Toward Culturally Responsive and Equitable Testing: Innovative Psychometric Analyses on Contextualized Measurement and Adaptive Testing

Nixi Wang

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Min Li, Chair

Chun Wang

Geneva Gay

Program Authorized to Offer Degree:

College of Education

University of Washington

Abstract

Toward Culturally Responsive and Equitable Testing: Innovative Psychometric Analyses on Contextualized Measurement and Adaptive Testing

Nixi Wang

Chair of the Supervisory Committee:

Min Li

College of Education

Measurement errors attributable to cultural issues are complex and challenging for educational assessments. We need assessment tests sensitive to the cultural heterogeneity of populations, and psychometric methods appropriate to address fairness and equity concerns. Built on the research of culturally responsive assessment, this dissertation explores the conceptual, analytical, and methodological aspects of culturally responsive and equitable testing. To that end, three coordinated studies were conducted to gather validity evidence based on cultural characteristics of contextualized items, students, and multigroup fairness from empirical and simulated datasets. The first study contributed to the understanding of culturally valid and equitable contextualized items, by implementing a coding rubric of item contexts followed by an illustrative analysis of science assessment items. The rubric of fifteen attributes was developed to represent context domains of cultural equity, context representation, and knowledge construction. The mixed methods, including item response theory, clustering profiles, and discourse analysis were conducted to explore context codes and student responses, along with

qualitative analysis of item contexts. The results found various associations of contextual characteristics such as cultural bias, linguistic bias, and others with item difficulty and differential item functioning (DIF) parameters. The second study implemented a Bayesian hierarchical explanatory item response model on a physics contextualized assessment. It examined the statistical relationships between the sociocognitive and sociocultural aspects of item contexts and student performance. Context features were systematically and iteratively coded, modified, and analyzed based on their feature importance. Results showed the statistical associations of context codes, including sociolinguistic familiarity, length of context, cognitive demand, and sociocultural bias on students' success probability given their latent proficiency. Conditional context effects accounting for students' racial backgrounds were also discussed. The third study considered context features as item attributes in general and explored the methodological issues of DIF in cognitive diagnostic computerized adaptive testing (CD-CAT) through a simulation study. Results revealed the performance of uniform and nonuniform DIF detection using different item selection algorithms, DIF-contaminated banks, and other various simulated conditions of the CD-CAT. Classification accuracy, exposure rate, and overlap rate were also explored. In summary, this dissertation demonstrated the significance and relevance of fair, culturally valid, responsive, individualized, and adaptive items and algorithms in educational measurement and testing. Practical and methodological implications were discussed for test and item development, validity, science knowledge, and fairness.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am indebted to many people who gave me invaluable support throughout the years of my PhD journey. I would like to thank my advisor, Dr. Min Li, for encouraging me and guiding me through the Measurement and Statistics program. Her support and advice were essential. This dissertation would not have been possible without her patience and direction. Thanks to Dr. Geneva Gay and Dr. James A. Banks, who brought me to the field of multicultural education and whose critical scholarship helped me grow my research identity. My committee including Dr. Chun Wang and Dr. Joseph Hellerstein have provided generous support and insightful feedback on my work. I thank Dr. Solano-Flores for challenging and stretching my intellectual thoughts. Many other professors in the College of Education such as Ken Zeichner and Elizabeth Sanders have extended their help either in research funding or knowledge sharing. I am thankful for Paula Wetterhahn's assistance and Dr. Evelyn Law's mentorship and trust.

To my incredible research community, including fellow graduate students, research collaborators, and intellectual friends from other institutions, I am grateful for our wonderful experiences and the flourishing of ideas. I thank Cinthia Palomino, Dongsheng Dong, Nate Abe, Klint Kanopka, Philip Hernandez, Xiaoming Zhai, Gage Kleinman, Manqian Liao and Noah Padgett from the American Institutes for Research internship, Andrea Brudvig, Aveline Vasu, Matt Davidson, and many others. They gave me the space to grow together and inspire each other.

I would like to thank the National Science Foundation for the gracious funding I received to work on the two grants: Building a Framework for Developing and Evaluating

Contextualized Items in Science Assessment (DECISA), led by Ruiz-Primo, Li, and Minstrell, NSF ECR program (NSF-1432406); and Automatic Profiling of Science Assessment items to Model Item Parameters: A natural Language Processing Approach, led by Li and Ostendorf, NSF ECR Core research program (NSF-1920512). Special thanks to the participants in my studies, without whom I would not have learned the insights from the collected student test scores.

Lastly, my most heartfelt gratitude goes to my big Family. My parents, my husband, my MIL, and other extended family members like Eileen and Chitra—they love me unconditionally and are willing to go extra miles to strengthen me with their substantial help and devotion. My ICF family including Yanlin, Heidi, Litai, John, Kristina, Sarah, Jenn, and others who surrounded me with memorable moments of laughter, food, music, and sharing. Peace and love.

DEDICATION

This dissertation is dedicated to the One who makes me and deserves all the glory.

Chapter 1. Introduction

The search to determine whether educational assessments lead to meaningful learning and evaluation must begin by identifying the needs and standards. Policymakers and researchers have reformed tests from the local to the large-scale national and international, from standardized procedural criteria to culturally responsive guidelines, and from traditional paper-pencil tests to computer-based, multimedia, and multimode technological systems of formative and summative assessments. The trend is distinct toward intelligent tutoring, learning, and assessment systems. For instance, large-scale assessments such as the National Assessment of Educational Progress (NAEP) and the Programme for International Student Assessment (PISA) extend testing beyond rote memorization of isolated facts to complex thinking, contextualized understanding and complex communication. We have seen innovations and reforms following those assessments taking momentum in implementing new conceptual frameworks and psychometric methodology. Furthermore, the changing demographic makeup in modern society brings multiple perspectives and provides test populations with cultural, ethnic, gender, and language alternatives. Individuals from diverse cultural backgrounds must be assessed effectively with the required knowledge, attitudes, and skills. Central among the social, educational, psychological, and measurement aspects of assessment are issues of equity and fairness (Bond, 1995; Cole, 1973; Camilli, 2006; Scott et al., 2013; Zieky, 2006).

Test fairness and bias are central issues of educational equity and diversity in the modern era of assessment. As a minimal request, the Standards (AERA, APA, NCME, 2014) require that assessments to measure the intended construct and minimize the possibility of student scores being affected by construct-irrelevant causes such as

linguistic, communicative, cultural, physical, or other characteristics. For the most part, efforts to determine bias focus on assessing the differences and variations in test scores between student subgroups around race, ethnicity, language, gender, culture, socioeconomic status, or special needs status. Such differences and variations are typically explained by construct-underrepresentation or construct irrelevant variance in a test. Still, many education researchers agree that determining the extent and nature of test bias is a necessary (but not sufficient) step toward developing fair and equitable assessments for all students (cf., Cole, 1996; Camilli, 2006).

There is a clear disciplinary boundary between psychometric-based work on test validity and bias and work addressing equity and student diversity in learning, curriculum, and assessment practices. Both recognize that assessments need to reduce barriers and increase access to success for all learners. However, psychometric research on equity and test fairness mostly focuses on interpreting social consequences and perceptions of test scores rather than improving the measures themselves for cultural validity. To bridge the gap between measurement and learning theories, I draw on sociocultural theories of learning (e.g., Rogoff, 2003; Vygotsky, 1978; Wertsch 1991) and culturally responsive teaching (Gay, 2000; Ladson-Billings, 1995) to reconceptualize contextualized measurement and adaptive testing.

1.1 Definitions and Goals

In this dissertation, I envision that culturally responsive and equitable principles are crucial to rectifying assessment issues. Culturally responsive and equitable testing (CRET) essentially seeks the question of "what makes an inherently 'good' and 'culturally valid' assessment item for learners from diverse backgrounds", and "what aspects of test measurement" can be culturally responsive. On both item and test levels, I discuss how psychometrics can be used for theory-driven research. On the item level,

for students who may have to surmount cultural and language barriers, culturally responsive, fair, and equitable item contexts are called to attention. On the test level, individualized and tech-powered testing infused with culturally responsive principles will provide multiple means of representation and engagements with the students, and empower and motivate students with positive and fair educational outcomes.

Importantly, CRET deems the knowledge and epistemology of assessment items as critical, holding the knowledge producers accountable in their agenda and calling into question the implicit standards and assumptions embedded in assessment. This is by no means to undermine objectivity but to undergird and uphold assessment knowledge for stronger cultural validity. Next, I will iterate the definitions and goals, including the two fundamental ideas of cultural validity and culturally responsive assessment, followed by a brief introduction about each chapter for how I approach the fairness and equity issues in CRET.

What is Cultural Validity? Sociocultural approaches to validity make explicit inferences about how culture plays a role in students' differentiated understandings of items and performance. For example, Solano-Flores and Nelson-Barber (2001) explain cultural validity as the:

> effectiveness with which [...] assessment addresses the sociocultural influences that shape student thinking and the ways in which students make sense of [...] items and respond to them. These sociocultural influences include the sets of values, beliefs, experiences, communication patterns, teaching and learning styles, and epistemologies inherent in the students' cultural backgrounds, and the socioeconomic conditions prevailing in their cultural groups (p.555).

The lens of cultural validity provides a validity framework for considering culture in items and test scores, reviewing implicit assumptions, and crystalizing cultural

differences. When properly conceived through culture validity, a test is most sensitive and legitimate to culturally diverse populations.

What is Culturally Responsive Assessment? Scholarship on culturally responsive assessment shows a variety of considerations of students' cultural ways of communicating and acting within and outside the classroom (Hood, 1998; Nelson-Barber and Trumbull, 2007) and assessments or teaching practices that teachers may need to adapt to meet their students' varying needs. For example, assessments may adapt and interpret what constitutes valid and desirable knowledge differently, as it differs within and across educational contexts (e.g., Solano-Flores and Nelson-Barber, 2001; Stobart, 2005). For another example, students from collectivist cultures may prefer demonstrating their knowledge through collaborative ways or assessment formats that are nonconventional such as oral histories. In addition, in performance-based assessment, teachers can facilitate students' participation in classroom assessment situations in different ways, such as conducting peer- or self-assessment, avoiding certain activities or peer-feedbacks which are not conducive to an identity-safe environment (Hunter et al., 2016). The practical approaches to culturally responsive assessments should include assessment practice, experience, process, and culturally valid assessment instruments and items.

In this research, I situate the definitions and discussions of cultural responsiveness in contextualized educational assessment and the relevant adaptive testing technology. As an illustration, in contextualized science assessment, scientific knowledge connected and organized around essential concepts (e.g., Newton's second law of motion) is conditioned, such as through figures and real-life examples, to specify its applicable contexts. In other words, a well-developed item context often supports understanding an item concept. In addition, it allows learners to quickly retrieve

knowledge pertinent to particular settings and assists in understanding how to transfer knowledge of the concept to other contexts. In this work, I contend that culturally responsible and equitable items should characterize and contextualize knowledge concepts in forms, scenarios, and processes familiar to and testable within the student's cultural frame of reference. Analytically, we need to carefully examine whether the way we construct our item contexts can affect or mediate test-taker's psychological, cognitive, and affective functions, and whether they can either facilitate comprehension or hinder understanding as a source of item bias.

On the test level, adaptiveness of testing can leverage items for fairness and for culturally diverse students as well. Unlike standardized tests that are largely considered impersonal, the advent of computerized adaptive tests allow testing to be more customized and engaging while providing precise measurement and better efficiency. Adaptive tests tailor items to the knowledge and ability of the individual so that each person is challenged at the level which is appropriate to them. Furthermore, digital and technological solutions offer opportunities for adaptation and personalization in their system design, such as web accessibility features. However, to utilize adaptive tests for CRET, we need to ensure algorithms are fair and bias-free. Test programs that highlight equity and cultural validity for the benefit of all students will likely be successful, valid, ethical, respectful, and helpful.

Yet, the term *culture* needs to be clearly defined and reflectively reviewed to facilitate the discussion. There are challenges in the perception and manifestation of CRET depending on how the *culture* is approached. Here, I would like to propose we need a way of rethinking *culture* across and beyond the ethnic, racial, gender, language, and social class lines, as a latent layer that accommodates the full complexity of power

in the society and politics of knowledge, instead of one that promotes separatism, simplification, stereotypes, and reification.

On the other hand, in psychometric testing, factors influencing student scores are intricate based on the differences in cultural identities, regions, and social backgrounds. Hence, there is a computational need to disentangle phenomena and observations on the levels of categorization, such as factors derived from the cognitive, psychological, and behavioral processes as experienced by groups of students. For analytical purposes, educational measurement coupled with robust and rigid theoretical frameworks helps to identify sources of bias and areas of relationships underlying group-specific score interpretations. With the manifest cultural groups, I anticipate that the illustrative analyses should suggest more equitable conditions for students from all cultural backgrounds.

In summary, in the face of educational diversity, there is an epistemological, conceptual, and methodological need to support research in educational measurement to dialectically and reflectively embrace theories of learning and to acknowledge and respect learners' cultural backgrounds and approaches to learning. A prescriptive approach, CRET aims for equity and cultural responsiveness by developing fair and culturally valid items and responsive test programs. Upon reviewing cultural validity, this research aims to delineate principles and applications of such concept to discuss what equitable, fair, culturally responsive, individualized, and adaptive testing might entail for psychometric methodology.

1.2 Chapter Overview

The first chapter is an introductory chapter that sets the background for my research and situates readers with a fundamental understanding of concepts and ideas. It includes a

brief introduction, the definitions of cultural validity and culturally responsive assessment, the background of the study, and an overview of the chapters.

The second chapter dives into the theoretical and conceptual considerations of equitable and culturally responsive items and tests. It draws insights from the relevant sociocultural and multicultural literature and provides a conceptual ground for the subsequent studies in this dissertation. Key ideas and parameters about the item measurement are mapped out considering fairness and cultural responsiveness.

The third chapter, a study on contextualized measurement, seeks to identify various contextual discursive attributes of cultural validity and applies the framework to item contexts pertaining to physics knowledge. This chapter also reveals some research gaps in the existing validity literature. It lays out specific questions that a multicultural perspective will focus on to contribute to equity conversations. The analysis weaves in item narratives, knowledge construction, and epistemologies too support assessment development and interpretation of scores. Finally, it suggests that test-based assessment and assessment knowledge must be situated in a bias-free, authentic, realistic, scientific, applicable, and constructive context, linking the functions of knowledge, values, empowerment, and practice.

The fourth chapter is a study investigating and analyzing a design-based contextualized assessment from a modeling perspective. Following the concepts developed in chapter three, the coding system systematically records descriptive, intertextual, and analytical coding of contexts at multiple levels to build context features of cognitive function, cultural equity, and physics knowledge associated with items. For a small selection of the sociocognitive and sociocultural context variables, a hierarchical item response theory framework is applied. The Bayesian analysis of context codes and responses reveals statistical relationships between context characteristics and student

performance. The results contribute to measurement research of calibrating context features and parsing out covariances related to item contexts.

