class cultural norms (Philips, 2006; Stewart & Haynes, 2016; Sireci & Randall, 2021; Randall, Poe, & Slomp, 2021). Standardized assessments are rooted in intelligence testing, which was often used to push racist agendas and propaganda. Taking French psychologist Alfred Binet's development of the intelligence quotient (IQ) test to screen young children for developmental disabilities, North American eugenicists' including Harry Goddard and Robert Yerkes, used this idea to justify their beliefs that intelligence was inherent and that humans could be ranked by gender, ethnicity, and class (Au, 2016; Stewart & Randall, 2021) using academic tests. Though the creators and early users of such tests would claim objectivity in the design and interpretations of these tests, they were used to justify racist scientists' claims that students of color (particularly African-American students) and low-income students are inherently less intelligent than White elite students. Therefore, large-scale assessments, especially for college admission into Ivy League schools, were designed using cultural and intellectual content created and circulated in majority White and elite spaces.

Even more recently, with the implementation of the No Child Left Behind (2001) Act and similar legislative decisions, *meritocracy* has become another justification for discriminatory testing practices in education. Au (2016) defines meritocracy as:

> [An ideology that asserts] that, regardless of social position, economic class, gender, race, or culture (or any other form of socially or institutionally defined difference), everyone has an equal chance at becoming 'successful' based purely on individual merit and hard work.

Such a claim would insinuate that the American educational system is just and equal for all and that all students can access the same resources for success. These ideologies completely put the fault of Black students' low test scores on the students, pushing the idea that these students are not capable nor deserving of the same academic rewards and higher education opportunities as their White peers. These same principles ignore the legislative and institutionally racist educational practices that launch economically privileged White students ahead and keep Black students behind (Stewart & Haynes, 2016; Au, 2015; Sireci & Randall, 2021; Perez, 2002). A single test score does not reflect the background of a student nor the student's potential. The idea of meritocracy is an excuse to ignore how White students disproportionally benefit from standardized testing. The adverse outcomes that African-American students and majority Black schools endure show that low performance on standardized tests have far-reaching effects on African-American students' academic trajectories that are impossible to ignore.

Adverse Outcomes of Testing on Black Students

This section discusses the adverse effects that many Black students experience when taking standardized tests that could affect their performance. In addition, I discuss the influence that standardized testing can have on their academic trajectory.

Stereotype Threat

Black students may experience stress when preparing for and taking high-stakes exams and after test administration, which manifests in various adverse psychological outcomes. Steele and Aronson (1995) coined the term *stereotype threat* to explain the "social-psychological predicament" that occurs when a person is aware of the negative stereotypes about their cultural or social group. A person experiencing stereotype threat may believe that people who are not part of their identity group perceive her as fitting the negative images and beliefs of her in-group. For example, a Black student may know that Black people are stereotypically characterized lazy and unintelligent. This student will begin to wonder if their teachers and classmates (especially those who are non-Black) see them in that same way. Therefore, Steele and Aronson mention that this recognition of the negative portrayals becomes a threat when a person is hyper-aware of the potential stereotyping, changing how the person thinks about themselves and how they behave in situations when these stereotypes are more salient.

The researchers evaluated Black and White college students' performance on some verbal test items from the Graduate Record Examination (GRE) assessment. Based on the condition that they were in, participants were told that the items were either diagnostic of their intelligence, a non-diagnostic problem-solving task, or a non-diagnostic but challenging task. The researchers found that Black students in the diagnostic condition performed significantly worse than White students in the same condition and Black students in the other two groups. The authors conclude that this finding supports the activation of stereotype threat for African-American students, which affected the way that these students took the exam.

Educators and test designers must consider how stereotype threat can manifest in African-American students' test-taking behaviors. As Steele and Aronson (1995) evaluated in

their studies, Black students in the diagnostic condition more commonly left item responses incomplete or skipped compared to students of any race in the other two conditions. Whether students missed test questions because of stereotype threat or lack of knowledge of the material, test evaluators will evaluate the lack of response in the same way. Therefore, these may lead to evaluators suggesting programs, workshops, or institutions that may not directly address their academic needs. For Black students, when these test responses are paired with an educator's negative perception of them, it can be more challenging for these students to "work" their way out of these positions.

Morris (2004) investigated the daily operations of two predominately Black urban public elementary schools in Atlanta and St. Louis that were notable for their students' outstanding performance. The researcher states that the principals of these schools had to continuously encourage their Black students and their faculty to work "twice as hard" when preparing for standardized tests. Though these tests did not always reflect what students were learning in their classrooms, their testing performances were published in local newspapers, often being compared to that of a suburban, White, middle-class school in the same areas. Therefore, the teachers were encouraged to spend extra time preparing their students for the standardized exams.

The educators at the two schools recognized that if the students at their schools reflected low performance on the tests compared to the White schools, they would contribute to the stereotype of African-American students being less intelligent than White students. Therefore, the students at these two schools are not only aware of the potential biases that their non-Black educators may have of them, but they are now aware of the judging eyes of the newspaper journalists and consumers who know nothing of their school experiences and education in the

first place. These students become unwilling bearers of an unfair and deeply rooted stereotype about African-American intelligence while having to sacrifice time dedicated to genuine learning experiences for testing preparation.

Loss of Genuine Learning Experiences in the Classroom

Participating in an education system that values test scores rather than genuine student engagement and learning can be incredibly damaging for African-American students. In many cases, educators must implement a curriculum that centers student achievement on standardized tests more than encouraging authentic learning experiences for students. Student interests, preferences for learning, and encouragement for creativity are compromised or wholly eliminated in preference for the test-taking format (Thompson & Allen, 2012).

Paulo Freire (1970) states in his pivotal book *Pedagogy of the Oppressed* that teachers rely on the "banking method of education" in the classroom. Teachers behave as the all-knowing carriers of information, while students are just empty receptacles where information is deposited. In this system, students' knowledge, as well as the quality of the teacher, is emphasized by how well students can recite the information they are taught and how little the students question what is taught. That is, students are led to believe there is a correct way of knowing and demonstrating intelligence. In turn, these standardized tests evaluate a specific and rigid way of demonstrating competency.

Research indicates that many school administrators and teachers believe that nationally-normed standardized assessments are among the least valid indicators of student learning. Given that NLCB and similar legislation place a massive weight on high test scores representing school reputation, resource acquisition, teacher job positions, and more, teachers are required to "teach

to the test." Therefore, many educators pivot their lessons and engagement with students toward standardized assessment preparation due to the No Child Left Behind Act of 2001 (Gunskey, 2007; Perez, 2002). Teachers lose interest in their lessons due to a lack of control over their curriculum, making many students disengage. This observation is especially accurate for students of color, where many aspects of their identities and cultures are rarely embraced in the general education culture in the first place. The less these students engage, the less prepared they may be for standardized exams. Therefore, these students may be misidentified into lower academic tracks.

Overidentification Into Lower Academic Tracks

Much research indicates that students of color, especially Black students, are disproportionately identified to be in lower academic tracks and special education classes than White students (Skiba, Knesting, and Bush, 2002; Stewart & Haynes, 2016; Choi, 2020). Though it must be acknowledged that various factors influence the placement of students onto various academic tracks, such as student grades and teacher recommendations, standardized testing can play a crucial role in the type of classes that students are placed in through the *tracking* process. Tracking, or ability grouping, separates students into specific ability-level classes based on their presumed proficiency in a subject (Stewart & Haynes, 2016). Since student scores on standardized exams are used to make inferences about students' abilities, these scores determine if students will be placed into the advanced, regular, or remedial tracks. However, as the information in the previous sections outlined, many educational exams exhibit a positive bias toward White students compared to Black students.

White students, especially those from higher education backgrounds, are more likely to be placed in advanced track classes in high school based on standardized exam scores compared

to Black students. In that regard, students in the higher-track classes receive a more challenging curriculum and thus a greater opportunity for academic achievement than those in the remedial or special education tracks (Hallinan & Kubitschek, 1999, Stewart & Haynes, 2016). Therefore, biased practices in testing can deprive Black students of quality education and opportunities for substantial growth, such as substantial college preparation.

Why Culturally-Relevant Assessment?

Recognizing the prevalence of African-American student underperformance on standardized testing compared to White students, examining what types of content empower Black students and how these considerations can be used in the assessment space is necessary. This section introduces the concept of culturally-relevant pedagogy (CRP) as a framework for culturally-relevant assessment and fair testing practices.

Culturally-Relevant Pedagogy (CRP)

As the previous studies have shown, the traditional assessment framework often puts Black students and other students of color at a disadvantage compared to White students. Black students often participate in an educational system where Eurocentric intellectual capital is prioritized on standardized testing. This capital is "privileged," indicating that students from White, middle to high-class backgrounds will likely have consistent access and exposure to this material outside of school through their home, community, and social group ties.

Further, these students are more likely to benefit from these assessments than those not in this group, such as having increased college acceptance rates (Stewart & Haynes, 2016; Perez, 2002; Au, 2016). Black students have an unfair advantage in testing that has little to do with their ability or amount of knowledge, as these standardized tests are designed to measure. Therefore,

test designers, researchers, and educators must change these exams' structure, administration, and use to benefit all students.

There are many definitions of fairness when considering standardized testing. Sireci and Randall (2021) discuss how the definitions of fairness in testing have changed. Overall, however, fairness can be considered as construct validity, or ensuring that the content of an assessment measures what it is meant to measure, like examinee ability. Increasing construct validity can help reduce bias in scoring and interpretation that lead to increased benefits for one group of students over others or reinforce negative stereotypes about marginalized students. Further, the researchers mention that throughout the history of national testing standards, testing agencies have significantly become interested in developing equitable testing material that eliminates racial/ethnic, linguistic, and ability status biases. These considerations are essential when considering ability grouping and college acceptance. Researchers and educators have explored ways to include more culturally sensitive and relevant material in assessments that allow students of color to use learning strategies and experiences outside White mainstream culture.

One pivotal framework to guide educators through culturally-relevant considerations concerning Black students is the culturally-relevant pedagogy (CRP) framework. Ladson-Billings (1995) presents CRP as a framework to encourage teachers in predominantly African-American classrooms to incorporate relevant and meaningful material from Black culture into their lesson plans and activities. This task aims to encourage and empower Black students to embrace their cultural identities and develop critical perspectives and solutions to the sociopolitical issues they face**.** This framework is critical in classrooms with non-Black teachers and majority-Black students.

Further, Ladson-Billings (2014) encourages users of CRP to recognize that culture is dynamic and changes through time, geographic context, and social context. Incorporating Black culture, ways of knowing, and experience in the mainstream curriculum is often a novel phenomenon in spaces where White culture is the norm. In many cases, Black students are directly and indirectly encouraged to assimilate into White culture. Dr. Jennifer Randall (2021) reflects on these experiences with her White 9th grade English teacher:

> I trusted her implicitly and followed all her directions — even when she told me — with absolute sincerity—that I simply had to change the way I spoke and wrote. She explained to me that I had to learn to communicate in standard English— that "ain't" and "gunna" and "finna" was the language of the ignorant and that I was too smart to use that type of language. (p. 594).