The fourth chapter is a methodological study investigating the methodological considerations of individual adaptiveness of measurement and consequent multigroup fairness concerns. For a long time, diagnostic assessment emphasizes feedback for learning benefits. Cognitive diagnostic, computerized adaptive assessment (CD-CAT) provides detailed information about students' skills and knowledge and enjoys the potential benefits of creating tailored and immediate feedback and recommendations on students' proficiency levels, highlighting each individual's strengths and potential barriers to learning. Following a simulation study of differential item functioning (DIF) in CD-CAT, the analysis offers practical and technological insights into DIF detection and test efficiency and discusses implications for fairness and bias issues in the advanced designs of testing.

Lastly, in the concluding chapter, I end with a discussion to connect the studies back to the high-level concepts. Readers are invited to evaluate the findings tied to the theoretical and practical implications of testing and envision future research programs. This chapter calls for future directions of measurement innovations, such as machine learning, for further exploration and the development of equitable, fair, culturally responsive, individualized, and adaptive testing.

Reference

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

Bond, L. (1995). Unintended consequences of performance assessment: Issues of bias and fairness. *Educational Measurement: Issues and Practice, 14*(4), 21-24.

Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement*, *10*(4), 237-255. https://doi.org/10.1111/j.1745-3984.1973.tb00802.x

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221-256). American Council on Education/Praeger.

Gay, G. (2010). *Culturally responsive teaching: Theory, research, and practice* (2nd ed.). Teachers College Press.

Hood, S. (1998). Culturally responsive performance-based assessment: Conceptual and psychometric considerations. *The Journal of Negro Education*, *67*(3), 187. https://doi.org/10.2307/2668188

Hunter, J., Hunter, R., Bills, T., & Cheung, I. (2016). Developing equity for Pāsifika learners within a New Zealand context: attending to culture and values. *New Zealand Journal for Educational Studies*, *51*(2), 197-209.

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, *32*(3), 465-491. https://doi.org/10.3102/00028312032003465

Nelson-Barber, S., & Trumbull, E. (2007). Making assessment practices valid for Indigenous American students. *Journal of American Indian Education*, *46*(3), 132-147.

Rogoff, B. (2003). *The cultural nature of human development*. Oxford University Press.

Scott, S., Webber, C. F., Lupart, J. L., Aitken, N., & Scott, D. E. (2013). Fair and equitable assessment practices for all students. *Assessment in Education: Principles, Policy & Practice*, *21*(1), 52-70. https://doi.org/10.1080/0969594x.2013.776943

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

Stobart, G. (2005). Fairness in multicultural assessment systems. *Assessment in Education: Principles, Policy & Practice*, *12*(3), 275-287.

Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wertsch, J. V. (1991). *Voices of the mind: A sociocultural approach to mediated action*.

Zieky, M. J. (2006). Fairness review in assessment. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 359-376). Routledge.

Chapter 2. Concepts of Culturally Responsive and Equitable Testing: Reimagining Items and Tests for Culturally Diverse Learners

Abstract

Educational measurement needs to address the complex and multifaceted social and cultural processes in test taking and test takers. This chapter conceptualizes key components of items and tests that help promote *culturally responsive and equitable testing* (CRET), which aims at providing accurate score inferences for culturally diverse groups of test populations. I focus on contextualized assessments for conceptualizing elements of items that are pertinent from a sociocultural perspective. In doing so, it establishes the inherent connection between the social and cultural situatedness of test questions that are raised in contextual scenarios and its pertinence to the test performance of culturally diverse learners. On the test level, I explore how adaptive testing can be conceptualized for culturally diverse populations through culturally responsive principles. The implications of the two components of CRET, item contextualization and test adaptation, are discussed in view of validity, fairness, and equity. Future directions of research are explored.

Keywords: Culturally Responsive Testing, Item Context, sociocultural theory, cultural validity, fairness

2.1 Introduction

Educational measurement, often in the forms of standardized testing, needs to address the complex and multifaceted social and cultural processes in test taking and test takers. In test taking, the ways that student respondents interact with test systems and produce observable responses are inseparable from their cultural participation styles, experiences, and preferences developed in the learning process. In test takers, the current demographic shifts (such as in the United States) require psychometricians and measurement researchers to examine more closely the experiences of student test-takers from various cultural, ethnic, and linguistic backgrounds.

Conventionally, we consider test performance to be caused mainly by persons' latent traits, true scores, or latent variable θ plus some measurement error. Alternatively, in this chapter, we see neither test questions nor student performance as idiosyncratic, but rather shaped by processes and interactions that arise between (socioculturally) and within (cognitively) individuals. Within individuals, there are sets of question characteristics weaving into repeated measures during testing that are interrelated and interacting with individuals' neurocognitive and psychological processes. In particular, research on contextualized assessment sheds light on uncovering the unique resource and role of *situating* problem tasks in rich contextual information (Ahmed & Pollitt, 2007; King, 2012; Ruiz-Primo and Li, 2015; Dong, 2020; Wang et al., 2019; Wang, T., 2016). These include stories, scenarios, situations, or fictional environments that are social and cultural in nature. Moreover, they can frame assessed knowledge in ways that are connected to students' sense-making and meaning-making. A "good" problem scenario may facilitate student comprehension and mental representation in problem-solving, whereas a "bad" problem presentation may contain item bias which hinders student performance (Aronson et al., 1999; Steele & Aronson, 1995).

Between individuals, the disparity between the minority groups and the mainstream students in academic performance of educational assessments is well documented (such as African-American students and White students in Vanneman et al., 2009). The achievement gap between groups is complicated by a complex range of factors, including individual characteristics such as language and socioeconomic status. For example, research supports that, for students of linguistic minorities, their success in problem-solving can be affected by their unfamiliarity with task texts due to their limited language fluency (Luykx et al., 2007; Solano-Flores & Trumbull, 2008). In response to those challenges, culturally responsive assessment is increasingly sought as an equitable paradigm to be mindful of cultural variations in students' ways of participating, thinking, knowing, and learning (Hood, 1998; Lee, C., 1998; Montenegro & Jankowski, 2017; Solano-Flores & Nelson-Barber, 2001; Solano-Flores, 2011, 2019; Trumbull & Nelson-Barber, 2019).

While culturally responsive assessment is espoused with many benefits, the scholarly discussion on cultural responsiveness of educational testing stays mostly on a pragmatic or program development level, mainly focusing on its pedagogical and procedural aspects. Seldom is the validity of score interpretations across multiple cultural groups pointed back to measures and normative knowledge that establish tests. Notably, though cultural responsiveness has been comprehensively conceptualized in teaching (see Gay, 2000; Ladson-Billings, 1995), it has not been sufficiently considered in the field of testing (Solano-Flores, 2019). Indeed, there is no research on the linkage between culturally responsive assessment and adaptive testing, despite the psychometric advantages of the latter. This chapter bridges the research gaps between theory and measurement related to this topic by applying cultural theoretical perspectives on testing measures--the items and the tests themselves. In the ensuing discussion, I posit that

items and tests are the important mediators for developing culturally responsive and equitable testing (CRET). I explore how the CRET is capable of addressing *fairness* through *item contextualization* and fulfilling *cultural responsiveness* through *test adaptation*. Conceptualizations about those items and test measurements will bring forth the cultural-theoretical and structural underpinnings of CRET. It is poised to strengthen the validity arguments of future tests of its kind for providing accurate inferences for a culturally diverse population of test-takers.

This conceptual chapter consists of three main sections. First, it aims to flesh out concepts of *item contextualization* in studying the interweaved relationships of sociocultural characteristics of contextualized items and the individual *situated* mind in testing. This serves to crystallize its theoretical cultural relevance to student performance later. Second, an expansion of the measurement considerations from equitable and fair item contextualization to culturally responsive test adaption is staged. I explore on the test level how key principles from culturally responsive theories can be conducted. Lastly, the implications and future directions are discussed.

2.2 Contextualized Item in Culturally Responsive and Equitable Testing

A substantial body of literature is reviewed to examine the seemingly intricate connection between the sociocultural characteristics of items and individuals' cognitive and psychological processes that produce observable responses. To illustrate this relationship, relevant concepts from the sociocultural theories are adapted and employed to elaborate on the contextualized items in the next sections.

2.2.1 *Contextualized Item as a Sociocultural Mediational Means*

A sociocultural perspective of learning highlights individuals' learning and development through internalizing social interaction and socially shared activities (Bruner, 1986; Rogoff, 2003; Cole, 1996; John-Steiner, 1985; Vygotsky, 1978; Wertsch 1991). As

Vygotsky (1978) explained, specifically, human higher mental functioning and psychological functioning have their origin in social activities. Cognitive processes and functioning in a learner's development, such as attention, memory, and problem solving, have cultural lines of development, with their meanings originating from social and cultural contexts, factors, institutions, and activities. To understand learners' mental functioning during testing, I need first to unravel the concept of mediation and contextualized items as semiotic means.

Notably, a defining property of human mental functioning by Vygotsky (1987) is its mediation by tools ("material or technical tools") and signs ("psychological tools"). Initially, Vygotsky expounded on the concepts of signs, or semiotic mediation, as a system of meaning-making that orient the physical and psychological world externally and internally. For example, signs of a traffic light, words in a paragraph, or an emergency exit mediate individuals' mental representations of meanings embedded in the social practice. Expanding Vygotsky's concept of mediation, Wertsch (1991) further emphasized the importance of semiotic mediation for the interdependence between individuals and social processes. He described that both forms of intrapsychological and interpsychological functioning are shaped by semiotic mediational means. This is critical for our conceptualization in this chapter and will be explained further. Likewise, Rogoff (2003) clarified that it is through cultural and collective experiences that a learner adopts socially shared experiences and acquires useful strategies and knowledge. In this perspective, she confirms how culture and language play an essential role in a learner's development. Moreover, people started to recognize that no aspects of sociocultural activity can be studied in isolation from others (p. 58). And it is a dynamic relationship where the culture and social practice itself can be transformed through individual interaction and participation with other members of their communities.

As Vygotsky (1978) highlighted the use of psychological tools, particularly language, that mediate the development of higher mental functions, it is important to note that such mediation is present during the educational testing process of students. I view that assessment items, each as a semiotic tool (Solano-Flores, 2021), mediate test takers' cognitive, mental, and psychological functioning. In other words, learners' mental functioning during testing, such as recall and differentiated interpretations of a task, can be linked to language that describes varied social, cultural, community-based, and institutional encounters and activities. It is because individual learners' mental functioning has irreducible social origin, and social interactions and contexts can be encoded through semiotic resources (van Leeuwen, 2004). Such a view echoes similar disciplinary lenses from areas like the learning sciences (e.g., Brown, 1994), cognitive science (Clark, 1998; Lave & Wenger, 1991), and sociolinguistics (Gee, 2008). For instance, on a practice level, a *situated* view of assessment "emphasizes questions about the quality of student participation in activities of inquiry and sense-making and considers assessment practices as integral components of the general systems of activity in which they occur" (Greeno et al., 1997, p. 37).

2.3 Defining Elements of Contextualized Item

From the previous discussion, we understand that semiotic mediation plays a critical role not only in learning development but also as a psychological process to assist problem-solving (John-Steiner & Mahn, 1996). Here, I attempt to define elements of contextualized items that have importance to such a process in culturally diverse learners as test-takers.

2.3.1 *Context Stimulus*

Throughout this dissertation, I define the core knowledge of an assessing item that a person is required to answer correctly as *item construct,* which solicits a person's latent

trait or true ability. The resources, including narratives, texts, visuals, and format surrounding a *situated* problem task, are called *item context,* not the assessed knowledge but what frames it. *Item context* is theoretically and psychometrically important, for it can be analyzed to ferret out item sources of construct-irrelevant variance that may otherwise remain hidden. These variances can create noise and unfairness regarding multiple groups. Here, I explain that the cultural embodiment of *item context* in assessment tasks is a semiotic embodiment of an external cultural operation or phenomenon. The embodiment of context acts analogous to "sign use" in sociocultural activity. Once perceived by a test taker, it becomes an instrument for psychological activity. From a Vygotskian perspective, it shares the same function as other sign uses in eliciting and transforming sociocultural learner's relevant knowledge, skill, or propensities, however conceived.

A token, which is often used in natural language processing, is an instance of a sequence of characters as a functional semantic unit. Broadly, they instantiate higher levels of sign types or language systems for mediation to happen (Mertz, 2013). Figure 2-1 is an example to show how tokenization can demarcate any language system and possibly classifies a string of characters describing social context into semiotic tokens, which can then be passed on to processing and categorization. As such, the entity of context embodiment, which are grouped tokens in the space of the assessment tasks, is defined as a "context stimulus".

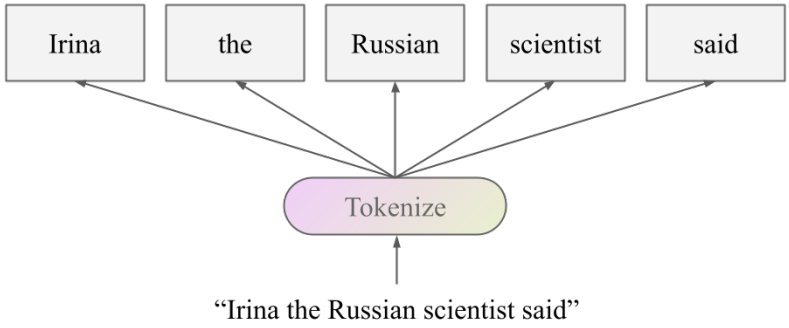


Figure 2-1. Tokens in Item Context (IC).

It is also argued here that context stimuli have relevance to historically and culturally formed systems of meanings and individuals’ cultural experiences. As a simple illustration, the context stimulus of "fishing" may be a leisure activity conceived by some groups of people but not others. It may also be a concept for indigenous groups and experiences as a cultural doing representing their daily living and community life. The different significance related to the context stimulus of "fishing" has different affordances, which may prepare individuals to act or not act in different ways in both the test setting and in the real world. In one case, when presented with other semiotic stimuli, such as actions and events about dining, fishing can be understood pragmatically as a way of providing food or as a source of protein. In another context, fishing solicits a different purpose where the indigenous tribes hold a ritual and ceremonies for the community to celebrate seasons of life or festivals as a cultural event. Thus, context stimuli represent and convey different meanings to the extent that test-takers share the commonly coded cultural experiences.