Teachers and others whom students look to as educational authority figures are also the gatekeepers on how students learn to operate in the world. Whether implicitly or explicitly, these individuals teach students what they deem valuable. When students of color do not see their experiences, languages, and learning patterns reflected in these lessons - or even worse, when these aspects of culture are corrected (as Dr. Randall's anecdote shows), educators reaffirm the myth that White American elite culture is the most valuable.

Culturally-relevant pedagogy, at last, puts students of color at the forefront of pedagogical considerations. Implementing such a framework requires users to recognize institutional racism within the United States educational system (Randall et al., 2021; Scott, 2023; Morris, 2004). As Scott (2023) summarizes about culturally and linguistically diverse (CLD) students in the traditional American education system, "…when students from CLD backgrounds are unable to meet academic and behavior expectations, teachers and school personnel effectively consider this to be a problem of the student and family as opposed to the integrity of the education system, learning environment, education program, or a combination of these factors." Educational

stakeholders must recognize how historical legislations and social practices of segregation, separation, and racial violence still exist in modern educational spaces. Most importantly, educators must understand how these factors influence the way African-American students are examined, tested, and evaluated for educational outcomes (such as attending college).

Approaches to CRP in Assessment

This section explores various approaches to implementing culturally-relevant material into assessments. I outline the main principles of culturally-relevant assessment and expound on how these can be incorporated into assessment design, content, and interpretation.

Implementing CRP in assessment requires a shift in the paradigm of standardized assessment design, implementation, and evaluation. Therefore, students of color must be empowered to implement culturally-relevant assessments. In turn, stakeholders must recognize that:

1. CLD students have cultural, familial, communal, and linguistic norms that influence how these students demonstrate competence on assessments (Randall et al., 2021; Yosso, 2005). These ways of knowing are not hindrances to student ability but assets. These practices should be used to inform test design and score interpretations.
2. Topics of race, racism, and culture need to be handled with sensitivity and care. Since the White American identity is presented as the norm in mainstream education, the histories, experiences, languages, and intellectual capital of students of color are often repressed, corrupted, or removed. Therefore, culturally-relevant assessment should be used to prioritize and empower the voices of people of color, especially where culturally-specific information is concerned (Stewart & Haynes, 2016; Seneca College, Humber College, Kenjgewin Teg, Trent University, & Nipissing University, n.d). Further, these topics should be introduced to develop students' sociopolitical awareness.

3. These practices are *culturally-sustaining* or implemented to preserve the ways of knowing that students of color have (Randall et al., 2021; Scott, 2023). Therefore, engaging in culturally-relevant assessment should be a common occurrence. Researchers and educators should work with communities of color to improve culturally-relevant assessment material.

As the previous section highlights, the experiences of Black students and other students of color are rarely ever subject to in-depth focus and investigation in mainstream education. Students of color are not the norm in the classroom, nor are they the norm in standardized testing. To assign a specific way of learning and demonstrating knowledge as "standard" in standardized testing is to exclude students who do not fit that mold. This racism puts students of color at a disadvantage and requires them to assimilate into White mainstream culture to be deemed academically apt. Further, this standard often ignores the *cultural capital* (the skills, knowledge, and assets from a particular culture group) that non-White students have.

Centering on the experiences of Hispanic and Latinx students, Yosso (2005) specifies several forms of cultural capital that students of color acquire through their familial and communal ties, such as *linguistic capital*. The researcher defines linguistic capital as "the intellectual and social skills attained through communication experiences in more than one language or style" that multilingual students develop through interactions with cultural members, culturally-relevant art and media, and other forms of linguistic expression. However, traditional exams often repress these skills and knowledge favoring White, English-dominant literature and contexts. Since White members of the socioeconomic elite value this language, multilingual students are often disadvantaged on exams since they may not engage with this information often outside of their school and testing spaces.

Therefore, implementing language from communities of color is essential to culturally-relevant assessment. Under Randall et al.'s (2021) sociocognitive lens of assessment justice, students would explore, use, and engage with languages from marginalized communities by exploring the histories, purposes, and uses within these communities. This lens requires that these languages are addressed as formal languages and that students can critically discuss the role that racism plays in the devaluation of such languages, such as African American Vernacular English (AAVE). Therefore, not only can multilingual students use the communication skills they have developed through formal and informal interactions, but all students have the opportunity to become effective communicators with members of other cultural groups, especially outside of their school communities and situations.

As researchers of CRP and similar frameworks recognize, the histories and structures of racism within the United States play a prominent role in how students of color interpret, engage with, and express comprehension in academic spaces. Also, in many aspects, the narratives of students of color are often repressed to favor the histories and perspectives of the dominant social group. Therefore, implementing topics of race and racism is crucial to developing culturally-relevant assessments. Such topics should prioritize the narratives of communities of color or allow the option for students of color to discuss and critically engage so that these students can develop their sociopolitical awareness (Ladson-Billings, 1995; Randall et al., 2021; Yosso, 2005).

Given these guidelines, test designers should be careful to introduce culturally and racially-laden topics with care to not continuously victimize marginalized communities (Stewart & Haynes, 2016; Seneca College, Humber College, Kenjgewin Teg, Trent University, & Nipissing University, n.d.) For example, imagine an eighth-grade history exam with social

justice movements as a main topic. For the short essay excerpt of the exam, all students are asked to read a short passage about the 19th Amendment and respond to the prompt: “How did the 19th Amendment benefit American women?” This example is limited in cultural relevancy and sensitivity, as the 19th Amendment did not wholly benefit African American women or other women of color under Jim Crow Laws. Such a passage would limit the perspectives of non-White women who participated in the women’s suffrage movement. However, a more inclusive, culturally-relevant example would allow students of various backgrounds to offer their perspectives and knowledge on the topic. For example, students could be asked to read excerpts from interviews with leaders from contemporary, minority-led social justice organizations (e.g., Black Lives Matter, Indigenous Land Rights movements, immigrant justice movements) and be prompted to write a short response about their excerpt of choice.

Lastly, culturally-relevant assessment practices should be treated as something other than a one-time consideration. This framework should not be used by test designers to check a box for racial inclusiveness but to be used to shift the paradigm of testing norms. Therefore, developing culturally-relevant materials should be a consistent practice. These materials are made to preserve the ways of knowing from students of color and ensure they can fairly engage with testing material using all of their strengths, not just those valued by White elite academics (Randall et al., 2021; Scott, 2023). This is an even more critical consideration as culture changes throughout generations, so test designers need to consistently work to alter and remove outdated or culturally inaccurate references used in assessments (Ladson-Billings, 2014).

Regarding implementation for Black students, I want to clarify that incorporating CRP and culturally-relevant assessment practices does not indicate that all Black students’ experiences and cultures are monolithic. Instead, engaging in CRP is a start to prioritizing the

voices of the students that make up the African diaspora. These students are allowed control of their narratives and how their stories, histories, and interests are represented in the educational system, empowering them to be more than passive receptacles of privileged, often inaccessible, and irrelevant information. Recognizing this is also not to claim that CRP is the ultimate solution to increased fairness in testing for Black students or that every Black student will respond positively to this implementation. However, it is worth recognizing the researchers that have explored CRP and similar frameworks throughout the decades that have seen increased engagement and positive academic outcomes for Black students (Ladson-Billings, 1995, 2014; Morris, 2004). Regardless, an approach to decentralizing a Eurocentric focus in education to allow students of diverse backgrounds to study and contribute effectively is a step in the right direction towards making the United States educational system safe and fair so that all students can succeed.

Assessment Modeling

This section introduces the most common assessment model framework, item response theory (IRT), compared to the framework of interest for this study, diagnostic classification models (DCMs). *Using an example of a fourth-grade reading exam, I give a general overview of the use and function of each framework and give an example of a model for each. Finally, I briefly comment on how DCMs can be used in culturally-relevant assessment and discuss the purpose of the simulation and empirical analysis I conduct for this study.*

Classical Test Theory

Educational stakeholders use test scores to make assumptions about students' ability and understanding of the subject they are testing on. Let us imagine that all the fourth-grade students in an urban elementary school are taking the North Carolina End-of-Grade (EOG) reading exam with 100 test *items* or test questions. This test is designed to measure student proficiency on grade-specific reading standards. Since reading ability is a *latent* trait or a characteristic that cannot be measured by visual observation, examinee scores on the items provide psychometricians a way to measure and interpret this group's ability.

The most fundamental and widely-used framework for scoring assessments is *classical test theory* (CTT). Under CTT, it is assumed that an examinee's score on an assessment can be determined using Equation 1:

$$Observed\ Score = True\ Score + Error \quad (1)$$

Therefore, the observed score that an examinee receives is the number of items that the student answered correctly. CTT assumes that an examinee's true score is what this student would receive on the test under perfect testing conditions with their prior ability. For example, let us assume that an examinee is expected to correctly answer 85 out of 100 questions on the reading

exam, having a true score of 85. Instead, she receives a score of 82. This score means that she answered 18 items incorrectly instead of the expected 15 items, generating an error of 3.

Item Response Theory (IRT)

Though these conclusions about student scores are straightforward under CTT, psychometricians are limited in making comparable interpretations about groups of students because CTT assumptions are specific to the test and the group of students being measured. However, psychometricians can use *item response theory* (IRT) to compare larger groups of examinee scores along a standardized scale. IRT refers to a family of statistical models that allow psychometricians to examine student ability as a continuous trait. Examinee ability (referred to as theta, θ) is calculated for each examinee. The examinee is pinpointed along a standardized continuum from low ability to high ability. Typically, this scale has a mean of 0 and a standard deviation of 1. Because the scale is standardized, results from data using IRT are not sample- and item-dependent as they are in CTT.

IRT psychometricians analyze all examinees' abilities by analyzing the properties of the test items and their *item response pattern* or the pattern of items that students answered correctly and incorrectly. The properties of the items include item discrimination, item difficulty, and the possibility of students obtaining a correct answer by guessing. These properties are referred to as item parameters. The *a*-parameter is item discrimination, or how well an item differentiates between examinees of varying abilities. The *b*-parameter, or *item difficulty* parameter, evaluates how easy an item is. Item difficulty allows psychometricians to determine the ability level that examinees need to have a probability of 0.50 (or a 50% chance) or above of answering an item correctly. The *c*-parameter is the guessing parameter or the likelihood that a student of very low ability will answer an item correctly based on pure chance.

These parameters can influence the way that examinees score on an item, which would be scored dichotomously as either correct ("1") or incorrect ("0"). Because of these bounds, the student likelihood of a correct response is calculated as a probability. This likelihood is a function of student ability and the item parameters examined in an IRT model. Therefore, the likelihood of a correct response is denoted as a percentage. There are four commonly-used IRT models: The One-Parameter Logistic (1PL) Model, the Rasch Model, the Two-Parameter Logistic (2PL), and the Three-Parameter (3PL). Psychometricians use an *item response function* (IRF), specifically an *item characteristic curve* (ICC), to visualize these parameters when analyzing the items in an assessment. For this study, I will discuss the Rasch model, as Garth-McCullough (2008) used this model in her empirical analysis.