While there are lots of ways of categorizing semiotic resources in literature, for simplicity, I group the context stimuli on two levels: representational and functional item contexts. Functional context stimuli for a test item include format, mode, language adaptation, translation, accommodation, and other functionalities that situate an item

construct. Representational context stimuli, mainly reserved to present narratives and texts, focus on the “situatedness” of problems in a sociocultural framework. Indeed, the dynamic relationships of context stimuli form a semiotic system that not only connects sociocultural patterns, activities, and knowledge in thinking but also possibly scaffolds what Wertsch (1985) calls proximal developmental activities.

Among representational context stimuli, I recognize four general categories that make up the sociocultural context stimuli at play. First, a contextualized item contains actor(s) as human being(s), human agent(s), or initiator(s) in a represented social context. They can probably be considered as cultural identities (Hall, 1990). Second, activities refer to intra- and inter-personal interactions and conditions exhibited in an item context about events, practices, and interactions within situations or time. Sociocultural activities or actions in society lead to personal participation, transformation, and appropriation of meaning (Cole & Engstrom, 2007). Extending them to an item context, the semantic meaning of activities associated with actors are the carriers of meanings, information, and representation of the sociocultural process at large. Third, there are social, cultural, and historical tools and artifacts that are parts of object descriptors in the context. They can be symbols, rules, or regulations that reflect social patterns in society. The fourth type of context stimuli includes embodied environment, situation, relationships, or scenarios where social interactions take place. They may be expressed as places and localities such as nature, a classroom, a school, or a society in the context of interest. Lastly, it needs to be noted that the four types of stimuli often interact and overlap, forming and exerting a dynamic relationship. The "sociocultural interconnectedness" of the above semiotic stimuli captures a dynamic relationship between changing meanings and semiotic stability within certain contexts.

2.3.2 *Context Stimuli as An Equitable Consideration*

Increasingly, students are exposed to a wide variety and quality of assessment items with regard to their contents and contexts. An equity orientation toward context construction would support the general claims frequently made by culturally responsive pedagogy (Gay, 2010) to benefit culturally diverse students as test-takers. Conversely, without providing culturally valid background knowledge and an equitable orientation for context stimuli, contents, topics, and texts can have more negative effects on historically and culturally marginalized learners.

Representative Item Context. Constructing effective and equitable context stimuli involves, first and foremost, culturally valid and relevant examples. In a learning context, learning scientists espouse the role of community and social relationships as essential for learning and development. Many studies and projects (Crawford, 1993; Gauvain, 2001) show that the academic success of culturally diverse student populations can be notably improved by linking to their own local cultural communities, customs, traditions, and artifacts. In an item context, we can revive and acknowledge those social networks and cultural communities that are beneficial for culturally diverse learners. For example, the stories and biographies of ethnic female scientists and mathematicians of color are relevant for female and ethnic minority groups. Similarly, weaving into the concept of *item context* with equitable and effective social and cultural stimuli help define the purview of problem tasks, including their settings, goals, purposes, and values.

Functional Item Context. Based on culturally responsive principles, constructing effective functional item context stimuli involves protocols and procedures of preparing various context stimuli to be “more synchronized with the mental schema, participation and thinking styles, and experiential frames of reference for diverse ethnic groups”

(Gay, 2010, p. xxiv). For example, to accommodate cultural diversity, item-level content can be administered in multiple languages. Diversifying context stimuli format and mode is recommended. Some cultural minority groups, as suggested, prefer visual, auditory, and tactile stimulations in their learning and sense-making (see also Irvine & York, 1995; Yamazaki, 2005). For example, Tuck and Boykin (1989) show African American students perform significantly better when exposed to varied problem-solving task formats, meaning, and academic tasks (i.e., spelling, vocabulary, mathematics, and picture-sequencing) in both low- and a high- variability contexts. The students were also assessed for task motivation in the two variability contexts. Results revealed that academic task performance and task motivation were superior when tasks were presented with greater variability.

In their work with Native American children, John-Steiner and Osterreich (1975) found that children use various learning styles and modalities to accomplish their learning goals, sometimes a combination of visual and verbal stimuli. Therefore, having multiple modes and formats of items such as visuals, procedural, or abstract contexts helps match diverse students' learning styles. By making context stimuli culturally responsive to each cultural learner, students can interact more and engage more with the context. In this way, it may also be possible that students as test-takers, by and large, assume more agency to perceive and conceive of information. All those considerations demonstrate the possibilities of flexible, functional *item contexts* for accommodating differences in students' linguistic and cultural backgrounds, helping contribute to a testing program or environment that is culturally safe, customized, comfortable, and efficient.

Another equitable consideration from culturally responsive pedagogy is to maintain high and holistic academic expectations of all students. Besides estimating and

evaluating the information between learners and their corresponding social stimuli, *item contexts* should consider providing new appropriate development stimuli to afford students' knowledge application and transformation from knowledge to actions. To various extents, *item contexts* should also be considered as part of the curriculum to afford students a diverse range of necessary experiences to broaden their viewpoints or capacities. As a result, students are more likely to be successful in future scenarios and applications. That is, provided that we gain information about the students and their current environment, *item contexts* can extend and enrich students' understanding and perspectives. It should also help enculturate students into more authentic and diverse scenarios, environments, and practices. When making sense of the knowledge of the world, test-takers as learners need to understand the objects and contexts of other people and cultures.

2.3.3 *Knowledge Construction*

Knowledge represented in assessment, as a system of meanings, is a sociocultural product that is created by human agents whose knowledge themselves is situated in their sociocultural background and positionalities (Banks, 1993; Harding, 1993). As Banks (1993) argues, knowledge construction is significantly influenced by the group experiences of the knowers. Thus, assessment instruments and items are influenced by the perspectives of the educators and scientists who write them. They differ as to which of the aspects of standards, human thinking, and learning topics they bring to the foreground, and consequently, in terms of the nature and instantiation of assessment arguments cast in light.

In view of knowledge epistemology, it is argued that even in disciplinary science, there is more than one way of acquiring, validating, representing, and contextualizing knowledge about the world. In science education, science epistemology

can be represented by varied yet equally valid approaches to exploring the same question/topic (Harding, 1993). For example, research indicates that the indigenous knowledge of science and nature represents well-rounded, dynamic, and holistic ways of knowing (e.g., Bang & Medin, 2010; Carjuzaa & Ruff, 2010; Lipka et al., 2014). The various ethnic- and culture-specific approaches to grasping knowledge regularities, developing ways of reasoning, and putting problems into a coherent framework are not at all subsumed or superseded by generic approaches of mainstream science. Hence, knowledge needs to be scrutinized and established through multiple lenses of epistemology.

The descriptive nature of semiotic resources in *item contexts* demands culturally valid and fairly distributed representations of knowledge. From a distributional perspective (Mislevy, 1994), it is true that the interpretations of context representations on the individual level often show wide variation. Some test-takers prefer the notions, concepts, and meaning in mathematics and science to be directly expressed by concise words, charts, pictures, and tables, while some may prefer descriptive expressions, specified goals, purposes, values, or dialectic discussion which require provoking higher mental functions. Yet, a key problem in the learning and assessment of culturally diverse groups is that students often feel the tests do not reflect their cultural experiences (Claypool & Preston, 2021; Trumbull & Nelson-Barber, 2019). They need to have a sense of belonging in their school life and have pride in their heritage culture.

Trumbull and Nelson-Barber (2019) argued that assessment needs to incorporate local cultural knowledge, tap on students' prior knowledge, and be attentive to the language and syntax familiar to local students and cultural communities. For instance, some analysis of science context (Wang et al., 2019) shows that context knowledge is more authentic when it connects to students' daily learning and living, such as hands-on

activities, experiments, daily practices and routines, and personal, community-based, and cultural experiences. Some science contexts have incorporated tools and building blocks that students are familiar with at home or in schools, and simulation models that are more interactive and relatable to students. In doing so, not only is knowledge responsively and justly distributed, but also student test-takers are more active than passive receivers of knowledge.

2.3.4 *Item Contextualization and Cognitive Performance*

In this section, I examine in more detail how the use of language, choice of words, and implicit narratives in contextualized items can affect test-takers' psychological, cognitive, and affective functions. A key concept in Vygotsky's theory is the notion of internalization, a process whereby individuals actively appropriate and reconstruct external, shared operations such as forms of mediation on the internal plane (Vygotsky, 1997). The linking of the narratives or semiotic means and the psychological mind is metaphorically comparable to the meditation and internalization of the mind in sociocultural processes. As Bruner (1962) puts it, "… the tools and aids that do are the developing streams of internalized language and conceptual thought that sometimes run parallel and sometimes merge, each affecting each other" (p. vii).

In Figure 2-2, semiotic mediation through signs and language (e.g., in *item context*) can be seen as the main intrapsychological function in testing. Likewise, social mediation through signs and symbols plays an important role in social beings' growth (in social contexts). To some degree, the following can all be internalized forms of thought after mediation:

- cognitive mental functioning such as memory, perception, attention
- higher psychological functions such as allocation, regulation, planning, and strategy of cognitive functions

- affective states such as interest and anxiety.

As Vygotsky (1978) declared, semiotic mediation and social mediation are inevitably intertwined to produce individual growth in a given historical and cultural context. Similarly, internalizing through semiotic *item context* stimuli can be analogous to internalization through social signs in the process of learning development

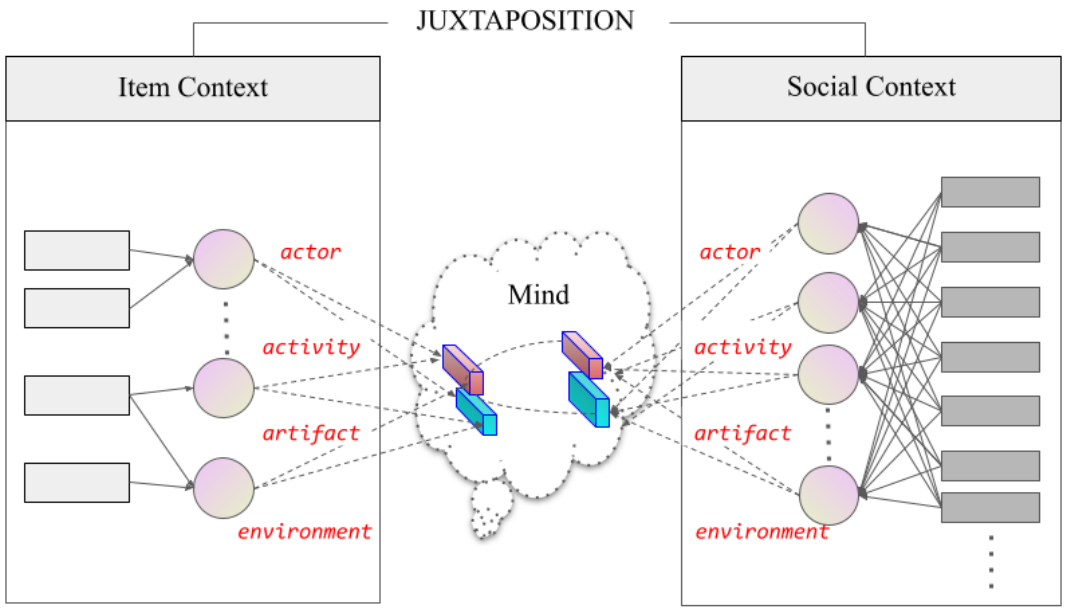


Figure 2-2. A juxtaposition of context internalization in testing and learning.

I view intersubjectivity as a shared psychological space, which is the result of juxtaposing internalized subjectivity through social and semiotic mediation, a critical understanding of the contextualized item model in this study. Basically, through mediation and internalization, the immediate semiotic item contexts and item parameters afford cognitive and behavioral performance by connecting psychological variations to their relevant social processes. For instance, a student learner's cognitive functions in understanding the force of friction may be mediated by the relevant social experiences when playing with the wheels and brakes of a toy car or by accessing semiotic representations of a toy car context.

It is worth noting that the properties of context internalization in testing and learning emphasize equal or similar structural affordances shared in the intersubjectivity space. The concept of affordance from social cognition (Lave & Wenger, 1991) gives us indications of how learners/respondents perceive and act on the external shared or situated stimuli instantaneously and without hesitation. But individual differences may exist between the structurally equivalent social and semiotic stimuli. To illustrate, say a learner's higher psychological plane, such as interest and attention to a specific experimental task, is mediated by the social relationships between the teacher, the group leader, and the learner. The student learner quickly internalizes the social meaning of authority and hierarchy behind those social subjects. Now through intersubjectivity, it is hypothesized that the learner as a test-taker has similar structural affordances of meanings conveyed by a similar task situated by words such as "scientist", "expert", "volunteer", and "student". Outcomes of the learner's psychological functions (e.g., allocation of attention, motivation) and affective state are comparable. While the notion of social hierarchy influences these, the impacting stimuli are not exactly the same.

This points to the important issue of item bias in contextualized assessment. Research supports that psychological functioning such as anxiety, negative thinking, and mind-wandering have negative effects on test performance among triggered or "threatened" individuals by influencing and coopting working-memory resources (Nguyen & Ryan, 2008; Pennington et al., 2016). These effects need to be minimized, by linking the psychological process and affective factors in which students or test-takers relate semiotic inputs or stimuli to broader contexts among multiple systems of culture between students, within schools, and in society. A large body of socio-psychological research (Aronson et al., 1999; Dee & Domingue, 2021; Steele, 1997; Steele & Aronson, 1995) shows that the activated psychological and common formation

of implicit bias in testing raises stereotype threats and heightened anxieties, which can result in the underachievement for students who belong to a cultural minority group. As students solve contextualized problems along the test, bias may be represented, activated, and perceived through the representation of social and cultural contexts. In this way, item contextualization depends in part on the sociocultural and psychological perspectives of testing in which warrants are framed.

2.4 Adaptive Testing in Culturally Responsive and Equitable Testing

Recognizing the underlying social and cultural mechanism in the process of context-reading and problem-solving yields a way to probe into the measurement components between test-taker, item, and test. That is, *item context* matters, and the way to construct *item context* equitably matters to leveraging testing for fairness for culturally diverse learners. Whilst item contextualization organizes semiotic resources which render the situated social and cultural process of learning governable, it alone cannot construct a culturally responsive test. The reason is that, in traditional testing, an item is administered to all students, regardless of the differences in individual learners. Hence, an *item context* may be perceived as biased or inaccessible to one group but not others. This is an inherent weakness of standardized testing, wherein one measurement instrument cannot be the best fit for everybody.

Such an issue can be described in measurement terms as generalizability. On the test level, a traditional non-adaptive test can address cultural heterogeneity by either increasing its generalizability across socio-cultural contingencies or limiting its generalizability to a specific target population (Solano-Flores, 2011). Generalizability can be threatened by a multitude of construct-irrelevant factors that potentially introduce measurement error. So, the goal is to minimize the discrepancy between a true trait level and the observed scores for each test-taker and any statistical bias against

groups. For instance, the discrepancies can be related to assorted reasons, such as test-takers' cognitive traits, motivation, mental health, stress, behavioral or psychological tendencies, and their social and cultural capital, which tend to have their origin in the sociocultural activities and encounters (Demmert, 2005; Kim & Zabelina, 2015; Trumbull et al., 2015).