In the Rasch model, examinee probability of a correct response is estimated using the difficulty (*b*) parameter while setting the discrimination parameter (*a*) equal to 1 for all items, dropping it from the equation entirely. Therefore, the Rasch model determines the examinee's probability of a correct response by assessing item difficulty, assuming that all the assessment items differentiate student ability similarly. The guessing *(c)* parameter is not included. Therefore, the equation for the Rasch model is:

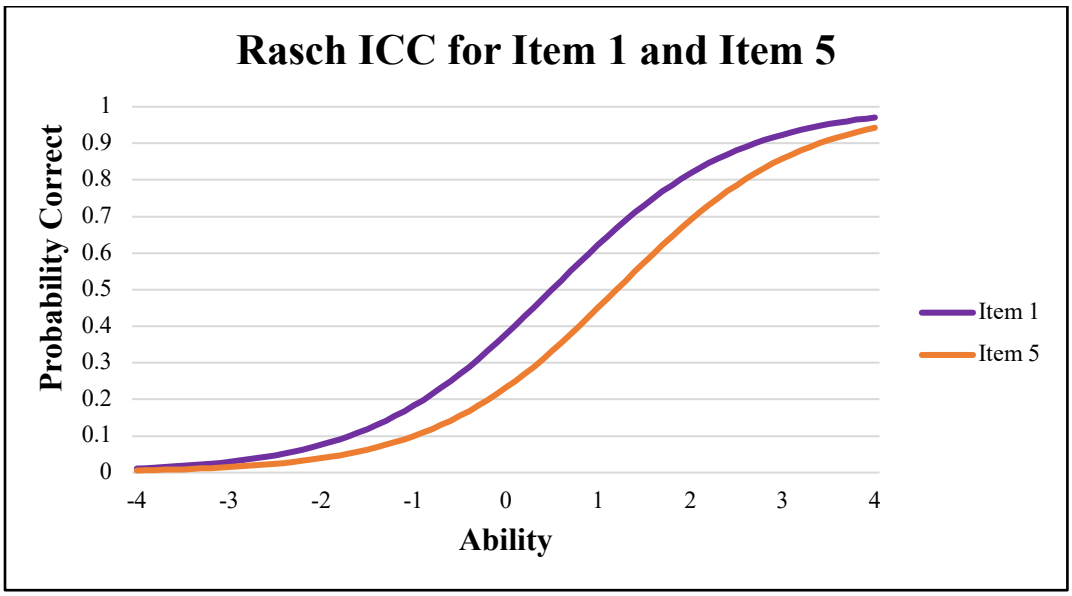
$$P\left(X_{ij} = 1 \middle| \theta_j, b_i\right) = \frac{e^{\theta_j - b_i}}{1 + e^{\theta_j - b_i}} \tag{2}$$

The left side of the equation is read as the probability *P* that examinee *j* will answer item *i* correctly (or that $X = 1$), given the examinee's ability θ and the difficulty of the item b. Since the Rasch model assumes that all items have the same discrimination value, examinee's total score on an assessment can be determined by the number of items that the student answers correctly.

Let's compare item 1 and item 5 in our fourth-grade North Carolina reading EOG. These items would be displayed in an ICC like Figure 1. θ is displayed on the x-axis, while examinee probability of a correct response lies on the y-axis. For item 1, $b = 0.5$ and for item 5, $b = 1.2$. The curves for both items have the same slope, indicating that item discrimination is the same for both items. The difficulty parameter affects the position of the curve along the x-axis. As discussed previously, item difficulty is identified by the ability level that a student would need to have a 0.50 (or 50%) chance of answering an item correctly. We see that for item 1, examinees need an ability level of 0.5 to have a 50% chance of a correct answer. For item 5, examinees need an ability level of 1.2 to have the same probability. Therefore, we would conclude that item 1 is easier than item 5 because examinees of lower ability have a higher probability of answering item 1 correctly than they do item 5. Also, the more examinee ability increases, the higher the probability of a correct answer on an item. This is expected of how students should perform on an item; in theory, the more knowledge that an examinee has about an item, the higher the likelihood that the examinee will answer the item correctly.

Figure 1

Rasch Model Item Characteristic Curve (ICC) for Item 1 and Item 5



IRT models are useful for psychometricians who are interested in how students of various abilities perform on various items based on the characteristics of the items. Based on the function of these models, IRT is best-suited for *norm-referenced* assessment contexts, or those contexts where interpretations about an examinee's ability are made compared to the other examinees who took the test. However, using information from an ICC alone cannot tell us what concepts each of these students are proficient on and where they differ, just essentially how *much* knowledge an examinee has about an item. In that case, another modeling framework would be needed to investigate this comparison further. This is where diagnostic classification models (DCMs) become useful.

Diagnostic Classification Models (DCMs)

Diagnostic classification models (DCMs) are a family of statistical models in which examinee ability is measured as a categorical trait instead of a continuous trait. Instead of placing students along a scale in comparison to each other, students are placed into categorical

classifications based on their proficiency on the *attributes,* or the skills and concepts, that an assessment measures. Traditionally, there are two mastery classifications that examinees are placed in: *master* and *non-master*. Proficient and non-proficient are also used as the class labels, respectively. In this context, examinees are given a *classification status* according to every attribute that the assessment measures. Masters are numerically represented as a "1" while non-masters are numerically represented as "0." As in IRT, this status is determined by an individual examinee's pattern of responses and how the item parameters for the model are constrained.

Let's say that the fourth-grade EOG included Reading for Literature and Reading for Informational Text as the two attributes on the test. There are simple items on the test in which only one attribute is tested in an item, as well as complex items, where multiple attributes are tested in that item. Psychometricians use a tool known as a *Q-matrix,* which is a table that specifies what items measure which attributes. The number of rows for a Q-matrix is determined by a vector of length A that looks like $[q_1, \dots, q_a, \dots, q_A]$, where A represents the number of attributes in an assessment. In the cells of each table, a 1 would be used to denote that the item measures the ath attribute and a 0 would indicate that the item does not measure that attribute.

Like IRT, an examinee's likelihood of a correct response on an item is determined by an examinee's item *attribute profile* and the characteristics of the assessment items. Most commonly when DCMs are used in educational contexts, examinee responses are binary, where an incorrect response would be scored as a 0 and a correct response scored as a 1. Depending on the type of DCM that a psychometrician uses, student probability of a correct response on an item is determined by how the main effects and interaction effects of an item are constrained. Main effects represent the boost in probability of a correct response that an examinee receives if

they are a master on a singular attribute. Interaction effects represent a boost in probability for examinees that are a master of multiple attributes.

The four main DCM models are the DINA, DINO, CRUM, and LCDM models. Under the DINA model, the main effects are constrained to 0, indicating that only masters of the interaction term will receive a boost in probability in answering the item correctly. Examinees who are non-masters for all attributes or only masters of one attribute have the same likelihood of a correct response. On the other hand, the DINO model constrains all main effects to be equal and the interaction effects are not affected by negative main effects. Therefore, examinees only need to be a master of one attribute in the item to receive a boost in probability of a correct response; masters of a single attribute and masters of the interaction terms have the same likelihood of a correct response. Finally, under the CRUM, the interaction effect is constrained to be 0, indicating that masters of each attribute that an item measures have no extra boost in the probability of a correct response.

For the focus of this study, I will be using the Log-Linear Cognitive Diagnosis Model (LCDM), which is the least restrictive of the three main models. The LCDM includes all main effect and interaction terms. The specifications for this model are the main effects are constrained to be greater than 0 and the interaction terms are constrained to be higher than any negative main effects. That is, examinees who are masters of an attribute will always have a higher probability of a correct response on an item than non-masters. Further, masters of multiple attributes will always have a higher likelihood of a correct response than examinees who are masters of only one attribute. Let's examine how the item parameters for a two-attribute model would be specified under the LCDM:

$$P(X_{ei} = 1|a_e) = \frac{exp\left(\lambda_{i,0} + \lambda_{i,1,(1)}a_1 + \lambda_{i,1,(2)}a_2 + \lambda_{i,2,(a_1,a_2)}a_1a_2\right)}{1 + exp\left(\lambda_{i,0} + \lambda_{i,1,(1)}a_1 + \lambda_{i,1,(2)}a_2 + \lambda_{i,2,(a_1,a_2)}a_1a_2\right)} \quad (3)$$

Firstly, the left side of the equation reads the probability *P* that examinee *e* will answer item *i* correctly, $X = 1$, given the examinee's mastery status for a specific attribute *a*. For the right side, let's discuss the subscripts of each of the terms. As on the left side, *i* is the item number of the assessment. The first numerical subscript is the specification that the term is either an intercept (denoted by a "0"), a main effect (denoted by a "1"), or an interaction term (denoted by a "2"). The intercept represents the likelihood of a correct response for non-masters. The main effect reflects the boost in probability that masters of attribute a will receive. The interaction term represents the boost in probability that the masters of each attribute that the item measures will receive. The parenthetical subscript specifies which attribute the term is representing (e.g., (1) represents attribute number 1). For interaction terms, each attribute will be specified and separated by a comma (e.g., $(a_1 a_2)$).

For items measuring more than two attributes, a main effect term for each additional attribute would be added. Psychometricians would have to add interaction effects for each combination of main effects as well. If there is an item with only one attribute, only the intercept and main effect term would be included in the equation.

Remember that items are scored dichotomously and that examinees receive a mastery classification per attribute based on how they score on items measuring those attributes. An attribute profile is a denotation the various master and non-master classifications that can exist per attribute in an item. The total number of attribute profiles for an item can be calculated using 2^A, where *A* is the number of attributes in that item. These profiles would be denoted using $a_e = [a_{e1}, a_{e2}, \ldots . a_{eA}]$, where $a_{eA} = 1$ if examinee *e* is a master and a 0 if not.

Item Parameters

Equation 3 is a way to represent item parameters, or the likelihood that an examinee will answer an item (or test question) correctly based on their attribute proficiency. Again, for an LCDM model, the main effects (or examinee proficiency on single attributes) are constrained to be greater than 0 while the interaction effects (or examinee proficiency on multiple attributes) are constrained to be greater than any negative main effects. To visualize these parameters, DCM psychometricians use an *item characteristic bar chart* (ICBC).

Again, since students are categorized into masters and non-master categories instead of being placed along a continuum like in IRT, a bar chart is a more appropriate visualization than a line chart. In Figure 2, we have an ICBC that represents item 5 in our reading exam, which measures both the Reading for Literature attribute ($A = 1$) and the Language attribute ($A = 2$). Therefore, there would be $2^2 = 4$ attribute profiles for this item. As this assessment is modeled using the LCDM framework where the interaction term is constrained to be higher than any negative main effects, examinees who are masters of both attributes (i.e., those with an attribute pattern of [1,1]) have the highest probability of a correct response on an item. Further, examinees who are masters of only attribute 2 have a higher likelihood of answering an item correctly than examinees who are only masters of attribute 1.