Given the students' cultural heterogeneity, multicultural educators would, in one-on-one assessment, use multifarious information about each student to engage them in effective evaluation and adjust their questions based on immediate interactions with the student. In a similar vein, adaptive testing creates this dynamic. It uses the maximum statistical information from input, which can be past observed responses and other supplementary data. Then, the best matching item is selected algorithmically for each student in real-time to best demonstrate their competencies. In contrast to a non-adaptive test in which certain items only give useful results for learners of preference, an adaptive test as a personalized assessment skips relatively easy items for more competent students and relatively difficult items for less competent students. As such, the outward adaptive technology scales up conditions for less standardized but more individualized, localized, and prime measurement conditions. Nevertheless, what is still lacking for adaptive testing is that by administering selectively, items can be individualized, effective, meaningful, and equitable to culturally diverse learners. Enabled by big data and learning analytics, adaptive tests are able to drive these unfettered areas of possibilities if properly designed based on culturally responsive principles.

On cultural responsiveness, the field witnesses substantive literature demonstrating the strength of culturally responsive teaching as a pedagogical tool in connecting to students' local, cultural, and community knowledge while adapting to

students' cultural frames of reference (see Gay, 2000; Ladson-Billings, 1994). According to Gay (2013), the key strategy of culturally responsive pedagogy is channeling the strengths and assets of culturally diverse students and communities and leveraging those strengths to improve their personal agency and educational achievement (p.68). Shifting from a deficit to a strength-based perspective, Ladson-Billing (1995) maintains that teaching that brings on students' cultural strengths motivates and engages students better in their school life. As in assessments, students bring in not only their prior knowledge and experience but also their cultural ways of thinking, engaging, and communicating. Assessments such as the Kaiapuni Assessment of Education Outcomes (KĀʻEO) are accordingly developed based on indigenous systems of language and knowledge (Hawaii State Department of Education, 2021; see also Bang & Medin, 2010; Lipka et al., 2014 for more) and ways of teaching and learning (e.g., Bang & Marin, 2015). Thus, the cultural responsiveness of assessment makes it student-centered by accepting their cultural ways of knowing and thinking.

In line with culturally responsive teaching and practice, adaptive testing should be conceptualized, with technical knowledge, to adaptively implement cultural responsiveness "using the cultural knowledge, prior experiences, frames of reference, and performance styles of ethnically diverse students" (Gay, 2010, p.31). For instance, culturally responsive teaching affirms the legitimacy of cultural identities, cultural heritages, and histories that affect students' dispositions, attitudes, and ways of knowing. Considering *item context* as a semiotic tool and sociocultural context stimuli as capital in semiotic forms, the responsive function of the CRET lies in the key process of selecting and appropriating socially and culturally available psychological tools and semiotic resources to a culturally diverse learner to assist their prime state of problem-solving (See Figure 2-3). Given the cultural identities of students taking the test, the

corresponding context stimuli can be compiled to adapt to students' cultural backgrounds and the required instructional coverage.

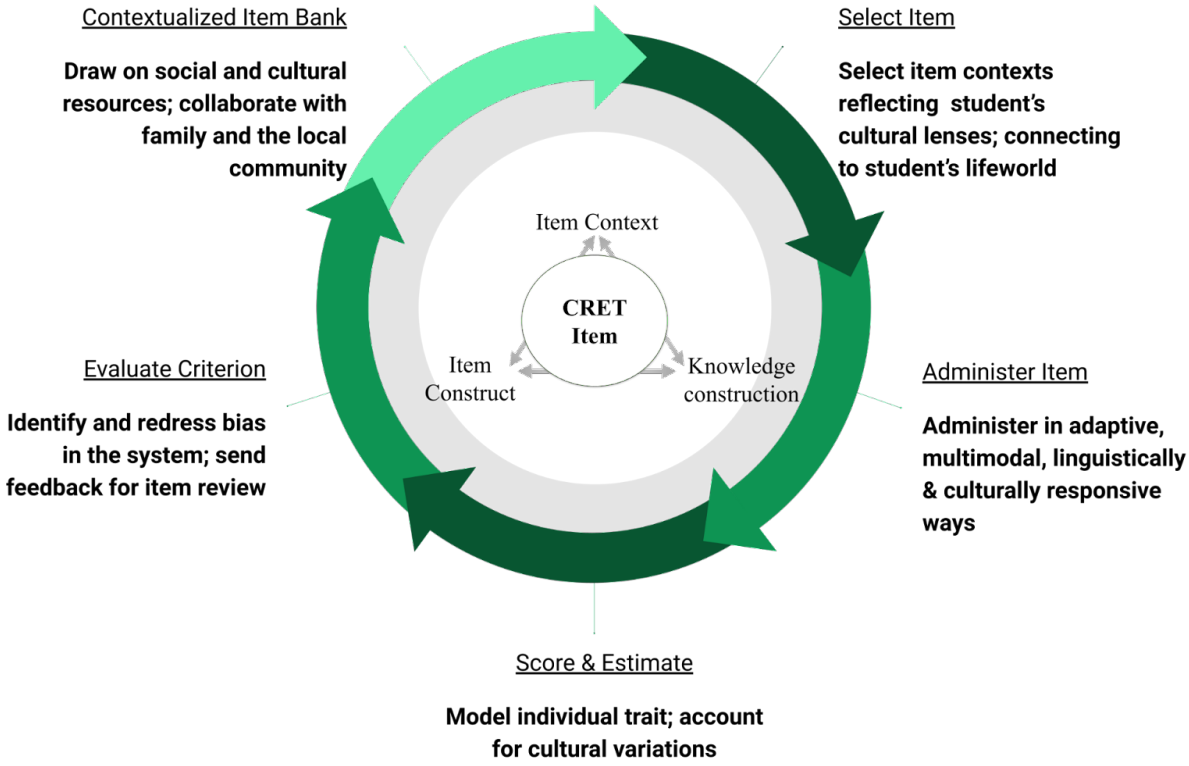


Figure 2-3. A conceptual framework for CRET via adaptive testing.

An efficient and culturally responsive approach increases test-takers' motivation and engagement with enhanced performance outcomes. For culturally responsive assessment, teachers can facilitate students' participation in assessment situations in different ways based on culturally responsive pedagogy. In a similar vein, adaptive testing can benefit from diverse assessment and response formats. They engage students better and accommodate student needs or cultural communication (Luykx et al., 2007; Noble et al., 2012; Solano-Flores & Nelson-Barber, 2001). Imagine how a responsive design of item administration can boost inclusiveness systematically. CRET can adaptively administer items appropriate for test-takers with special needs and accommodations such as text-to-speech, pop-up glossary, and extended time.

Collectively, the item bank can be constructed by test developers and researchers, drawing upon the cultural community, experts, teachers, and students to align the assessment instruments with local cultural-community practices (Scott et al., 2013). From a knowledge perspective, CRET is attuned to respect students' prior knowledge by appropriating cultural knowledge of students and the community and their cultural practices, indirectly encouraging academic achievement.

Evaluations of student scores from adaptive tests should be based on criteria of accuracy, fairness, and cultural validity. An equitable consideration for a test and an item shifts the emphasis of the intended test population away from the majority demographic group to the inclusion of all groups. It focuses more on those groups who might be disproportionately affected by the negative consequences of testing. A responsiveness consideration can evaluate on the test level if the overall construction of different historical and cultural context stimuli encourages different psychological routes to a given developmental endpoint (Miller, 2011). With that, each item context and the overall context representation of the test will be culturally appropriate, balanced, and valid, especially for culturally diverse groups.

2.5 Conclusion and Discussion

This chapter attempts to conceptualize test elements to empower CRET, a possibility of testing that acknowledges the cultural significance and legitimacy of content and contextual knowledge. It adopts a pedagogy of cultural responsiveness and demonstrates equity considerations. It provides a common conceptual ground for the three case studies that come later in this dissertation, which explore the technical details of fairness in contextualized assessments and computerized adaptive testing. Implications of the CRET point back to the test developers, researchers, and teachers.

For researchers in measurement, *item contextualization* creates a new window to examine the relationship between individuals' mental and psychological functioning and the broader social processes. However, using information processing theory alone in test calibrations and psychometric analyses cannot explain complex phenomena, such as why some test-takers can manage an overload of cognitive complexity at ease while others are having trouble. Culturally responsive testing would remedy this by constructing applied and authentic contexts that increase local responsiveness, cultural representation, and equitable knowledge construction. From a quantitative approach, there are myriad configurations of context details to embody social context, in which we can test out statistical relationships of the psychological and the social. The caveat is that reducing the individuals represented in the context to group identities, such as race and gender, risks stereotyped monolithic identities that are isolated, fixed, and flattened. Nevertheless, linking student, item, and process characteristics together in latent cultural characteristics of groups and adaptive administration will reduce oversimplification and enhance cultural nuances.

Culturally responsive principles need to apply beyond the representation of item contexts to the structures of tests, settings, standards, scoring rubrics, the interpretation, consequences, and use of scores, and so forth. The development of normative documents needs to continually include culturally- and ethnically- diverse scientists, communities, and students to enact the knowledge agency for culturally diverse learners. Regarding social consequences of testing, public communication from test publishers should demand a variety of validity evidence, including score interpretations and use consequences of scores concerning cultural validity (Shepard, 1993; Solano-Flores, 2011). The equitable principle of item contextualization expects test developers to publish student interviews and reports validating the interpretation and uses for a

range of cultural characteristic item contexts. The culturally responsive principle would require extra attention for reporting group differences in test performance. Any meaningful inferences should be framed by the corresponding contextual differences between groups, as well as with transparency the intended test design, purpose, social consequences, and caveats from claims.

For knowledge construction more broadly, such as in large-scale assessments, it is critical to ensure cultural validity to build knowledge that is accessible and equitable for all students rather than unintentionally favoring one group or the mainstream population. For instance, research indicates that incorporating various examples of cultural scientists in the curriculum helps students develop a clearer understanding of the role of culture in the history of science (e.g., Bang & Medin, 2010). Likewise, test developers can conduct and publish qualitative interviews to represent scientific topics and the broader science-situated context. Documentation, including cognitive interviews, can confirm the benefits of the CRET and demonstrate other areas of holistic outcomes. For instance, with equitable representation of context, students learn to acknowledge and appreciate the contributions made by individuals from other different cultures.

Some may argue that to make assessments accessible to all students, regardless of racial or cultural background, we must remove any construct-irrelevant variable, including culture-specific contextual items. This may unintentionally privilege or hinder the test performance of cultural minority groups. Many scholars observe that current psychometric analyses, such as differential item functioning, do not unravel the root problems of item bias (Yildirim & Berberoğlu, 2009; Zumbo, 2007). By programming and experimenting with *item context* configurations, the CRET has the possibility of

finding specific context stimuli that are associated with item bias. Thus, it helps mitigate bias and optimize understanding of targeted constructs equitably.

Indeed, more research is needed to implement and prove the actual possibilities of the CRET. For future studies, more conceptual studies about the CRET testing processes and technical parameters need to be delineated. CRET calls for the re-designing of algorithms in the current computerized adaptive testing to adaptively use culture-specific contextual items, as well as more measures and indicators to be developed to investigate whether such culturally relevant materials will increase test engagement, interest, and inclusion for a culturally diverse test population. More qualitative and quantitative research considering cultural validity is needed to bridge the CRET ideas with the data evidence. Altogether, these ideas and evidence will underscore the potential significance that equitable and culturally responsive testing has in addressing educational diversity and equity issues.

Last but not least, the illustrations through CRET are by no means to discard or diminish the culturally responsive practices among teachers and schools. Teachers can contribute to the input components of adaptive tests through pre-specification of standards and testing conduct. In return, the instantaneous output of CRET enriches student profiles and can be part of the teacher's formative feedback as a fair and responsive process. In a classroom testing setting, teachers can localize and individualize *item contexts* through their knowledge of student characteristics. In giving specific contexts or scenarios that they know students have good encounters with, teachers enhance communication in giving assessment feedback. Therefore, professional development can incorporate the method of teachers analyzing, evaluating, and responding to cultural situations before, during, and after the testing process. It can

be said that CRET cultivates students' motivation and engagement with testing through both the test instrument and teacher practice.

Reference

Ahmed, A., & Pollitt, A. (2007). Improving the quality of contextualized questions: An experimental investigation of focus. *Assessment in Education*, *14*(2), 201-232.

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

Bang, M., & Marin, A. (2015). Nature-culture constructs in science learning: Human/non-human agency and intentionality. *Journal of Research in Science Teaching*, *52*(4), 530-544. https://doi.org/10.1002/tea.21204

Bang, M., & Medin, D. (2010). Cultural processes in science education: Supporting the navigation of multiple epistemologies. *Science Education*, *94*(6), 1008-1026. https://doi.org/10.1002/sce.20392

Banks, J. A. (1993). The canon debate, knowledge construction, and multicultural education. *Educational researcher*, *22*(5), 4-14.

Beaumont, C., De Valenzuela, J. S., & Trumbull, E. (2002). Alternative assessment for transitional readers. *Bilingual Research Journal*, *26*(2), 241- 268. https://doi.org/10.1080/15235882.2002.10668710

Brown, A. L. (1994). The advancement of learning. *Educational Researcher*, *23*(8), 4-12. https://doi.org/10.3102/0013189x023008004

Bruner, J. (1962). Introduction. In L. S. Vygotsky, *Thought and language* (pp. v-x). MIT Press.

Carjuzaa, J., & Ruff, W. G. (2010). When Western epistemology and an Indigenous worldview meet: Culturally responsive assessment in practice. *Journal of Scholarship of Teaching and Learning*, *10*(1), 68-79.

Clark, A. (1998). *Being there: Putting brain, body, and world together again*. MIT Press.

Claypool, T. R., & Preston, J. P. (2021). Analyzing assessment practices for Indigenous students. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.679972

Cole, M. (1996). *Cultural psychology: A once and future discipline*. Harvard University Press.

Cole, M., & Engestrom, Y. (2007). Cultural-historical approaches to designing for development. In J. Valsiner & A. Rosa (Eds.), *The Cambridge handbook of sociocultural psychology (*pp. 484-507*)*. Cambridge University Press.

Dee, T. S., & Domingue, B. W. (2021). Assessing the impact of a test question: Evidence from the "Underground Railroad" controversy. *Educational Measurement: Issues and Practice*, *40*(2), 81-88.

Dong, D. (2020). *What do we know about context: An integrated analysis of context characteristics of science assessment item*, [Doctoral dissertation, University of Washington, Seattle, WA]. ProQuest Dissertations and Theses. Retrieved from http://hdl.handle.net/1773/45482.

Gauvain, M. (2001). *The social context of cognitive development*. New York: Guilford.

Gay, G. (2010). *Culturally responsive teaching: Theory, research, and practice* (2nd ed.). Teachers College Press.

Gay, G. (2013). Teaching to and through cultural diversity. *Curriculum Inquiry*, *43*(1), 48-70. https://doi.org/10.1111/curi.12002

Gee, J. (2008). A sociocultural perspective on opportunity to learn. In P. Moss, D. Pullin, J. Gee, E. Haertel, & L. Young (Eds.), *Assessment, equity, and opportunity to learn* (Learning in doing: Social, cognitive and computational perspectives, pp. 76-108). Cambridge University Press. doi:10.1017/CBO9780511802157.006

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1997). Cognition and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 15-47). Simon and Schuster Macmillan.

Hall, E. T. (1990). Unstated features of the cultural context of learning. The *Educational Forum*, *54*(1), 21-34. https://doi.org/10.1080/00131728909335514

Harding, S. (1993). Rethinking standpoint epistemology: What is 'Strong Objectivity'?. In L. Alcoff & E. Potter (Eds.), *Feminist Epistemologies*. Routledge.

Hawaii Department of Education. (n.d.). Kaiapuni assessment of education outcomes (KĀ'EO). Retrieved April 14, 2022, from https://www.hawaiipublicschools.org/TeachingAndLearning/Testing/KAEO/Pages/home.aspx#

Hood, S. (1998). Culturally responsive performance-based assessment: Conceptual and psychometric considerations. *The Journal of Negro Education*, *67*(3), 187. https://doi.org/10.2307/2668188

Irvine, J. J., & York, D. E. (1995). Learning styles and culturally diverse students: A literature review. In J. A. Banks & C. A. Banks (Eds.), *Handbook of research on multicultural education*. Macmillan.

Jacob, R., & Parkinson, J. (2015). The potential for school-based interventions that target executive function to improve academic achievement: A review. *Review of Educational Research*, *85*(4), 512-552.

John-Steiner, V. (1985). *Notebooks of the mind: Explorations of thinking*. University of New Mexico Press.

John-Steiner, V. P., & Osterreich, H. (1975). *Learning styles among Pueblo children: Final report to National Institute of Education*. University of New Mexico, College of Education.

John-Steiner, V., & Mahn, H. (1996). Sociocultural approaches to learning and development: A Vygotskian framework. *Educational Psychologist*, *31*(3), 191-206. https://doi.org/10.1207/s15326985ep3103&4_4

King, D. (2012). New perspectives on context-based chemistry education: Using a dialectical sociocultural approach to view teaching and learning. *Studies in Science Education*, *48*(1), 51-87.

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, *32*(3), 465-491. https://doi.org/10.3102/00028312032003465

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

Lee, C. D. (1998). Culturally responsive pedagogy and performance-based assessment. *The Journal of Negro Education*, *67*(3), 268. https://doi.org/10.2307/2668195

Lipka, J., Mohatt, W. G., & Ilutsik, E. (2014). *Transforming the culture of schools: Yup¡k Eskimo examples* (2nd ed.). Routledge.

Luykx, A., Lee, O., Mahotiere, M., Lester, B., Hart, J., & Deaktor, R. (2007). Cultural and home language influences on children's responses to science assessments. *Teachers College Record*, *109*(4), 897-926.

Mertz, E. (2013). *Semiotic mediation: Sociocultural and psychological perspectives*. Elsevier.

Miller, P. (2011). *Theories of developmental psychology* (5th ed.). Worth Publishers.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika*, *59*(4), 439-483. https://doi.org/10.1007/bf02294388

Montenegro, E., & Jankowski, N. A. (2017, January). *Equity and assessment: Moving towards culturally responsive assessment* (Occasional Paper No. 29). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).

Nguyen, H. H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, *93*(6), 1314.

Noble, T., Suarez, C., Rosebery, A., O'Conner, M.C., Warren, B., & Hudicourt-Barnes, J. (2012). "I never thought of it as freezing": How students answer questions on large-scale science tests and what they know about science. *Journal of Research in Science Teaching*, *49*(6), 778-803.

Pennington, C. R., Heim, D., Levy, A. R., & Larkin, D. T. (2016). Twenty years of stereotype threat research: A review of psychological mediators. *PloS one*, 11(1), https://doi.org/10.31234/osf.io/j9ae5

Rogoff, B. (2003). *The cultural nature of human development*. Oxford University Press.

Rogoff, B. & Wertsch, J. (Eds.). (1984). *Children's learning in the "zone of proximal development": New directions for child development* (No: 23). San Francisco. Jossey-Bass.

Rose, D. H., Gravel, J. W., & Gordon, D. T. (2014). Universal design for learning. *The Sage Handbook of Special Education: Two Volume Set*, 475-489. https://doi.org/10.4135/9781446282236.n30

Ruiz-Primo, M. A., & Li, M. (2015). The relationship between item context characteristics and student performance: The case of the 2006 and 2009 PISA science items. *Teachers College Record: The Voice of Scholarship in Education*, *117*(1), 1-36. https://doi.org/10.1177/016146811511700111

Scott, S., Webber, C. F., Lupart, J. L., Aitken, N., & Scott, D. E. (2013). Fair and equitable assessment practices for all students. *Assessment in Education: Principles, Policy & Practice*, *21*(1), 52-70. https://doi.org/10.1080/0969594x.2013.776943

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, *19*, 405. https://doi.org/10.2307/1167347

Sigman, M., & Ruskin, E. (1999). Chapter IV. Social and emotional responsiveness. *Monographs of the Society for Research in Child Development*, *64*(1), 54-65. https://doi.org/10.1111/1540-5834.00005

Solano-Flores, G. (2011). Assessing the cultural validity of assessment practices: an introduction. In M. R. Basterra, E. Trumbull, and G. Solano-Flores (Eds.), *Cultural validity in assessment: Addressing linguistic and cultural diversity* (pp. 3-21). Routledge

Solano-Flores, G. (2019). Examining cultural responsiveness in large-scale assessment: The matrix of evidence for validity argumentation. *Frontiers in Education*, *4*. https://doi.org/10.3389/feduc.2019.00043

Solano-Flores, G. (2021). The semiotics of test design: Conceptual framework on optimal item features in educational assessment across cultural groups, countries, and languages. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.637993

Solano-Flores, G., and Trumbull, E. (2008). In what language should English language learners be tested?. In R. Kopriva (Ed.), *Improving large-scale achievement tests for English language learners* (169-200). Erlbaum.

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

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*(6), 613-629.

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.

Trumbull, E., & Nelson-Barber, S. (2019). The ongoing quest for culturally-responsive assessment for Indigenous students in the U.S. *Frontiers in Education*, *4*. https://doi.org/10.3389/feduc.2019.00040

Tuck, K., & Boykin, A. W. (1989). *Task performance and receptiveness to variability in Black and White low-income children* [Paper presentation]. The Eleventh

Conference on the Empirical Research in Black Psychology. Washington, DC, United States.

van Leeuwen, T. (2004). *Introducing Social Semiotics*. Routledge. doi: 10.4324/9780203647028

Vanneman, A., Hamilton, L., Baldwin Anderson, J., and Rahman, T. (2009). Achievement gaps: How Black and White students in public schools perform in Mathematics and Reading on the National Assessment of Educational Progress, (NCES 2009-455). National Center for Education Statistics. https://files.eric.ed.gov/fulltext/ED505903.pdf

Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Vygotsky, L.S. (1987). Thinking and speech. In R. W. Rieber & A. S. Carton (Eds.). *The collected works of L.S. Vygotsky* (N. Minick, Trans.) (pp. 37-285). Plenum.

Wang, N., Li, M., & Dong, D. (2019, April). *Measuring equitable contextualized items in science assessment* [Paper presentation]. American Educational Research Association Annual Meeting. Toronto, Canada.

Wang, T. (2016). *Examining sequence of contextualized items in science-experimental evidence on English learners (ELs) and Non-ELs* [Doctoral dissertation, University of Washington, Seattle, WA]. Retrieved from http://hdl.handle.net/1773/35566

Wertsch, J. V. (1985). *Vygotsky and the social formation of mind*. Harvard University Press.

Wertsch, J. V. (1991). *Voices of the mind: A sociocultural approach to mediated action*. Harvard University Press.

Yamazaki, Y. (2005). Learning styles and typologies of cultural differences: A theoretical and empirical comparison. *International Journal of Intercultural Relations*, 29(5), 521-548. https://doi.org/10.1016/j.ijintrel.2005.07.006

Yildirim, H. H., and Berberoğlu, G. (2009). Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. *International Journal of Testing*, *9*(2), 108-121. https://doi.org/10.1080/15305050902880736

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

Chapter 3. Understanding Cultural Contexts in Contextualized Science Assessment: Concepts, Measures, and an Illustrative Analysis

Nixi Wang [1], Min Li [1], Dongsheng Dong [1]

[1] *College of Education, University of Washington, USA*

Abstract

Cultural validity in measurement has grown out of the classical validity theory to attest to culturally valid interpretations of test scores, which is substantial to address test fairness, bias, and equitable test development. This study proposes an equitable and culturally valid framework for addressing the cultural contexts of items, which incorporates three domains, including cultural equity, context features, and knowledge construction of item context. Using PISA science contextualized assessment items as an example, we employ the newly developed set of rubrics to code cultural context attributes. The analysis is illustrative for item reviews that evaluate item context in contextualized assessment. A mix-methods approach using a clustering algorithm, profiling, and case analyses identified important item contextual features, including cultural characteristics, linguistic accessibility, and others, in predicting item difficulty and Differential Item Functioning (DIF) parameters. The results discuss relationships between context representation, science knowledge, cultural validity, and fairness. We thus advocate in educational measurement a cultural and critical perspective regarding the cultural validity of items or item contexts. Implications point to test development and validation.

3.1 Introduction

A test's validity (Kane, 1992, 2006) is closely tied to considerations of fairness, wherein fair and valid items have functions of invariance as a desired item property. An invariance assumption holds when an item performs in the same way across different subpopulations, and does not produce systematic bias in test scores against certain groups. While such an assumption is normally tested through established psychometric and statistical procedures such as Differential Item Functioning (DIF), we know less about the causes of DIF, but rendering items for content review. According to Messick (1987), a validity argument for assessment is "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). Importantly, integrative judgments of test items should strive to address whether items are fair to all student populations. We argue that, with the achievement and opportunity gaps noticeable in education, we need to examine the epistemological underpinnings of assessment knowledge to be to be fair and equitable for culturally diverse groups of students.

The concept of cultural validity sees what the test intends to measure as mediated by sociocultural factors that shape assessment knowledge, as well as by students' ways of knowing and thinking (Solano-Flores & Nelson-Barber, 2001; del Rosario Basterra et al., 2011). When an assessment is administered to a culturally heterogeneous population, close attentions to the test's cultural validity claims are legitimately called. Solano-Flores (2011) explicated the principles of cultural validity in four aspects: theoretical foundations, population sampling, item views, and test views. For item reviews, one can argue that cultural validity is seriously threatened when items are biased against minority cultural groups. To conduct item reviews for cultural

validity in educational assessment, including STEM assessments, a sociocultural/multicultural perspective foregrounds knowledge subjectivity and multicultural differences. Feminist theorist Sandra Harding (1993) once explained the "strong objectivity" of scientific knowledge, and acutely pointed out that knowledge is inseparable from its producer, and when the knowledge construction is preconditioned on social and political factors and actors, its embedded epistemology and the ideology should be open to scrutiny.

For a long time, item context had rich supplemental information surrounding an item construct, through either a relevant scenario, phenomenon, or problem set-up. We consider item context as a part of the knowledge pertaining to test validity, particularly cultural validity, and acknowledge the cognitive effects and cultural meanings generated from item context during students' test taking. When it activates misunderstandings or wrong affective cues, item context can also be a possible venue for causing construct-irrelevant variance and DIF. In this perspective, we explore worthwhile questions of "what an equitable and culturally valid item context is", and "how we can evaluate cultural characteristics of the context in relation to item parameters for validity and equity".

3.2 Cultural Validity and Science Contextualized Knowledge

Contextualizing scientific knowledge in the PISA science assessment has been a common practice to assess students' complex thinking and transfer of knowledge (OECD, 2017; Haladyna, 1997). Research supports that contextualized items in science assessment promote students' cognitive process and understanding of scientific concepts, practice, and transfer of knowledge (Ahmed & Pollitt, 2007; Ruiz-Primo & Li, 2015; Little & Jones, 2010). For example, the utility of real-world context in science

test items has pedagogical benefits in building an authentic assessment from students' prior knowledge (Kusimo et.al, 2000; Cooper & Dunne, 2000; Boaler, 1994). The PISA assessment has developed a related scientific framework for eliciting students' applied knowledge on interpreting science phenomena and scientific practices.

The challenge, however, lies in potential inequities due to the variation of context interpretations by social and cultural groups. Equity is the belief that student achievement increases when culturally diverse students receive opportunities that allow them to draw on their social and cultural literacies in order to be academically successful (Stembridge, 2019). Historically, Bernstein's study (1996) suggested that children from the working-class background were more likely than their peers from the middle-class backgrounds to misrecognize the intended nature of certain specialized problem-solving contexts. For instance, children from minority groups or the working-class background tended to sort food items by reference to how the food items appeared together to them in their households and everyday life. There were also modal differences in how working and middle-class children make sense of and sort items. Indeed, many researchers point out the significance of social class and how it influences and intersects with cultural variables in various ways. That is, how students interpret test items do not always match the assumptions and intentions in the items (Cooper & Harries, 2002; Noble et al., 2012). In their study of constructing a legitimate 'realistic' math item, Cooper and Dunne (2000) showed that when presented with a contextualized math task, the students made sense of the tasks through their encountered experiences in everyday life. The sheer appearance of children's names, were not mere tokens to the 11-14 year-olds but something they incorporated into their scheme-making for concepts to be relatable. Similarly, Boaler (1994) found that a plausible reason for girls'

underachievement in math assessment was that girls might find certain types of 'realistic' items more difficult to negotiate than boys did.

Equity is the belief that student achievement increases when culturally diverse students receive opportunities that allow them to draw on their social and cultural literacies in order to be academically successful (Stembridge, 2019). Solano-Flores and Nelson-Barber (2001) showed examples that cultural assets, values, attitudes, and experiences from students' cultures and communities have created in them differed ways of sense-making of items and different problem-solving patterns. For Native Americans and Indigenous children, for instance, their attitudes, identities, knowledge, and behaviors toward science are intertwined with their daily interactions within and across the cultural borders between their community and the school (Honey & Grotzer, 2013; Nelson-Barber & Trumbull, 2007). American Indian/Alaska Native students may prefer assessment formats that require reflection and integrating multiple perspectives rather than those force a single answer, such as multiple-choice and true/false, which according to cultural minority researchers, the latter may respect more of their ways of knowing and thinking (Macias, 1989). In addition, the tension between a student's minority culture and school science in contextualized tasks, could lead to their perception of gaps or confusion about the boundary between their life and science, which can further lead to the underperformance of minority groups in science tests (Medina-Jerez, 2008; Parrott et al., 2000). On the other hand, Nunes et al. (1993) showed the positive role of street culture when Brazilian street children performed better in mathematical problem solving and calculation in a street setting than in a more formal test setting.