Let's discuss the parameters of Figure 2 more specifically. To find the probability of a correct response on an item, we have to calculate the logit of a correct response. The x-intercept shows the four attribute profiles that we have for this model: [0,0] representing the non-masters of neither attribute, which is the intercept for our equation, [1,0] representing the main effect for masters of Attribute 1, [0,1] representing the main effect for masters of Attribute 2, and [1,1],

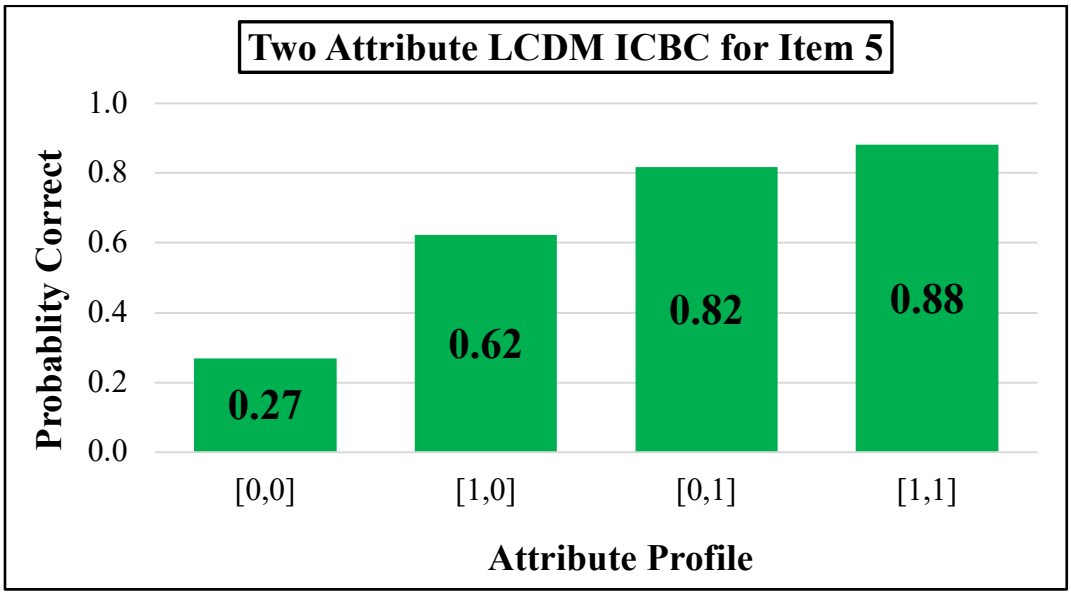
representing the interaction effect for masters of both attributes. For our item 5 example, our parameters estimate for each profile are -1, 1.5, 2.5, and -1, respectively.

Next, we consider the logit of each parameter estimate. Therefore, our intercept has a logit of -1, which corresponds to a probability of 0.27. Since main our main effect estimate represents the boost in probability of a correct response that examinees of Attribute 1 receive, we add the estimate of the intercept estimate and that of Main Effect 1 estimate together, which would be $-1 + 1.5 = 0.5$, which corresponds to a probability of 0.62. We repeat this process using the logits of the intercept and Main Effect 2, which would be $-1 + 2.5 = 1.5$, corresponding to a probability of 0.82. Finally, we combine the parameter estimates of the intercept, Main Effect 1, Main Effect 2, and the Interaction Effect to find the probability for the interaction effect, which would be -1 $+1.5 + 2.5 - 1 = 2$, which corresponds to a probability of 0.88.

Let's interpret these student profile probabilities to understand what these numbers mean practically. In the context of our reading assessment example, the ICBC allows us to visualize the boost in probability of a correct response that examinees who are proficient on the Reading for Literature attribute (attribute profile [1,0] and the Language attribute (attribute profile [0,1]) receive on the exam. As we can see from the ICBC, students who are non-masters of either attribute have the lowest probability of answering Item 5 correctly with a 27% chance of answering item 5 correctly. On the other hand, students who are masters of Attribute 1 have a 62% chance of answering the item correctly, while students who are masters of Attribute 2 have an 82% chance of answering the item correctly. Further, as expected with an LCDM, examinees who are proficient in both skills have an increased boost in probability of a correct response.

Figure 2

Two Attribute LCDM ICBC for Item 5



Person Parameters

Examinee responses for each item is scored and calculated into *posterior probabilities,* which is the likelihood that an examinee is a master for each attribute, respectively. These posterior probabilities represent the person parameters for a DCM. The cutoff value for examinee classification is typically .50. Therefore, if an examinee that has a posterior probability of .50 or higher for an attribute would likely be classified as a master on that attribute. Therefore, that examinee would be given a mastery classification of 1, which represents "master." On the other hand, students with posterior probabilities under .50 would be classified as 0, or "non-masters." However, this value can be set at the researcher's discretion based on their desires to reduce false-positive master classifications and false-negative non-master classifications.

For example, let's explore the posterior probabilities and mastery classifications for two students who took the reading exam, Mya and Bianca. The mastery classification cutoff value is

0.50. As Table 1 shows, Mya has posterior probabilities that are over .50, where $a_1 = .989$ (Reading for Literature) and $a_2 = .768$ (Reading for Informational Text). Therefore, she is given a master classification for both attributes, denoted as 1. On the other hand, Bianca has a posterior probability that are higher than .50 for a_2 only ($a_2 = .496$), making her a master of Reading for Informational Texts. It is worth noting that for Bianca, the posterior probability for a_1 is close to the 0.50 cutoff. A benefit of DCM feedback is that educators can see how individual students perform on specific attributes, which can help educators better understand how to implement effective support to target the skills that students may be lacking. For example, we can see that Bianca seems to demonstrate considerable proficiency on the Reading for Literature attribute, but she may benefit from additional review using test questions or activities targeted at this attribute.

Table 1

Student Mastery Classifications for Reading Exam

Student	a_1 **Posterior**	a_2 **Posterior**	a_1 **Classification**	a_2 **Classification**
Mya	.989	.768	1	1
Bianca	.496	.988	0	1

In conclusion, IRT and DCMs allow psychometricians to measure examinee ability on educational exams and further understand what students are learning in their classrooms. DCMs allow psychometricians to investigate examinee ability according to the skills or concepts that they are proficient in. In terms of test design, DCMs require fewer items on an assessment to be as reliable as tests designed under the IRT framework (Templin & Bradshaw, 2013). This

statement makes sense when considering the overall goals of each model. The goal of IRT models is to locate examinees on a theoretically infinite ability continuum. Because the continuum is so finely-grained, IRT models need many items to accurately and reliably determine a student's position on that continuum. On the other hand, the goal of DCMs is to determine if students have sufficient knowledge on a particular concept or skill, not necessarily how *much* of that skill or concept that the examinee knows. Logically, it wouldn't require a researcher to analyze as many items as it would in IRT to determine if an examinee is proficient on a particular attribute.

Further, the feedback that researchers and educators receive from assessments designed under the DCM context allow stakeholders to understand student's ability at an individualized level. Again, examinees are classified into a master or non-master category for each attribute. Therefore, DCM models are best suited for *criterion-referenced* assessments, or those exams where examinees are required to meet a certain cut-off score to be determined proficient on the skill. With these results, educators can determine the skills that each student is proficient in as well as student's areas of growth where they need additional support. DCM feedback can help educators design more targeted interventions for each student that appropriately addresses each student's needs.

Lastly, researchers who are interested in how examinees perform according to various demographic variables (such as race) can use DCM feedback to explore any patterns in how different groups of students perform on specific attributes. These considerations could be especially interesting when considering culturally-relevant assessment. Using the DCM context, researchers could compare how Black students perform on certain attributes on an assessment with more culturally-relevant elements compared to a traditional assessment. Further, DCM

researchers could explore if there are any covariates or external factors that would affect Black students' ability on these attributes. These reasons are why I use a DCM to investigate Black students' reading comprehension using culturally-relevant material.

Purpose of the Present Study

The empirical data for this study comes from Ruanda Garth-McCullough's dissertation research published in her 2008 article *Untapped Cultural Support: The Influence of Culturally Bound Prior Knowledge on Comprehension Performance*. She explored the relationship between 117 Black 8th-grade students' reading comprehension scores based on their prior reading ability and culturally-relevant prior knowledge. Prior reading ability was categorized as low, medium, and high. Culturally-relevant prior knowledge refers to students' familiarity with language, traditions, social interactions, and more that are commonly exhibited by people of a particular culture.

The researcher selected short reading passages centered around three different cultures: African-American, Chinese-American, and European-American, respectively. Each passage contained over 25 cultural references, including phrases, traditions, and beliefs. The students read these short texts and were administered reading comprehension exams for each respective passage. By conducting an ANOVA analysis using a Rasch model, Garth-McCullough found that, for the African-American passages, medium-level readers with high cultural knowledge had higher reading comprehension scores than high-level readers with low cultural knowledge. In other words, Black students' familiarity with relevant Black cultural knowledge appeared to compensate for their reading ability. This finding is exciting for advocates of culturally-relevant material, as it demonstrates that assessment content that Black students can relate to allows

students of various reading abilities a fair chance to demonstrate competency that may not otherwise be recognized in a traditional exam.

Given the finding stated above, Dr. Garth-McCullough's work became the inspiration for the present study. I chose to conduct a secondary analysis of her work by analyzing the data in a DCM context. By doing so, I wanted to explore how students of with various levels of prior cultural knowledge would be classified despite reading level, especially for African-American-centered texts. Given that the researcher considered student prior reading ability and prior cultural knowledge as covariates in the analysis, I first designed a simulation study to identify the best method of including a covariate in a DCM analysis.

CHAPTER 3

SIMULATION STUDY METHODS

In Chapter 3, I describe the purpose and design of the DCM simulation analysis study and empirical analysis.

Purpose of the Analysis

The main objective of the simulation study is to explore different methods of including a covariate in a DCM estimation. I evaluated the methods in terms of type I error, power, classification reliability, and classification accuracy. For Method 1, the covariate is included into the model estimation of the DCM. The second and third methods are post-hoc methods. For Method 2, I regress the covariate onto estimated examinee posterior probabilities. Lastly, for Method 3, I regress the covariate onto estimated student mastery status.

This study uses the log-linear cognitive diagnostic model (LCDM), which is least restrictive of the DCMs. The LCDM model categorizes examinees who answers an item correctly as a master, regardless of how many attributes the item measures or what attribute the examinee demonstrates proficiency on in that item. Also, since I am using a one-attribute model, there was no need for me to examine the interaction effect between multiple attributes, so no additional restriction was necessary.

Software.

Data were generated and analyzed using R Studio, Version 4.3 (R Core Team, 2023). The models were estimated using Mplus, Version 8 (Muthen & Muthen, 2012-2017).

Study Design

Constant Factors.

Number of Attributes

Since the purpose of this study is to examine the effect of a covariate on classification, I am using a one-attribute model to directly explore this effect. While most applications of DCMs have included multiple attributes, using one attribute allows me to isolate the impact of the covariate and better assess the estimation methods mentioned above.

Base Rates

A base rate is a value that designates the proportion of masters to non-masters on the exam. I will use a base rate of .50, indicating that 50% of examinees will be masters on the assessment. This value is most used in DCM simulation designs (Bradshaw & Madison, 2016).

Manipulated Factors

Sample Size

I test sample sizes of 100, 250, and 500 students to represent low, medium, and high conditions respectively. Though these chosen sizes are relatively smaller compared to other DCM simulations (Sen & Cohen, 2021, Bradshaw & Madison, 2016; Lin, Xing, & Park, 2020), the goal and design of my model is more simplistic than the models in these studies. Because I am testing a one-attribute model, I expect to begin seeing a differentiation in the evaluation metrics at 250 students.

Test Length

I generated assessments with 5, 10, and 15 items. These test lengths represent short, medium, and longer diagnostic assessments.

Effect Size

Effect size is generated as the correlation between attribute classifications and the covariate. I will test an effect size of 0 to evaluate Type I error. I will also test effect sizes of .25, .50, and .75 to evaluate the power of each method.

Factor Conditions

Given that I have four manipulated factors (sample size, test length, and effect size) with various levels (three levels, three levels, and four levels, respectively), I have a total of 36 independent factor conditions in this study ($3_{sample\ size}\ x\ 4_{effect\ size} x\ 3_{test\ length} = 36\ factor\ conditions$). I will hereinafter refer to these conditions as "factor conditions."

Table 2

Simulation Factor Conditions

Sample Size	Effect Size	Test Length
100	0	5
250	.25	10
500	.50	15
-	.75	-

Covariate Conditions

Given that I estimated a covariate into the model estimation directly and using two post-hoc methods, I had two covariate conditions to consider; the "with covariate" condition, which is where the covariate is directly estimated in the model (Method 1), and the "without covariate" condition, which is where I regress the covariate onto examinee posterior probabilities (Method 2) and examinee classification status (Method 3). From hereinafter, the covariate conditions will be referred to as With Covariate for Method 1 and Without Covariate for Methods 2 and 3.

Number of Replications

I generated 100 replications for the With Covariate condition and the Without Covariate condition for a total of 200 replications per factor condition.

Evaluation Metrics

Reliability

Consider the idea that an examinee is given an assessment on one occasion and is later given that same assessment with no memory of the content of the test. Therefore, the student would have no recall of the material on the exam and would be relying solely on their own knowledge as they did on the first occasion. Reliability is a measurement of how often a students' score on one administration of an exam will be the same on a second administration of the exam. In the DCM context, reliability is the quantification of classification consistency, or a value representing how often masters are classified as masters on repeated administrations of an assessment, and vice versa. I will be using the tetrachoric reliability metric for this study (Templin & Bradshaw, 2013). The tetrachoric reliability metric is used when student

performance is measured as a categorical trait. Tetrachoric correlation returns a proportion for what the reliability value would be if the correlation was on a numerical scale.

Accuracy

Accuracy measures how often the model-estimated classifications agree with the generated classifications. It is represented as a proportion of agreement between an examinees' estimated classification status and their true classification status, so a model with perfect accuracy would have a value of 1.

Type I Error

Type I error is the probability of the model rejecting a null hypothesis when it is actually true. In this study, Type I error measures the probability of finding a significant covariate effect when effect size equals 0, which means there is no covariate effect to be detected. It will be represented as the proportion of replications that rejected the null hypothesis when it is actually true over the number of replications that failed to reject the null hypothesis. I will evaluate Type I error by setting the overall significance level for each model at $\alpha = 0.05$.

Power

Power is the sensitivity of a model, or the probability of the model detecting a covariate effect when there is one to be detected. It would be represented as the proportion of replications that failed to reject the null hypothesis. Power will be measured at effect size 0.25, 0.50, and 0.75 for each model.

CHAPTER 5

SIMULATION STUDY RESULTS

The purpose of the simulation was to a covariate in a DCM. I evaluated a direct model estimation method and two post-hoc methods:

1. Method 1; With Covariate – Covariate included in estimation
2. Method 2; Without Covariate – Covariate regressed on posterior probabilities
3. Method 3; Without Covariate – Covariate regressed on mastery classifications

Accuracy

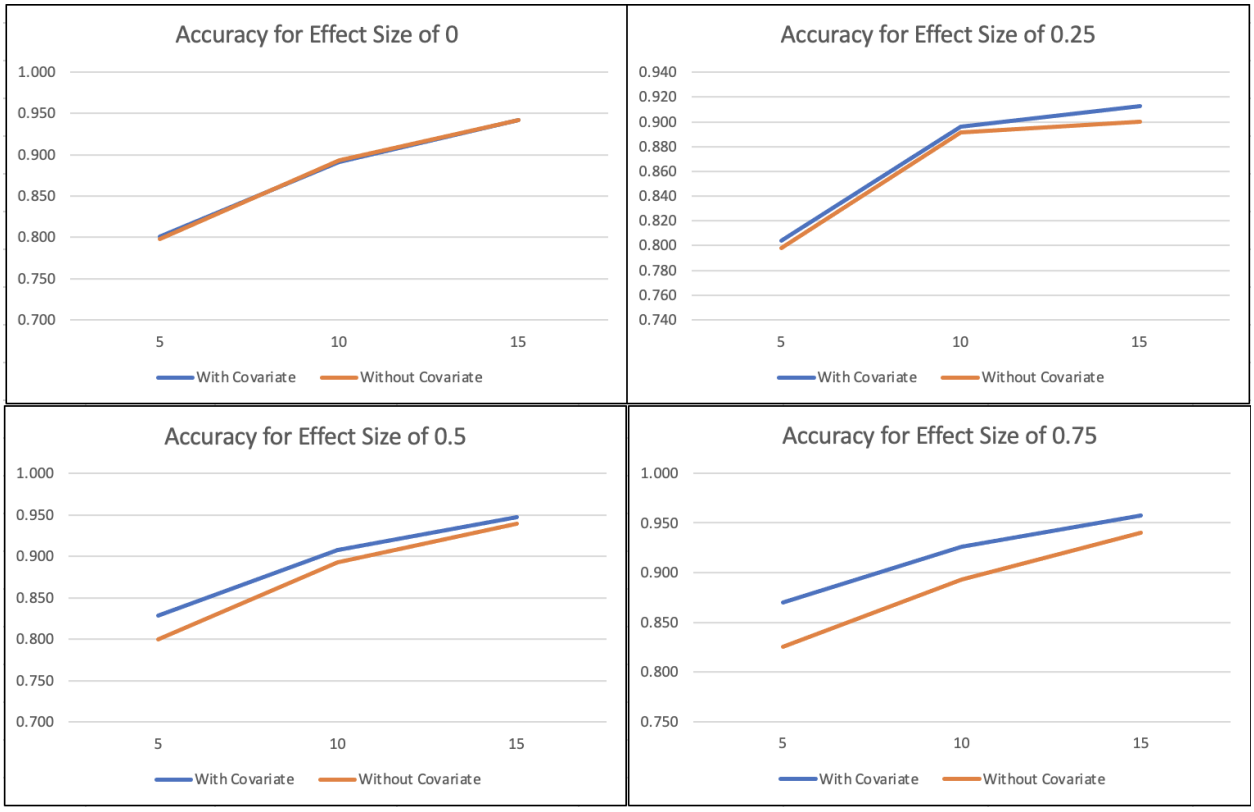
Accuracy measures how often the model-estimated mastery classifications matched the generated classifications. Accuracy was calculated to determine if including the covariate directly in the model (Method 1) would produce higher mean values than including the covariate in post-hoc methods (Methods 2 & 3). Therefore, accuracy was calculated for the With Covariate method separate from the Without Covariate methods. Methods 2 and 3 share average accuracy values.

As predicted, the With Covariate had the highest mean accuracy values for most of the factor conditions ($Min = .801, Max = .958$), though the Without Covariate condition produced accuracy values that were within close range of With Covariate condition ($Min = .798,\ Max = .942$). For both conditions, accuracy increased as test length increased. Per each effect size condition, the 5-item condition produced the lowest accuracy for each method, while the 15-item condition yielded the highest average accuracy. Further, as visualized in Figure 4, as effect size increased, the difference in mean accuracy between each covariate condition became more evident. This finding indicates that as the covariate effect on examinee classifications increases,

the With Covariate produces more noticeable improvement in model accuracy than the Without Covariate condition does.

Figure 4

Mean Accuracy For With Covariate and Without Covariate Condition



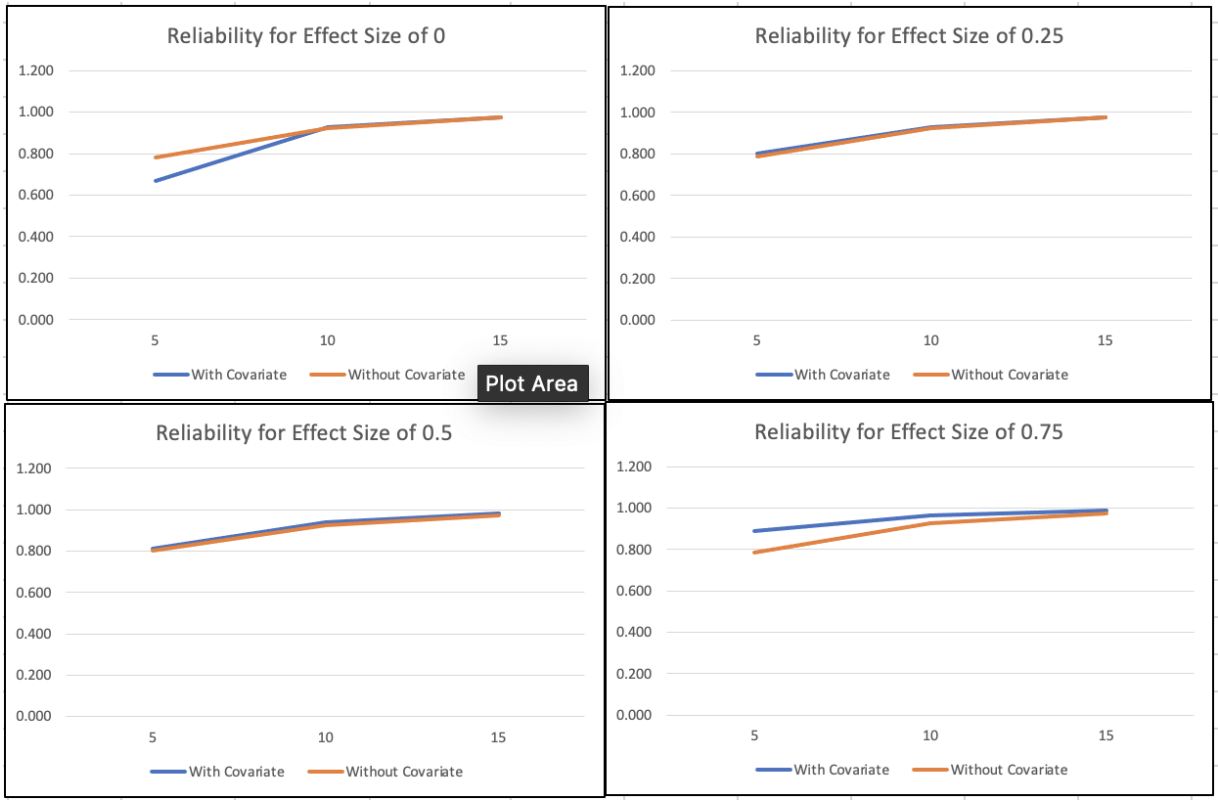
Reliability

Reliability is a measure of examinee classification consistency on repeated administrations of the same test. The minimum average reliability value for the With Covariate condition was 0.671 and the maximum was 0.987. On the other hand, the minimum score for the Without Covariate condition was 0.783 and the maximum was 0.975. Per each effect size condition, the With Covariate condition maintained a slightly higher average reliability score than the Without Covariate as test length increased, though the difference in means would often be tenths of a percentage apart.