Lastly, if we look at cultural validity from the perspective of measurement error, the literature on DIF indicate some sociocultural characteristics of items and test-takers

as the underlying causes. When checking for fairness, researchers often test items for DIF between gender, language, racial/ethnic, and other cultural groups. Some studies (Huang et al., 2014; Yildirim & Berberoğlu, 2009) confirmed that cultural difference is one of the three most common factors to contribute to DIF in large-scale assessments, along with factors of assessment linguistic and curriculum coverage differences. Additionally, similar to examples of stereotype threats of race and cultural identity (Steele & Aronson, 1995), environmental, situational, or social cues embedded in a test item and/or testing situation can be sources of bias, which despite being subtle and implicit, may pose important challenges for valid inferences of test scores (Zumbo, 2007; Millsap, 2012). In their study of DIF evaluation for the PISA 2003 math problems, Yildirim and Berberoğlu (2009) found that, given there was equivalence in translation and adaptation of tests, an item that involved an algebraic manipulation that entailed reproduction skills favored the Turkish sample. Items that required reflection skills such as interpretation of graphs and mathematical communication were found to favor the U.S. students. Rosario Basterra et al. (2011) contend that if the DIF is associated with linguistic and cultural cognition, test developers need to modify or eliminate the items flagged for ethnic-racial DIF.

In conclusion, examining the literature shows that without considering students' cultural backgrounds, our tests will be a very inefficient way to measure, and is a missed opportunity to offer a more accurate picture of the true ability of the test taker. However, evidence for cultural validity are usually collected from qualitative case studies or cognitive interviews. Only in recent years has more research begun systematic item programming and context profiling, such as those for item illustration (Solano-Flores et al., 2016) and item context's cognitive and sociolinguistic features (e.g., Ruiz-Primo & Li, 2015). However, none of the efforts has addressed the cultural

characteristics of science item contexts. In what follows, we present our conceptual framework and a newly developed set of rubrics to examine cultural validity in item contexts.

3.3 Theoretical Framework

We present our theoretical framework for cultural validity from an equity stance in multicultural education. A multicultural perspective in understanding consequential validity is useful from a cultural and critical lens, which analyzes the relationship and interplay between testing, science knowledge, power, and the society. We draw on the critical race theory (Ladson-Billing & Tate, 1995), and feminist knowledge (Harding, 1993; Collins, 2002) to propose an equitable and culturally valid science assessment framework for all students. Critical race theory scholars employ an interpretive lens of multicultural education through the lens of the broader social relationships, and help interrogate the power, privileges, and knowledge deeply embedded in our academic schooling and achievement disparities between students of color and the mainstream students in the U.S. (Banks, 1993). According to those theories, a hegemonic culture of testing, which includes its testing practices and knowledge of tests, needs to be challenged as it favors the preferred dominant cultural group and marginalizes the knowledge and epistemology of those minority groups and communities. Moreover, based on the feminist views of knowledge, we perceive knowledge assessed by an item as part of the social structure and a vehicle for equitable knowledge construction. Therefore, we contend that a culturally valid context must be an equitable context that reflects multicultural ideals and commitments in knowledge construction and context representation.

For standardized tests, an equitable item context requires common patterns of contextualized knowledge that students from different racial, ethnic, linguistic, geographic, and other cultural backgrounds would have equal access to encounter. Viewing student achievement and context understanding as a sociocultural product of various complex factors such as the acquisition of available cultural capital, we want to disentangle and examine the cultural codes and cultural assumptions embedded in testing and test items. Specifically, by placing an item context at the center of analysis, we propose three domains of an equitable item context from multicultural theories (Figure 3-1). Such a proposition aims to gauge a context's functional, cultural, and psychological considerations regarding equity for culturally diverse students.

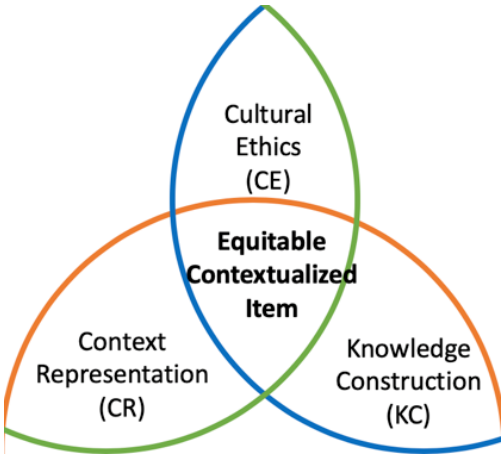


Figure 3-1. Conceptual framework of equitable contextualized item (ECI).

In this way, our orientation considers assessment items as the critical venue for reforming standardized tests and enacting equity for societal change. Based on multicultural principles, we use an analytical framework for the context coding of science items. Table 3-1 lays out specific analytic elements for sources of cultural bias, contextualized representation, and the knowledge and epistemology about coding science contexts.

Table 3-1

Analytical framework for contextualized science items

Dimension	Component	Multicultural Perspectives
	The Microsystem level	
Cultural equity	Cultural identity representation: Race, ethnicity, gender, sexuality, disability, social class, religion, language, immigration, and citizenship status	Diversity requires the understanding and representation of different cultural identities in science, including how minority scholars have influenced the history and development of science.
	The Mesosystem level	
Context representation	Cultural assets: culturally representative materials, tools, artifacts Student experiences: behavior, roles, cultural practices, tradition, community practice, rituals, and custom Social capital: networks, social relations; Cultural capital: skills, education, digital literacy Sociocultural locations: school, family, community, neighborhood, church, tribe	Contextual dynamics reflect the social reality and respect the multicultural environment of science. Context is inclusive of science practices, experiences, and locations of communities of color. Science facts and artifacts are equally accessible for cultural and linguistic minority groups.
	The Macro system level	
Knowledge construction and epistemology	Chronological, historical, and political context: political and social development, movement, protest, and public policies, regulations, and laws Social values domain: beliefs, attitudes, tenets, agenda, ideologies, cultural norms, religious assumptions Language and agency: cultural codes, symbolism, post-colonial power, postmodern politics of difference Knowledge construction: cultural orientation, epistemological orientations	Critical thoughts require deconstructing structural conditions and discourses embedded in scientific knowledge, especially the interplay of complex factors and power relations. Political orientation over cultural identity representations embodies individual agency, advocacy, democracy, and emancipatory agenda.

3.4 Cultural Validity of Item Context: Rationales and Key Concepts

3.4.1 *Cultural Equity*

Equitable considerations of *item context* highlight multicultural principles. Cultural

Equity (CE) includes an examination of cultural, psychological, and implicit bias, as well as norms and agency regarding human subjects represented in an item context. Cultural biases include both explicit and implicit stereotypes, bias, and beliefs in certain prescribed viewpoints that readers can evidentially detect from discourse, narratives, and expression. Cultural bias in testing setting can inadvertently increase one's awareness about themselves with potential negative thoughts or stress around the stereotype, which may suppress emotions and thoughts, reduce working memory, and thus reduce the test performance in the influenced group (Steele et al., 1995; Bair & Steele, 2010). We have flagged five unfavorable areas to consider cultural equity as below.

***Explicit stereotypes*:** the extent to which item context has explicit negative stereotypes, prejudice, or xenophobia against subordinate racial, ethnic, gender, language, disability, and social class groups. Inequitable item contexts have words in this manner exhibiting stereotypical judgments or overgeneralization against a certain group's status, knowledge, or practices.

Privileges: the extent to which item context illustrates privileges of dominant groups that stem from status differentiation, education, social positions, power, masculine pride, racial hierarchy, and linguistic or physical superiority. For example, inequitable contexts highlight the privileges of a dominant group of scientists by a higher level of their resources, achievement, independence, intelligence, and leadership that are inaccessible to other groups.

Agency: the extent to which item context gives personal or collective agency through describing specific intentions, events, or actions about a person's character or a group. Comparatively speaking, the agency assigned to a mainstream cultural identity while not on other subjects present in a context has symbolic inequality and may be

inequitable. Item reviewers should tag language, either text or visual image, that assigns mainstream or dominant person identity with good reasoning, logic, awareness, proactivity, judgment, and capability.

Norm perpetuation: the extent to which item context perpetuates a status quo, social inequality, or power structure in the society. Perpetuating inequitable norms or pathologizing certain groups continue to place the minority, the subordinated, and the victimized as inferior. Displaying cultural identity maintaining its role, duties, and practices just as it is; or normalizing minority groups from a deficit-based point of view, rather than strength-based views, is inequitable for culture.

Implicit bias: the extent to which item context contains insidious bias information that, with or without the perceiver's awareness, may activate negative messages, feelings, or collective memory of the test taker. Psychologically, according to Greenwald and his colleagues (1995), implicit bias is subject to automatically triggered evaluation. It can be so cognitively or psychologically hidden that victims would encounter stereotype threats or a self-fulfilling prophecy that negative affects on test takers (Steele, 1997).

The above five categories are examinable when an item context is considered a sociocultural context or contains text or image stimulus of human beings in the context. Issues of cultural equity illustrate equitable considerations of science contextualization and call to evaluating cultural biases that might negatively impact a student's performance in an assessment.

3.4.2 *Equitable Context Representation*

Not only should item context fulfill the purpose of contextualization, but an equitable and inclusive context creates, represents, and sustains meanings from student' identities, with student's communities, and for student's learning. Orner (1996) has argued the

importance of contextualization because knowledge and capability develop in the situatedness of learning, which does not occur in a vacuum. Multicultural scholars indicate that academic content should be relevant to students' cultural experience to make such content accessible, meaningful, and relatable (Gay, 2010; Ladson-Billings, 1995), and the need for connecting, validating and appreciating differences in students' backgrounds and drawing on students' funds of knowledge (González et al., 2006). Thus, we conceptualize the five functions of a culturally valid and equitable context in science assessment on its familiarity, constructivity, applicability, accessibility, and coherence characteristics.

Familiarity: the extent to which the purpose of science in a context is familiar, relatable, and easy to understand for all students. A familiar and relatable science context connects the scenario with what builds on students' prior knowledge, be it from students' life or closely aligned with the school curriculum. In principle, we take into consideration if the objects, artifacts, and materials used in the context are familiar and relevant to the local demographics of students who participate in the test. More broadly, if students from different groups have various cultural backgrounds, an equitable context from the multicultural perspective should pay special attention to cultural minority students, groups from low-income backgrounds, and their cultural knowledge and practices.

Applicability: the extent to which consequences or concepts of science scenarios be applied to students' and their communities' everyday life, hobbies, interests, and benefits. An applicable context should be authentic, practical, hinged on students' interests, and common to encounter in the natural or social world. An equitable and applicable context helps motivate all students as they apply science context knowledge and skills in their daily life, for themselves, and for their community development.

Constructivity: the extent to which elements of item context are constructive, helpful, and structured in scaffolding scientific thinking and skills. Constructive context has additional information, term definitions, explanations, and other resources to promote understanding, and bridges prior knowledge with constructive schemes. Drawing on students' funds of knowledge, an equitable constructive context taps on multicultural students' daily learning by doing, "hands-on" activities, experiments, community practices, routines, and personal, social, and cultural experiences. Additionally, constructive context can assess procedural science knowledge, incorporating helpful modules, building blocks, tools at home or school, and creative art projects, in a way to be more applicable and relatable to all students.

Cohesiveness: the extent to which the details of item context capture a coherent story or message. For example, for a scientific phenomenon or event, the five elements of "what", "who", "where", "why", and "how" are clearly specified and essentially described in item context. For a scientific experiment, the setting, action, intention, causes, and effect are clear and organized to render a coherent story.

Accessibility/linguistic complexity: the extent to which language and design of a context are, in principle, accessible for all students, including linguistic minority students and students with special needs. For instance, students who need accommodations have different modalities for assessment items and context available, such as in audio, visual, and human-assisted resources. For this study, however, an equitable context focuses on a reasonable reading load and cognitive demand for all students.

3.4.3 *Knowledge Construction*

If we are to construct a culturally valid and equitable science context, we need to tackle the metanarratives, epistemology, and knowledge construction of science knowledge

represented in *item context* as well. Based on the culturally responsive theory, we have five categories of knowledge.

Transmitting of knowledge: the extent to which it reflects the knowledge producer as the group in power with an agenda. It may reflect problematic or predominantly calculated white/Anglo-Saxon, androcentric/patriarchal, middle-class, or religious groups of values, worldviews, codes, and norms. *Linear and field-independent knowledge structure***:** the extent to which knowledge is represented in abstract, schematic, theory-based, oversimplified, and monotonous ways. *Complex and field-dependent knowledge structure***:** the extent to which knowledge is represented in resourceful, multi-modal, interactive, multidimensional, and dynamic ways. *Transformative and multicultural epistemology*: the extent to which context knowledge counters hegemonic ways of Western-oriented knowledge, or is critical, culturally sustaining, reflecting diverse ways of knowing and thinking, or knowledge for the greater social good. *Hegemonic and oppressive epistemology*: the extent to which knowledge reflects suppressive, partisan, exclusive, hegemonic, and materialistic ideologies. Experts and researchers who have solid knowledge in multicultural theories and epistemology, in considering whether students from different racial, ethnic, and cultural backgrounds would have equal access to encounter and excel, should evaluate the knowledge construction of item context.

3.5 Method

3.5.1 *Analytic Sample*

We have collected 145 publically released PISA science assessment items, including the Field Trial items and student responses for the U.S. datasets with sample sizes varied from 3846 to 5712 students for each corresponding year in 2000, 2006, and 2015. Since

PISA assessment items are presented in a testlet structure where multiple items share a common scenario or stimulus, we define the top level of context structure where all items nested within have one common scenario as the "general context". A secondary level of context shared by a sub-cluster of items is defined as "sub-testlet context", and a bottom level as "item context" if applicable. Specifically, Table 3-2 shows the breakdown of context levels for the PISA public item pool. Excluding those without cognitive statistics available, we select 44 general contexts, 9 sub-testlets, and 48 item contexts for our item context analysis.