Reliability follows an interesting pattern when comparing the With Covariate condition to the Without Covariate condition (see Figure 5). Starting out with the 0-effect-size-five-item-condition, the Without Covariate condition yielded a higher mean reliability value than the With Covariate condition, yet both conditions maintained indistinguishable differences in mean reliability until the .75 effect size condition. There, the With Covariate condition had a higher mean reliability value than the Without Covariate condition for a 5-item exam, though the average values once again became indistinguishable as sample size increased. Overall, I conclude that although both covariate conditions fall within the same range of average reliability, the With Covariate condition leads to slightly higher reliability in estimating examinee classification than the Without Covariate condition as effect size and test length increase.

Figure 5

Mean Reliability For With Covariate and Without Covariate Condition



Type I Error and Power

Type I error was determined by calculating the proportion of replications that rejected the null hypothesis, when the effect of the covariate was 0. Power was calculated as the proportion of replications that rejected the null hypothesis when the effect of the covariate was non-zero. For the explanation of both factors, I compared the values for Method 1, Method 2, and Method 3 respectively. I used a significance level of $a = 0.05$.

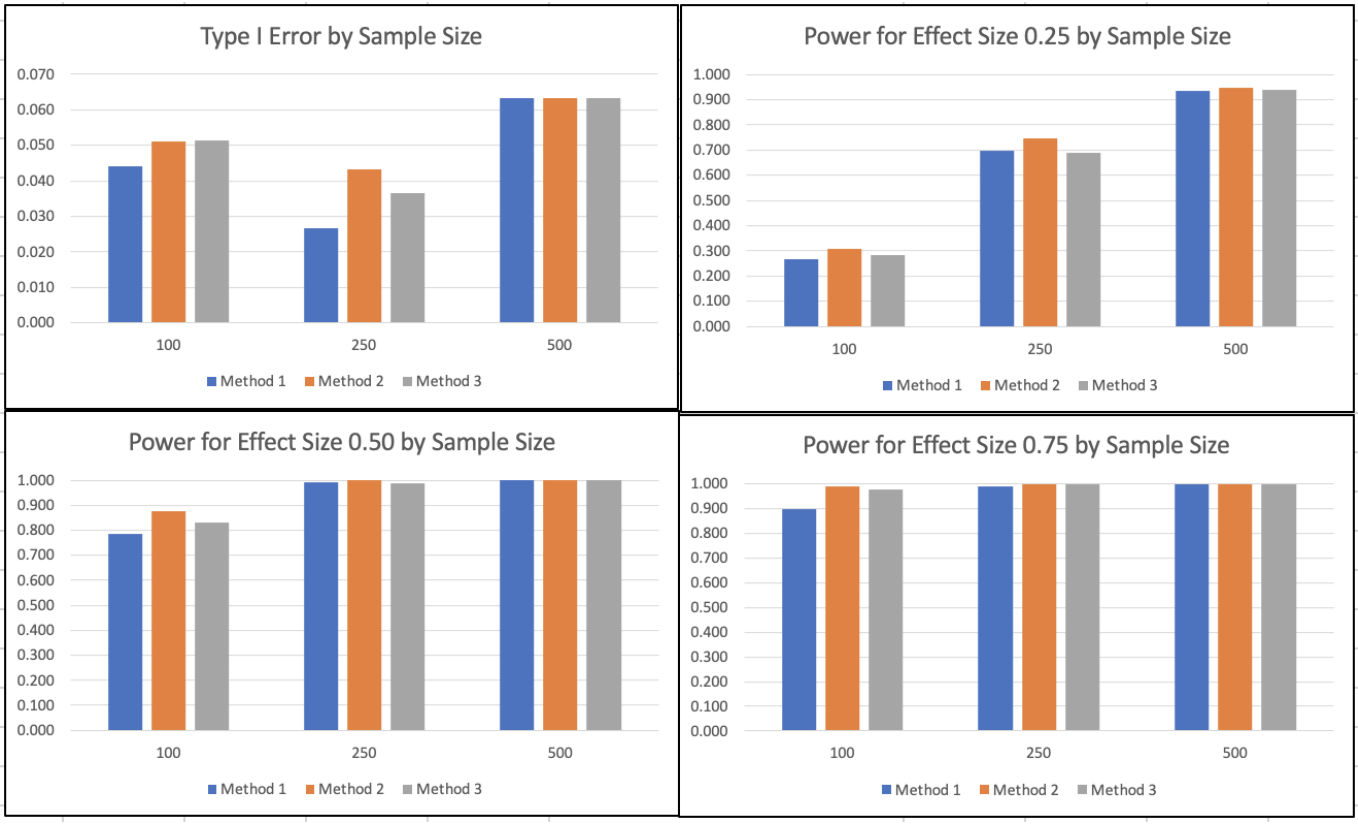
Overall, Method 2 had the highest average Type I error value ($Min = 0.043, Max = 0.063$) across each sample size condition, and Method 2 yielded the lowest values ($Min = 0.027, Max = 0.063$). When considering Type I error for an assessment administered to 100 examinees, both Method 2 and Method 3 yielded a Type I error of approximately 0.051. This result indicates that for every 100 examinees, both post-hoc methods rejected the null hypothesis approximately 5.1% of the time when the null hypothesis was actually true. Considering the same number of examinees, Method 1 yielded a slightly lower Type I error of 0.044, indicating that a significant effect was determined 4.4% of the time when there was none to be detected (see Table 1). Interestingly, Type I error decreased for all three methods for a 250-examinee condition before increasing to 0.063 across all methods in a 500-examinee condition.

Power increased for each method as effect size increased, indicating that more models rejected the null hypothesis as the effect of the covariate increased. As expected, all methods yielded the highest power for an assessment of 15 test questions and 500 students. Overall, Method 2 had the highest power as effect size and test length increased ($Min = 0.307, Max = 1.00$) and Method 1 had the lowest power out of all three methods ($Min = 0.267, Max = 1.00$). In addition, Method 1 had considerable low power for the 0.75-effect size, 100-participant condition (power = .900*)* compared to Method 2 (power = .990*)* and Method 3

(power = .977*),* though this is the only condition where Method 1 had a considerable drop in power compared to the other methods (see Figure 6). This finding indicates that a model where a covariate effect is included in the model estimation is slightly less capable of picking up on covariate effects than the post-hoc methods.

Figure 6

Type I Error and Power for Each Method Across Sample Size



In conclusion, the results of the simulation suggest that Method 1 is the best method for including a covariate in a DCM. Despite Method 1 having the slightly lower power compared to Method 2 and Method 3, all three methods had strong power that remained in range of each other. Further, Method 1 had the highest average accuracy and reliability scores compared to the

two methods. Therefore, Method 1 was chosen as the empirical analysis method to analyze Garth-McCullough's (2008) study on culturally-relevant assessment material and Black students' reading comprehension scores.

CHAPTER 6

EMPIRICAL ANALYSIS METHODS

Purpose of the Analysis

Garth-McCullough (2008) explored the relationship between Black middle school students' prior reading ability (hereinafter referred to as "prior reading level"), their prior cultural knowledge or familiarity with the language, beliefs, and daily experiences of a cultural group (hereinafter referred to as "prior knowledge"), and the cultural orientation (or context) of a short reading passage to determine the students' reading comprehension scores. The researcher collected participant reading scores based on their responses to two texts from three cultural orientations: African-American, Chinese-American, and European-American. The purpose of my empirical analysis was to replicate the analysis in a DCM context and uncover potential notable findings. More specifically, I was interested in exploring how mid-level readers with high prior knowledge compared to high-level readers with low prior knowledge on an African-American-centered text.

Participants

The original study in 2001 in four urban middle schools in a midwestern city. There was a total of 117 participants in the original study. From the original participant count, I removed students from the data analysis if they had no prior reading level score or prior knowledge score reported. Participants were also removed from each orientation dataset if there were no responses recorded. After removal, there were a total of 96 participants for the African-American text, 95 for the Chinese-American text, and 92 for the European-American text.

Prior Knowledge Instrument Design

The original prior knowledge instrument consisted of seven sections that prompted students to demonstrate their knowledge about cultural references. These sections were Yes/No, Free Association, Multiple Choice, Translate, Synonyms, Antonyms, and Experience and Interest.

Reading Comprehension Instrument Design

There were 30 test items for the African-American text and the Chinese-American text and 29 items for the European-American text. Each test item was multiple choice. Students were asked to answer questions about events from the corresponding text and students had to choose the correct answer choice.

Data Cleaning

In each original dataset, correct item responses were scored as "2" and incorrect responses were scored as "1." For this study, correct responses were scored as "1" and incorrect responses were scored as "0." For each participant in each cultural orientation category, I changed any missing item responses to "0" in the dataset. This change was made to allow Mplus to properly read in the data text files. Also, in the original dataset for the Chinese-American text, there were six item responses that were left as letter choice text responses (i.e., the student responses were not scored as correct or incorrect.) For fairness, I scored those responses as incorrect as I did not have the answer key for the texts.

Covariates

The covariates for examinee classification in this study were prior knowledge and prior reading level. Garth-McCullough created a demographic instrument collected to collect information on students' prior reading level. She also developed a preliminary instrument to gauge students' prior knowledge scores for each cultural orientation.

Prior Reading Level and Prior Cultural Knowledge

Student prior knowledge scores were determined based on their responses to a preliminary instrument that Garth-McCullough created to evaluate student familiarity with the three cultural orientations. Therefore, participants had three separate prior knowledge scores; one for each cultural orientation. I coded each participant into high ("2"), medium ("1"), and low ("0") prior knowledge groups by finding the maximum and minimum score for each text I chose for analysis and creating a range of scores for each group.

Software

I used Mplus to analyze the student responses for each of the three texts. Since Method 1 was chosen as the best method for including a covariate on model estimation, the prior knowledge and prior reading level covariates were regressed onto the model classifications for each text.