Table 3-2

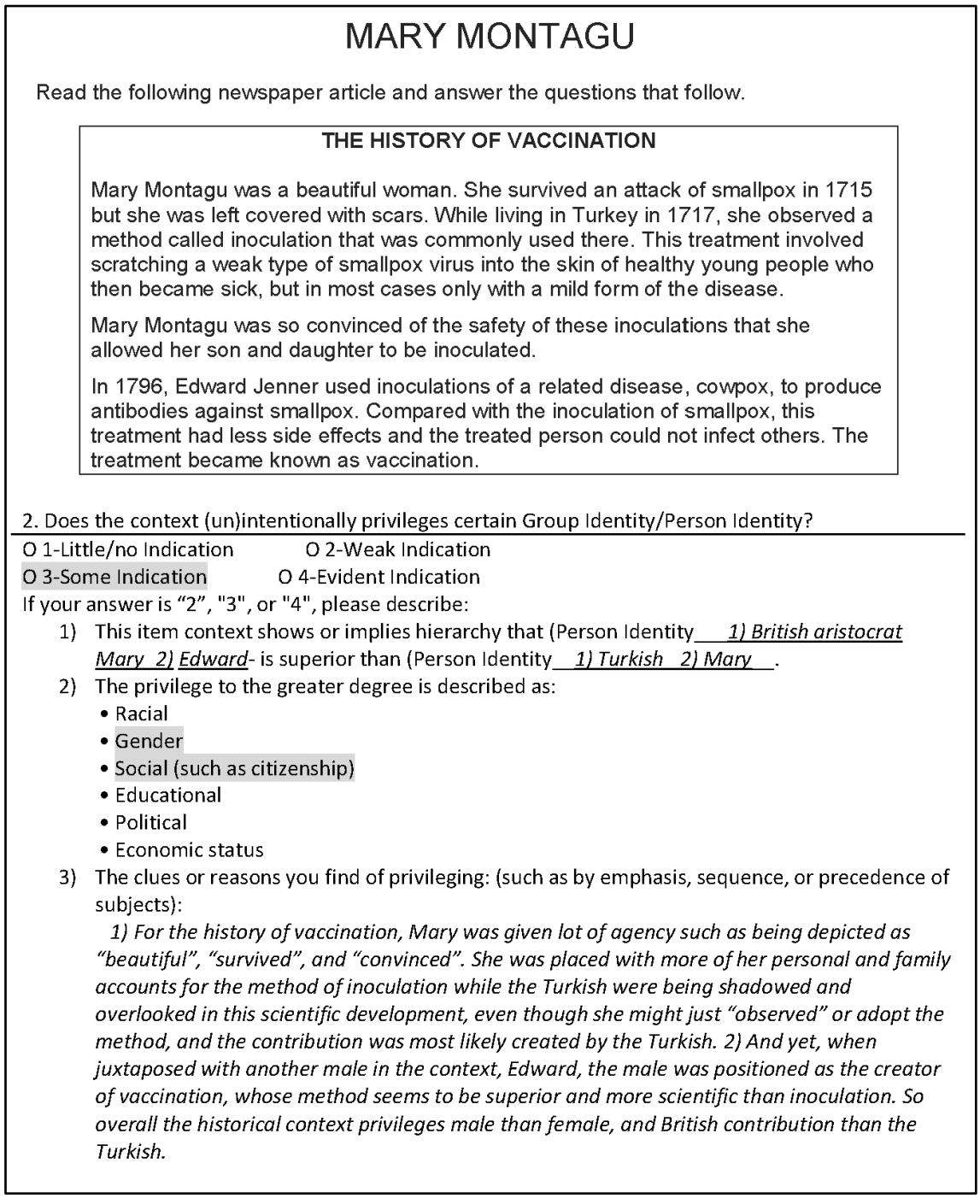
Breakdown of context levels for the publicly released PISA science item

Context level	**N**
General context only	55
General and sub-testlet context only	19
General and item context only	58
General, sub-testlet, and item contexts	6
Item context only	4
Total	142

3.5.2 *Developing Coding Rubrics*

Context Profiling Approach. Our rubrics of context profiling, the Equitably Contextualized Items (ECI) coding rubrics, elicits quantitative and qualitative codes for each context unit from human experts. The rubrics for fine-grained cultural characteristics of context categorization and discourse analysis are developed alongside the conceptual framework, a codebook, and procedural rules to ensure reliable and valid, descriptive and evaluative, and comprehensive hand-coding on multiple levels of

contexts. Figure 3-2 is an example of context profiling for a general context level on the "privilege" category for the domain of cultural equity.



MARY MONTAGU

Read the following newspaper article and answer the questions that follow.

THE HISTORY OF VACCINATION

Mary Montagu was a beautiful woman. She survived an attack of smallpox in 1715 but she was left covered with scars. While living in Turkey in 1717, she observed a method called inoculation that was commonly used there. This treatment involved scratching a weak type of smallpox virus into the skin of healthy young people who then became sick, but in most cases only with a mild form of the disease.

Mary Montagu was so convinced of the safety of these inoculations that she allowed her son and daughter to be inoculated.

In 1796, Edward Jenner used inoculations of a related disease, cowpox, to produce antibodies against smallpox. Compared with the inoculation of smallpox, this treatment had less side effects and the treated person could not infect others. The treatment became known as vaccination.

2. Does the context (un)intentionally privileges certain Group Identity/Person Identity?

O 1-Little/no Indication O 2-Weak Indication
O 3-Some Indication O 4-Evident Indication

If your answer is "2", "3", or "4", please describe:

1) This item context shows or implies hierarchy that (Person Identity *1) British aristocrat Mary 2) Edward*- is superior than (Person Identity *1) Turkish 2) Mary*.
2) The privilege to the greater degree is described as:
   - Racial
   - Gender
   - Social (such as citizenship)
   - Educational
   - Political
   - Economic status
3) The clues or reasons you find of privileging: (such as by emphasis, sequence, or precedence of subjects):
   1) For the history of vaccination, Mary was given lot of agency such as being depicted as "beautiful", "survived", and "convinced". She was placed with more of her personal and family accounts for the method of inoculation while the Turkish were being shadowed and overlooked in this scientific development, even though she might just "observed" or adopt the method, and the contribution was most likely created by the Turkish. 2) And yet, when juxtaposed with another male in the context, Edward, the male was positioned as the creator of vaccination, whose method seems to be superior and more scientific than inoculation. So overall the historical context privileges male than female, and British contribution than the Turkish.

Figure 3-2. An example of general context coding on cultural equity— "privilege".

Instrument. The ECI instrument is on three domains of cultural equity, context representation, and knowledge construction. We sought out a panel of external experts in multicultural education and cultural studies for feedback and validation. The

instrument has 15 Likert-scale questions (1-Little Indication to 4- Evident Indication) and 20 more context classification questions under each domain. In addition, the instrument includes open-ended questions for the coder's reasoning and evidence. As a result, the generated codes represent a total of more than 150 possible variables and evaluative comments, indicating a rich and wide data structure.

Psychometrically, our instrument satisfies an acceptable reliability ($\alpha = 0.71$) of standardized Cronbach's alpha from the classical test theory (Cronbach, 1988). A confirmatory factor analysis from the pilot data has confirmed a possible factor structure for the ECI instrument. Specifically, a 3-factor structure supports the model fits of $\chi 2$ (87) = 237.9, $p < .05$, CFI = 0.86, TLI= 0.83, RMSEA = .16, and SRMR = .28; a 1-factor model with fit indices of $\chi 2$ (90) = 112.1, $p > .05$, CFI = 0.83, TLI= 0.95, RMSEA = .08, and SRMR = .26; whereas a 2-factor structure has fits of $\chi 2$ (89) = 153.2, $p < .05$, CFI = 0.94, TLI= 0.93, RMSEA = .12, and SRMR = .21. For the rest of the following sections, we chose the 2-factor structure to explore possible associations of ECI characteristic variables and cultural validity.

Coder Training. Coder training for a team of multicultural education and science education researchers on the multicultural knowledge of science education is conducted. The process spans over a year for coder recruitment, knowledge training, and consensus calibration. With the help of a well-defined codebook and other reference materials, coders were asked to code independently in a coherent and accurate manner. After training, the collaborative coding was conducted among a group of three coders to monitor consensus codes with adjudication of discrepant codes.

***Expert Coding*.** The consensus codes are used for data analysis, although we select independent codes for checking interrater agreement. For each two expert coder/rater pairs' agreement, the quadratic weighted Kappa (Cohen's Kappa) is around

0.63 ($z = 6.55$, $p < 0.001$). According to Landis and Koch (1977)'s commonly cited guidelines, the value from 0.61 to 0.80 suggests substantial agreement. With this indicator, the study went on to collect consensus codes, where each of the three coders took up an equal amount of random context units for profiling coding, cross-validated by a second reviewer to flag potential discrepancies, and a final check through team discussion for consensus.

3.6 Data Analysis

3.6.1 *K-means Clustering*

K-means clustering (Hartigan & Wong, 1979; Berkhin, 2006) with principal component analysis (Jolliffe, 2011) is a widely used statistical technique for unsupervised learning tasks. We used K-means clustering to characterize PISA context prototypes under the 2 component factors of ECI items after the CFA analysis. The clustering algorithm uses the 15 ECI feature attributes (See the rubrics of each feature in Appendix 1.A) to perform prototypes classification and represent clusters by optimizing distance and similarity functions (in K-means, it is the squared error function minimizing the Euclidean distance to cluster centroids). In the first phase, k initial centroids are selected randomly. We used the PCA factor scores as a variable in the regression method and iteratively updated the centroids until convergence. As the K increases, there are various enhanced extensions to k-means for regularization over k, e.g., Akaike (AIC) or Bayesian information criteria (BIC). The number of clusters, K, is selected through the gap statistic (Tibshirani et al., 2001) comparison and the purpose of being relatively easier to classify and interpret the small-sample item-level data through a fixed *a priori* from the conceptual perspective. Cluster membership assignment is followed by profile analysis for exploring equitable and inequitable prototypes.

3.6.2 *Differential Item Functioning*

In the PISA's technical reports, the DIF analysis was performed in the multi-facet model of ConQuest (Wu, Adams, & Wilson, 1998), and the IRT models are varied from Rasch to two-parameter logistic for measurement models (Adams & Wu, 2003; OECD, 2009; 2017). The Item by Group interaction term is then added to the standard model. We use the unidimensional version of the multidimensional random coefficient multinomial logit model (MRCLM; Adams, Wilson, & Wang, 1997) to specify the Rasch partial credit models, with γ_i representing the DIF item parameter vector accompanied by G groups, indicating item difficulty differences between the focal and reference group. We performed three categories of multiple-group DIF: gender (female, male-reference group), race (White-reference group, Asian, Black, Hispanic, other), and language (English, Spanish-reference group, other) using PISA's variable "Language At Home". All items were formulated through the partial credit model ("item + item*step"), and estimated using the Marginal Maximum Likelihood algorithm in the TAM package in R (Robitzsch et al. 2020). Note the means of the item difficulties were constrained to 0, and the population distribution takes account of the impact in group ability differences through latent regression of θ onto the group variable. The standard settings of the TAM function tam.mml.mfr were used for the analysis of DIF, and we adopt Paek and Wilson's (2011, p. 1028) threshold values of a Rasch DIF framework based on the ETS classification standard, where A is considered a negligible DIF, B a medium DIF, and C a large DIF:

- A if $|\gamma| \leq 0.426$ or if H_0: $= 0$ is not rejected below 0.05 level
- B if $0.426 < |\gamma| < 0.638$ and if H_0: $= 0$ is rejected below 0.05 level
- C if $0.638 \leq |\gamma|$ and if H_0: $= 0$ is rejected below 0.05 level

3.6.3 *Case Discourse Analysis.*

In a mix-method manner, we followed those items with contexts identified as problematic or inequitable by the above clustering and DIF results. The qualitative discourse analysis enables us to look deeper into the concept of cultural validity and its association with item characteristics. Knowledge of both measurement effects and case studies contributes to the interpretation of overall cultural validity and internal construct validity.

3.7 Results

3.7.1 *Sociocultural Representation in PISA Science General Contexts*

We first report descriptive statistics in relation to what kind of human actors, social contexts, and environments are represented in our sample. Part of the ECI rubrics requires coder's judgment of subject's sociocultural identity associated with their names and pronouns. The result would represent a 'master narrative' of science PISA item contexts to some extent. Overall, the representation of human identities in those contexts is quite skewed. Table 3-3 provides the information on a White-centered, male-dominant sociocultural representation in the science contexts. The PISA general contexts are mostly social contexts. However, almost half (47.8%) of the social contexts are surrounded by White identities or Western contexts. A total of 8 PISA general contexts explicitly identify with a country or nationality name, and all of them are situated in the Western countries, including France, Canada, Australia, Athens, or "the British", and "the Dutchman".

Table 3-3

Context Categories and Cultural Identity Represented in General Contexts

Variable	Category	N of Contexts	Percentage
Context Category†			
Cultural context	Western	12	27.3
	Non-Western	2	4.5
Social context	Social	23	52.3
	Non-social	15	34.1
	Social (Minimal) ‡	6	13.6
Context proximity	Everyday life scenarios	11	25.0
	Practical science/phenomenon	33	75.0
Subject Identification†			
Race/ethnicity	White/Caucasian	11	47.8
	Asian	1	4.3
	Black	§	§
	Hispanic	§	§
	Other	§	§
Gender	Male	9	39.1
	Female	3	13.0
	Male & Female	2	8.7
Social class	Middle Class	7	30.4
	Non-Middle-Class	2	8.7
	Non-identifiable	14	60.9
Person/Group	Individual	13	44.8
	Group	14	48.3
	Group & individual	2	6.9

† N = 45 general contexts for context category and N = 23 general contexts for subject identification; the category of "Non-identifiable" is left out.
‡ With almost negligible social information.
§ Data shows zero counts.

We also observed the intersectionality of represented cultural identities. 8 general contexts (34.8%) among all social contexts have the White, male, or middle-class identities intersected with one another. 5 general contexts (21.7%) have the White female characters, versus only one general context is indicative of minority, and male character (4.3%). With respect to the gender representation alone, males are more than two times as likely to be present in the sample contexts, and more likely to be associated with the language of "privilege" or "agency" than the females, of which, 60% are also intersected with a "White" identity. Although a small sample, this revealed a pattern

from the publicly released PISA item pool that speaks to the skewed representation of PISA's science contexts.

Indeed, the absence of other cultural groups or people of color in the representation of science contexts is concerning. Only one general context is associated with an Asian name for cultural identity representation, and no information is associated with any African or Hispanic persons (names). The qualitative judgment of expert coders suggests an over-sampling problem, for PISA as an international assessment program, of the Western, Educated, Industrialized, Rich, and Democratic (WEIRD) society people, which are only 12% percent of the world population.