CHAPTER 7

EMPIRICAL ANALYSIS RESULTS

Significance of Covariates for All Cultural Orientations

For the African-American text, the equation below shows the regression equation:

$$logit(proficiency) = .375 - .052(prior\ knowledge) + 1.269(reading\ level) \quad (4)$$

This indicates that the more knowledge an examinee has about African-American culture, the lower their proficiency on reading comprehension. On the other hand, the higher a student's reading level, the lower their reading proficiency. However, neither prior knowledge ($p = 0.75$) nor prior reading level ($p = 0.067$) were significant predictors of examinee proficiency status. That is, examinee proficiency classifications were not related to familiarity with the cultural information in the text nor their prior reading level. It appears that African-American-culture-centered reading passages leveled the playing field for the Black examinees. This finding will be further explored in the examinee classification analysis for this text.

On the other hand, for the Chinese-American text, the regression equation for is:

$$logit(proficiency) = 2.633 + .172(prior\ knowleddge) + .858(reading\ level) \quad (5)$$

Indicating that the more cultural knowledge that an examinee had and the higher their reading level, the higher their reading proficiency. Both prior knowledge ($p = 0.007$) and prior reading level ($p = 0.045$) were significant predictors for the Chinese-American text.

Finally, the regression equation for the European-American text is:

$$logit(proficiency) = 3.18 + .619(prior\ knowledge) + 1.15(reading\ level) \quad (6)$$

When considering the European-American orientation, prior knowledge ($p = 0.002$) and prior reading level ($p = 0.002$) were found to be the most significant predictors of examinee classification compared to the other two cultural orientations. In sum, results indicate that both an examinee's prior cultural knowledge and reading ability were of an equal determinant for students to be classified as a master in reading comprehension for the European-American text. Table 3 shows the parameter estimates and standard errors for prior cultural knowledge and prior reading level for each cultural orientation.

Table 3

Estimates of Prior Knowledge and Prior Reading Level for All Cultural Orientations

Text Cultural Orientation	Prior Knowledge	Prior Reading Level
African-American	.052 (.164)	-1.27 (.692)
Chinese-American	-.172 (.064)***	-.858 (.427)**
European-American	-.62 (.2)***	-1.15 (.38)***

***p>.01, **p>.05

Examinee Classifications for African-American Text

There was a total of 96 examinee responses analyzed for the African-American text with 31 low-level readers, 38 mid-level readers, and 27 high-level readers. Further, the minimum prior knowledge score for the African-American text was 3 while the maximum was 16. Therefore, I separated examinees' prior knowledge scores from 3-7 (low), 8-11 (medium), and 12-16 (high). Table 4 shows the examinees in master and non-master classification based on examinee prior knowledge and reading level groups combined.

Mid-level readers with high prior knowledge make up the greatest percentage (42%) of the master class for the African-American text. This finding indicates that mid-level readers who had a great familiarity with the cultural references in the text performed better than high-level readers, regardless of the high-level readers' prior knowledge classification. This conclusion is consistent with Garth-McCullough's (2008) finding where mid-level readers with high prior cultural knowledge had the highest mean score on the African-American texts overall than high-level readers with low prior knowledge. Despite this group of students not being the strongest readers of the participants, they were still able to demonstrate comprehension that excelled that of the strongest readers.

In addition, low-level readers with high prior knowledge were more likely to be classified as masters than low-level readers from the low and mid-level knowledge groups. These conclusions indicate that students of lower reading levels were able to make connections and interpretations from the African-American text that were beyond the reading strategies that were used by high-level readers. Bringing in content that relates to Black students' daily experiences and interests can give students across different ability levels the opportunity to engage and demonstrate competence by providing them a point of reference for the reading material that they would most likely lack if it was not relevant to their identities and cultural practices (Ladson-Billings, 1995; Garth-McCullough, 2013).

Table 4

Cross Tabulation of Examinee Mastery Classifications for African-American Text

	Low Prior Knowledge		Mid Prior Knowledge		High Prior Knowledge		
Prior Reading Level	**Master**	**Non-Master**	**Master**	**Non-Master**	**Master**	**Non-Master**	**Total**
Low	0	2	0	9	6	14	31
Mid	2	0	4	1	24	7	38
High	5	0	7	1	9	5	27
Total	7	2	11	11	39	26	96

Examinee Classifications for Chinese-American Text

For the Chinese-American orientation, there were 31 low-level readers, 38 medium-level readers, and 27 high-level readers out of the 96 total examinees. The minimum prior knowledge score was 1 and the maximum was 26. The range for the low prior knowledge category was 1-8, the middle range was 9-17, and the high range was 18-26. High-level readers with high prior knowledge were the largest group in the master classification (32%). This is not surprising as Garth-McCullough found that high-level readers with high cultural knowledge had the highest scores of any other group on the Chinese-American texts. However, it must also be pointed out that there were more masters in the low-level reader, high cultural knowledge group (6 masters) than non-masters (1 non-master). This finding suggests that a low-level reader's familiarity with Chinese-American cultural concepts gives them a considerable boost in probability of answering these items correctly**,** increasing their chances of being classified as a master.

Table 5

Cross Tabulation of Examinee Mastery Classifications for Chinese-American Text

	Low Prior Knowledge		Mid Prior Knowledge		High Prior Knowledge		
Prior Reading Level	**Master**	**Non-Master**	**Master**	**Non-Master**	**Master**	**Non-Master**	**Total**
Low	0	4	8	13	6	1	32
Mid	0	1	15	4	13	2	35
High	2	0	1	3	21	1	28
Total	2	5	24	20	40	4	95

Examinee Classifications for European-American Text

Out of the 92 student responses used in the European-American text analysis, 27 students were low-level readers, 35 students were mid-level, and 30 students were high-level readers. For the European-American text, the minimum prior knowledge score was 1 and the highest was 8. The prior knowledge categories for the European-American text were 0-3 (low), 4-6 (medium), and 7-8 (high). Interestingly, mid-level readers with mid-prior knowledge were the largest group for the master classification (31%). It should also be noted that in the high prior knowledge group, there were zero non-masters across each reading level group.

Table 6

Cross Tabulation of Examinee Mastery Classifications for European-American Text

	Low Prior Knowledge		Mid Prior Knowledge		High Prior Knowledge		
Prior Reading Level	**Master**	**Non-Master**	**Master**	**Non-Master**	**Master**	**Non-Master**	**Total**
Low	1	12	6	7	2	0	28
Mid	4	4	18	6	2	0	34
High	2	3	15	2	8	0	30
Total	7	19	39	15	12	0	92

CHAPTER 8

DISCUSSION AND CONCLUSIONS

I explored the relationship between culturally-relevant assessment material on African-American students' reading comprehension scores in a diagnostic classification model context. First, I conducted a simulation study to investigate the best method of including a covariate in DCM estimation: estimating the covariate directly in model estimation (Method 1), regressing the covariate onto examinee posterior probabilities (Method 2), and regressing the covariate onto mastery classifications (Method 3). It was hypothesized that Method 1 would consistently have the highest accuracy and reliability rates and the highest power out of the other two methods. Further, Method 1 would have a Type I error of 0.05 or lower for each sample size condition. The simulation study revealed that Method 1 yielded higher average accuracy and reliability rates than the two post-hoc methods. However, though each method had an appropriate Type I error for each sample size condition, Method 2 maintained the highest model power for each effect size condition. I concluded that Method 1 was the best method of covariate inclusion in DCM estimation because of the high accuracy and reliability values. This method was used to conduct a secondary analysis of Garth-McCullough's (2008) research on Black students' reading comprehension scores on culturally-relevant assessment material.

Initially analyzed using IRT, Garth-McCullough investigated Black middle-school students' reading comprehension scores on reading passages from African-American, Chinese-American, and European-American cultural orientations. She investigated the effect of students' prior cultural knowledge and reading level on their scoring. One of the most notable findings from this study is that mid-level readers with high prior knowledge of African-American culture scored significantly higher on those texts than high-level readers with low prior knowledge. This

finding was replicated in the present study, as there was more mid-level, high-prior-knowledge readers classified as masters for the African-American text than there were high-level, low-prior-knowledge readers classified as masters.

Implications

The findings from the empirical analysis present a benefit in DCMs and the categorization of student ability on particular attributes, especially when considering external factors on examinee scores. Since examinees' overall ability is broken down by their performance on specific concepts and skills, educators and researchers can evaluate students' growth areas while recognizing their strengths. This is an appealing consideration for students of lower ability. For example, as the empirical analysis shows, culturally-relevant assessment material can level the playing field for students of average reading ability, allowing them to tap into funds of knowledge outside the testing and school environment. These findings align with schema theory and the examinees' reading comprehension. In short, schemas serve as cognitive "frames" or categories that people use to organize and interpret new information based on their prior experience or knowledge of the information (Lefa, 2014). People use these preconceived images and thoughts about people, places, things, and concepts to make sense of new information; people determine how this information fits into their schemas. When considering reading comprehension, it is more than a student's passive recall of the events in a text.

Students unconsciously connect their prior knowledge about a subject and the reading material to make sense of the reading (Garth-McCullough, 2008). Regardless of a student's prior reading ability, they can still make sense of the material in the reading passage if they already have a schema for the material. Therefore, it would make sense that students of average reading ability excelled in African-American-oriented texts despite not being the strongest readers. These

students were familiar with the African-American passage's language, setting, events, and characterizations, allowing them to comprehend the culturally salient material of the texts. Also, despite most low-level readers being classified as non-masters for the African-American text, there were a greater number of low-level readers with high prior cultural knowledge for the African-American text orientation than there were for the Chinese-American and European-American texts.

To capitalize on the fact that many of these non-masters had high familiarity with African-American culture, educators could develop culturally-relevant material for any interventions to help low-level readers improve their reading scores. For example, if a low-level reader that struggles with recognizing structure in prose text indicates that they enjoy hip-hop music, a teacher could incorporate lyrics from rap songs into lessons on prose structure. This intervention can help the student engage in the material more because it aligns with his interest, use his prior knowledge to connect the text and the skill, and allow him to express his knowledge in a way that a traditional assessment may not measure.

Limitations

One major limitation is the number of conditions in the simulation study. Any simulation has finite conditions that may not reflect all assessment or research scenarios. This stimulation study was meant to provide some initial insights into the inclusion of covariates in DCM analyses. Future work could examine other factors such as manipulation of base rates, number of attributes, and Q-matrix. Another limitation is of the simulation study is the low number of examinees I generated for my low, medium, and high sample size conditions. Though I observed a trend in how the With Covariate condition performed compared to the Without Covariate condition for the evaluation metrics, the sample sizes I chose were not comparable to other DCM

studies, nor were they reflective of high-stakes assessment sizes. Therefore, I cannot determine if the simultaneous inclusion of the covariate in the model estimation would hold in a larger pool of examinees. I would continue this study by generating a larger number of examinees for each sample size condition, such as 500, 1,000, and 2,500 examinees.

Another area for improvement in the generalizability of this study is the recency of the empirical data. Given that Garth-McCullough's original study was conducted over 20 years ago, the relevancy of the cultural references used in the reading passages are likely outdated. As Ladson-Billings (2014) mentions, culture is dynamic, and each generation of students brings various shared experiences, perspectives, opinions, and cultural developments to their learning environments. This is a fascinating thought with technology and social media development. With the prevalence of social media, students have significant exposure to different cultures in terms of food, dance, language, ways of dressing, and much more at the palm of their hands. Therefore, this study may not reflect how Black students would be classified on each cultural text today. It would be interesting to explore Black students' familiarity with other cultural texts today compared to African-American texts with more relevant cultural knowledge.

Future Directions

Application of DCMs

DCMs are useful for standards-based curricula as the mastery classifications can be used to provide actionable and interpretable feedback to stakeholders. These models are helpful to highlight students' strengths while pinpointing their areas for growth, allowing teachers to create and implement personalized review plans that optimize their students' learning. DCMs also have great potential to benefit students in other forms of testing, such as computer-based testing.

Computer-based testing typically encompass various activities to increase student learning (especially outside of class time) with activities such as instructional modules, learning materials, and game-based learning. In many cases, students are given feedback on the concepts or skills that they should further review with suggested activities, allowing them to identify their improvement areas (Achieving the Dream, 2022; Wu, 2016). In this case, DCMs could serve as a psychometric foundation for the results that students and teachers receive during learning and testing modules. Incorporating DCMs into computer-based testing can help students take direct control of their review times outside of class. For example, the software could suggest review questions, games, and other activities based on attributes the student had not yet mastered. In that case, students can study with greater purpose and show significant improvement in their ability during the next testing session.

Further, DCMs should be considered for intelligent tutoring systems (ITS) that educators use to evaluate and track student learning. As Murphy (2019) explains, the purpose of ITS is to learn, predict, and adapt to a student's learning based on how they perform on tasks. Considering this purpose, DCMs can be a great asset to intelligent tutoring systems as the purpose of implementing DCMs is to determine student mastery on specific skills and concepts already. Further, this framework can be used to develop student- and teacher-friendly reports on student progress.

This feedback can also decrease the time teachers spend on whole-class reviews when it where it is not needed, reducing the amount of time that high-ability level students engage in review they do not need, and allowing those students who need more support to receive targeted instruction (Wu, 2016). Further, educators can take greater advantage of student-led review sessions where students who are masters on one attribute tutor non-masters. These sessions can

encourage peer learning and peer relationship development and allow students who are shyer about seeking help from the teacher to grow in their weaknesses as well.

Applications of Culturally-Relevant Assessment

The application of culturally-relevant assessment material should be further explored in standardized testing contexts and other forms of testing, such as oral assessments, student portfolios, or extended projects. As Gunskey (2007) suggests, multiple means of assessing student learning should be implemented to capture student learning in ways that show students' progress of learning over time (rather than a once-per-year-large-scale test) and emphasizes each student's academic strengths and interests. Many advocates for testing reform encourage modes of testing that more realistically align with how and what students are learning in the classroom and those that emphasize student voice. As discussed, culturally-relevant assessment material is designed to allow students whose interests, heritage, and participation in current events are underrepresented in the mainstream curriculum to engage with material that more directly represents themselves.

This study could be extended to incorporate reading passages from across the African diaspora. Again, it should be recognized that African-American culture is not monolithic. Black students with immediate family members who are immigrants from other countries, or familial and cultural ties to non-American culture in general, could have additional ways of interpreting and engaging with educational material (Yosso, 2005). Therefore, including materials in assessments, such as reading passages from diasporic authors, would diversify the perspectives and considerations presented in the educational realm. It would also be interesting to examine Black examinees' performance on African diasporic texts with which they may not have direct experience considering ideas of cultural exchange within the diaspora.

Further, I hope that this paper encourages researchers to consider the implications of culturally-relevant assessment material for students of other cultural groups as well. As previously mentioned, the use of culturally-relevant material could be implemented to encourage the success of multilingual students on assessments, especially where the use of English-only reading selections limits their understanding or ability to demonstrate competence on words and themes (Yosso, 2005) More than anything, using culturally-informed assessment material should allow students more plentiful and diverse opportunities to demonstrate their knowledge in ways that traditional assessments may not provide.

As discussed earlier in the paper, culturally-relevant material should be purposefully and continuously implemented within educational contexts. Though the study focused on the application of the such material on large-scale exams, these methods could very well be used on teacher-designed classroom assessments. Such efforts could begin with having direct conversations with the students, their families, and members of the community so that teachers can better understand their students' personal contexts. Teaching practices that stem from culturally-relevant pedagogy are used to amplify the voices of students of color. Therefore, teachers should create the spaces for students to do so, allowing them the freedom to discuss the interests and values that influence how they learn. Further, on tests that give DCM-based feedback on student performance, teachers can collaborate with students on interventions that make the information more relevant to their interests and values, helping students improve their scores on attributes on which they need support.

In conclusion, as the educational landscape changes, more research and educational organizations are looking toward solutions to dismantling racist practices that have kept students

of color behind. I want to present a quote from Ladson-Billings (2014) on her purpose in creating CRP:

> As a researcher, it is part of my responsibility to help scholars see African-American students as agents in the classroom worthy of study and emulation. In other words, I hoped to help scholars and practitioners learn *from* and not merely *about* African-American students.

Let us start considering Black students as active contributors to and leaders in the academic realm rather than a reference group to measure other students' success against or as poster children for adverse academic and behavioral outcomes. Their ways of learning and seeing the world are just as valuable as their White peers. As educators and researchers, we must check our biases against African-American students and help change the narratives around their academic potential and success.

REFERENCES

Achieving the Dream (2022). Lessons learned from ATD network colleges in the Every Learner Everywhere initiative: Full report. https://achievingthedream.org/wp-content/uploads/2022/05/atd_ele_adaptive_courseware_new_models_to_support_student_learning.pdf

Au, W. (2016*).* Meritocracy 2.0: High-stakes, standardized testing as a racial project of neoliberal multiculturalism. *Educational Policy, 30*(1), 39-62.

Bradshaw, L. P., & Madison, M. J. (2016). Invariance properties for general diagnostic classification models. *International Journal of Testing,* 16, 99-118.

Choi, Y. W. (2020, March 31). *How to address racial bias in standardized testing*. https://www.nextgenlearning.org/articles/racial-bias-standardized-testing

Friere, P. (1970). Pedagogy of the oppressed. In Mekerta, S., Busdiecker, S., Leeds, A., Pierre, A., & Williams, E. (Eds.), *African diaspora and the world: Readings for ADW 111.* Spelman College.

Garth-McCullough, R. (2008). Untapped cultural support: The influence of culturally bound prior knowledge on comprehension performance. *Reading Horizons: A Journal of Literacy and Language Arts, 49*(1), 1-30.

Garth-McCullough, R. (2013). The relationship between reader response and prior knowledge on African-American students' reading comprehension performance using multicultural literature. *Reading Psychology, 34*(5), 397-435. doi: 10.1080/02702711.2011.643531

Guskey, T. R. (2007). Multiple sources of evidence: An analysis of stakeholders' perceptions of various indicators of student learning. *Educational Measurement: Issues and Practice, 26,* 19-27.

Hallinan, M. T., & Kubitschek, W. N. (1999). Curriculum differentiation and high school achievement. *Social Psychology of Education,* 3, 41-62.

Hedges, L. V., & Nowell, A. (1999). Changes in the Black-White gap in achievement test scores. *Sociology of Education, 72*(2), 111-135. https://www.jstor.org/stable/2673179

Ladd, H. F. (2017). No Child Left Behind: A deeply flawed federal policy. *Journal of Policy Analysis and Management, 36*(2), 461-469.

Ladson-Billings, G. (1995). Toward a theory of culturally-relevant pedagogy. *American Educational Research Journal, 32*(3), 465-491.

Ladson-Billings, G. (2014). Culturally-relevant pedagogy 2.0: The remix. *Harvard Educational Review, 84*(1), 74-84.

Lefa, B. (2014). The Piaget theory of cognitive development: An education implications. *Educational Psychology, 1*(1), 1-8.

Lin, Q., Xing, K., & Park, Y. S. (2020). Measuring skill growth and evaluating change: Unconditional and conditional approaches to latent growth cognitive diagnostic models. *Frontiers in Psychology,* 11, 1-12, doi: 10.3389/fpsyg.2020.02205

Morris, J. E. (2004). Can anything good come from Nazareth? Race, class, and African-American schooling and community in the urban South and Midwest. *American Educational Research Journal,* 1, 69-112.

Murphy, R. F. (2019). *Artificial intelligence applications to support K-12 teachers and teaching: A review of promising applications, opportunities, and challenges.* RAND Corporation. https://www.rand.org/content/dam/rand/pubs/perspectives/PE300/PE315/RAND_PE315.pdf

Perez, C. (2002). Different tests, same flaws: Examining the SAT I, SAT II, and ACT. *The Journal of College Admission,* 20-25.

Philips, M. (2006). Standardized tests aren't like t-shirts: One size doesn't fit all. *Multicultural Education, 14*(1), 52-55**.**

Randall, J., Poe, M., & Slomp, D. (2021). Ain't oughta be in the dictionary: Getting to justice by dismantling anti-Black literacy assessment practices. *Journal of Adolescent & Adult Literacy, 64*(5), 594-599.

Robinson, K. (2010). Black-White inequality in reading and math across K-12 schooling: A synthetic cohort perspective. *The Review of Black Political economy, 37*(3-4), 263-273.

Seneca College, Humber College, Kenjgewin Teg, Trent University, & Nipissing University. (n.d.). Culturally responsive assessments. *Designing and Developing High-Quality Student-Centred Online/Hybrid Learning Experiences.* Pressbooks.

Scott, L. A. (2023). If I ruled the world: Imagining culturally sustaining pedagogy in education. In Hunter, W., Taylor, J., Scott, L. (Eds.), *The Mixtape Volume #1: Culturally Sustaining Practices Within MTSS Featuring the Everlasting Mission of Student Engagement.* Council for Exceptional Children.

Sireci, S. G., & Randall, J. (2021). Evolving notions of fairness in testing in the United States. In Clauser, B. E., & Bunch, M. B, *The history of educational measurement* (pp. 111-135). Routledge. https://doi.org/10.4324/9780367815318

Skiba, R.J., Knesting, K., Bush, L. D. (2002). Culturally competent assessment: More than nonbiased tests. *Journal of Child and Family Studies, 11*(1), 61-78.

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

Stewart, S. & Haynes, C. (2016). An alternative approach to standardized testing: A model that promotes racial equity and college access. *Journal of Critical Scholarship on Higher Education and Student Affairs, 2*(1), 122-136.

Templlin, J. & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30,* 251-275.

Thompson, G. L., & Allen, T. G. (2012). Four effects of the high-stakes testing movement on African-American K-12 students. *Journal of Negro Education, 81*(3), https://doi.org/10.7709/jnegroeducation.81.3.0218

U.S. 107th Congress. (2001). H.R.1 – No Child Left Behind Act of 2001.

Wu, H. (2019). Online individualised tutor for improving mathematics learning: a cognitive diagnostic model approach. *Educational Psychology, 30*(10), 1218-1232.

Yosso, T. J. (2005). Whose culture has capital? A critical race theory discussion of community cultural wealth. *Race, Ethnicity and Education, 8*(1), 69-91.