3.7.2 *ECI Functions: Cultural Characteristics by Clusters*

The descriptive statistics of ECI Likert-scale scores are provided in Table 3-4. Assuming equal distances between ordinal variables here, the mean and standard deviation give a sense of the overall behavior of contexts in the three domains. Domain I have higher 'means' because non-social contexts were given a perfect score in the item sample. We have further grouped the 15 Likert-scale items into the equitable and inequitable categories for an additional binary scale (LI, MI = 0, SI, STI = 1).

Table 3-4

Descriptive Statistics of ECI Measurement about the PISA Contexts

			ECI Context Code Frequency (%)			
Context Function	**Mean**	**SD**	**LI**	**MI**	**SI**	**STI**
Domain I: Cultural Equity*						
Explicit Stereotypes	3.65	.76	0 (0)	13 (16.88)	1 (1.3)	63 (81.82)
Privileging Group Identity	3.61	0.85	4 (5.19)	6 (7.79)	6 (7.79)	61 (79.22)
Agency Over Others	3.38	0.90	3 (3.9)	13 (16.88)	13 (16.88)	48 (62.34)
Perpetuation of Representation	3.27	1.01	6 (7.79)	13 (16.88)	12 (15.58)	46 (59.74)
Implicit Bias	3.23	0.96	4 (5.19)	16 (20.78)	15 (19.48)	42 (54.55)
Domain II: Context Features						
Familiarity	3.14	0.68	0 (0)	13 (16.88)	40 (51.95)	24 (31.17)
Applicability	2.82	0.87	5 (6.49)	22 (28.57)	32 (41.56)	18 (23.38)
Constructivity	2.48	0.77	8 (10.39)	29 (37.66)	35 (45.45)	5 (6.49)
Cohesiveness	2.73	0.84	5 (6.49)	25 (32.47)	33 (42.86)	14 (18.18)
Accessibility	2.71	0.86	3 (3.9)	33 (42.86)	24 (31.17)	17 (22.08)
Domain III: Knowledge Construction						
Knowledge Subjectivity	3.51	0.82	2 (2.6)	10 (12.99)	12 (15.58)	53 (68.83)
Linear Knowledge Structure	2.90	0.88	3 (3.9)	25 (32.47)	26 (33.77)	23 (29.87)
Dynamic Knowledge Structure	2.81	0.86	7 (9.09)	16 (20.78)	39 (50.65)	15 (19.48)
Equitable Epistemology	3.32	0.64	0 (0)	7 (9.09)	38 (49.35)	32 (41.56)
Inequitable Epistemology*	3.03	1.01	8 (10.39)	14 (18.18)	23 (29.87)	32 (41.56)

Note: N = 101 contexts across three levels. LI= Little Indication, MI= Mild Indication, SI = Some Indication, EI = Evident Indication; frequencies and percentages in parentheses are reported.
* The categories are reversely coded to show LI to EI as from inequitable to equitable and culturally valid.

We applied both k-means and hierarchical clustering on general- and item-level data, and the gap statistic in Figure 3-3B and the silhouette statistic estimated the number of clusters to be 2 as the elbow in the monotonically rising statistic. However, since we are interested as well in identifying subgroups under the 2-dimensional data, we chose k = 4 for exploring the detailed prototypes under equitable and inequitable dimensions.

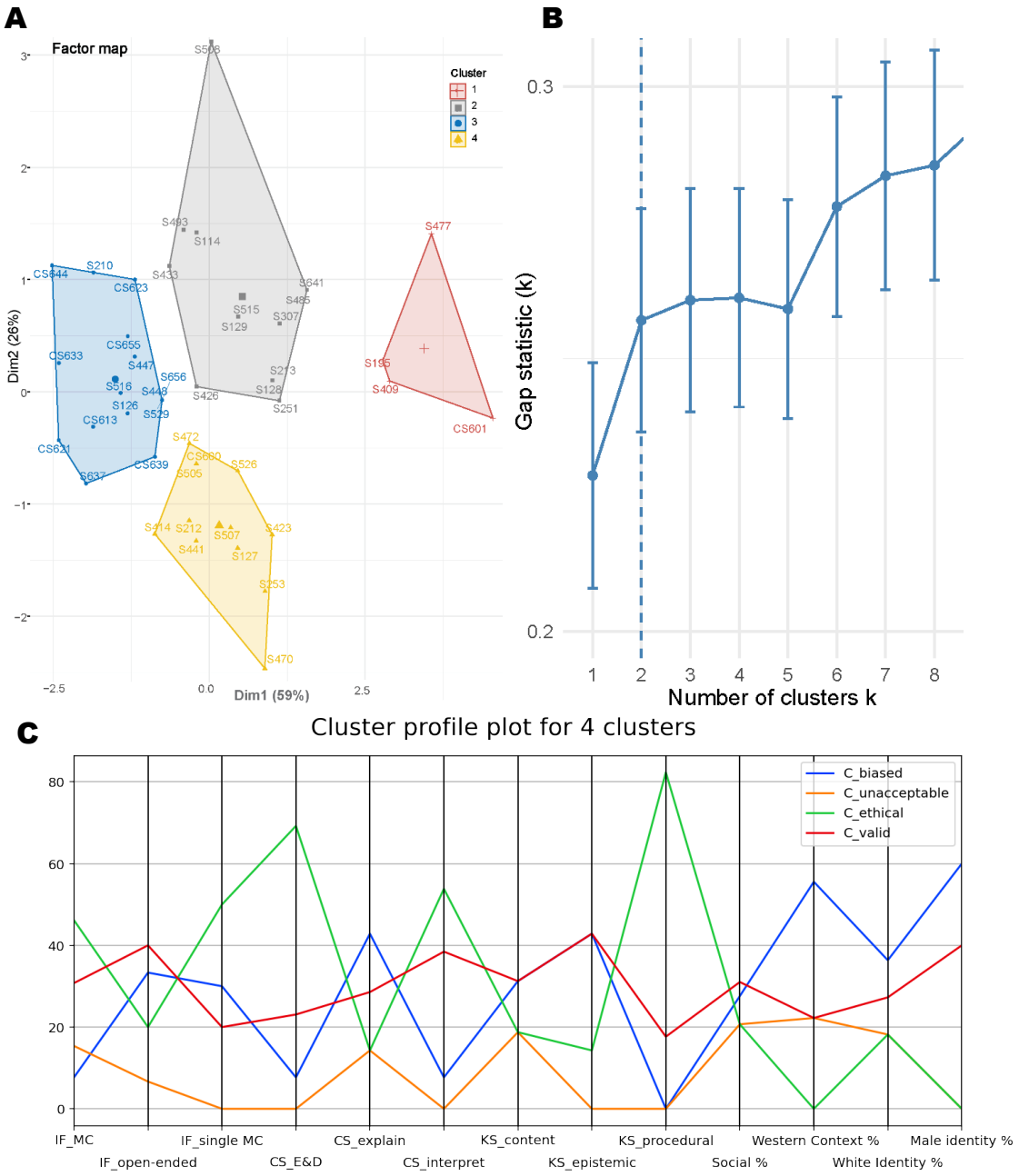


Figure 3-3. K-means clustering and cluster profiles.

Note. **3-A**: visualizations of 4 clusters of 44 general contexts on the coordinates of dimension 1 and 2. For better interpretation, we relabeled cluster numbers 1-4 for the rest of analysis as to represent culturally unacceptable, biased, ethical, and valid. **3-B**: Gap statistic for numbers of k clusters for general contexts. Dashed line gives the optimal estimated k. **3-C**: visualization of cluster profiles with PISA context features averaged in percentages by category. Colored lines represent clusters 1-4. IF_MC: Item format- Multiple choices questions; IF_Open-ended: Open-ended questions; from PISA science framework, IF_Single MC: Single multiple choice; CS_E&D: Evaluate and design scientific enquiry; CS_explain: Explain phenomena; CS_interpret: interpret data and evidence scientifically; KS_content: Content knowledge; KS_procedural: Procedural knowledge; KS_epistemic: Epistemic knowledge.

By characterizing each group of general contexts as single data points, we can get a somewhat clearer picture. In Figure 3-3A, the four clusters are mapped onto the two principal components (explaining 85% of the variance), and are separated by each dimensional coordinate from zero. Cluster centroids are means of ECI scores. Based on the clustering, we see that Cluster 1 has a significant higher mean in the lower indication categories than the rest (Little Indication: $Mean = 4$, $SD = 1.19$, $p < 0.05$; Mild Indication: $Mean = 7$, $SD = 1.99$, $p < 0.05$). Cluster 2 has the highest percentages over the representation of "explicit Western context" and "White identity". Cluster 3 has significantly higher scores than Cluster 2 in most of the categories of cultural equity, according to the pairwise Turkey HSD tests. Cluster 4 has the highest means over most of the ECI categories (11 out of 15 categories). Based on the alignment of dimensionality and the aforementioned cluster characteristics, the clusters can be interpreted as the culturally unacceptable and biased (clusters 1 and 2), and culturally ethical/equitable and valid (clusters 3 and 4). With this satisfactory solution, we then applied the clustering algorithm to the item-level data, and further characterized the four prototypes with categories of the PISA science framework, item format, and sociocultural context in percentages accordingly (Figure 3-3C). For example, we see that items focused on "explaining phenomena scientifically" tend to have a more biased context than those "evaluate and design scientific enquiry" and "interpret data and evidence".

3.7.3 *Cultural Validity, Item Difficulty, and DIF*

The interpretation of clustering on the 2-dimensional space based on the ECI scores facilitates cultural validity arguments for items, including item's cognitive difficulty, DIF, knowledge system, and format profiles. Based on the 48 publicly available items, we found that many ECI scores are positively associated with better student

performance. Specifically, linguistic accessibility/complexity, context cohesiveness, implicit bias, agency, and applicability are the top five sociocultural attributes positively associated with student performance at item level. Kendall's tau coefficients of rank correlation with items' p-values ranging from 0.09 to 0.36. Among those contributing characteristics, linguistic accessibility or linguistic complexity has shown to be a significant factor associated with the item parameter ($z = 3.20$, $p < 0.001$). Similarly, for the estimated item difficulty values from the PISA items' cognitive metadata, the top five attributes of linguistic accessibility/complexity, context cohesiveness, applicability, implicit bias, and context's constructivity are positively associated with easier items in the sample.

One of the important goals for cultural validity on contextualized items is the evaluation of DIF as indicated by differential predicted probabilities in getting answers correct by groups at the same level of latent ability. We re-calibrated PISA 2000, 2006, and 2015 students' responses from the Rasch/partial credit unidimensional models before estimating the Rasch DIF from the TAM package. The estimated item difficulties were mostly comparable to the PISA's official item statistics (Figure 3-4A), with models rendering EAP-reliability of around 0.90 for each calibration. Among the 48 PISA publicly released science items, we plotted the expected score curves to confirm the uniform DIF. Across the sample, the gender DIF is the least uncommon, with items mostly shown in the culturally unacceptable cluster in the more difficult item range (Figure 3-4B). Similarly, the culturally unacceptable cluster tends to show more language DIF items, items being inequitably easier for English speaking group (Spanish is the reference group), in the more difficult item range.

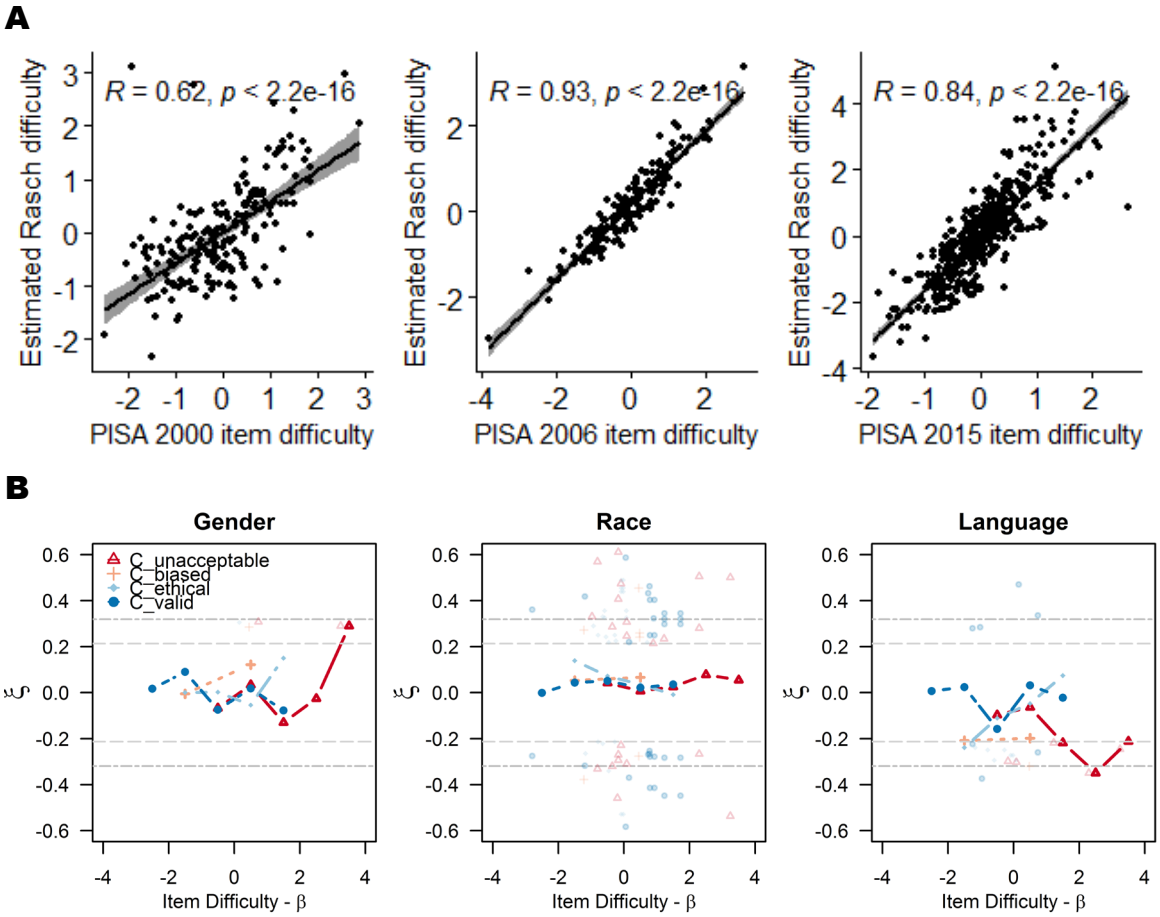


Figure 3-4. Estimated item difficulty and DIF parameter ξ.

Note. **4-A**: Plot of estimated Rasch parameter against the reported item difficulty in PISA technical reports in years 2000, 2006, and 2015. Pearson product-moment correlation coefficients (*R*) were computed. **4-B**: point connected lines are DIF parameter ξ ($\frac{1}{2}$ item difficulty difference from focal group to reference group) for demographic groups averaged over (0, 1] ranged bins of item difficulty, by 4 cluster groups in the 48 public items. Grey dashed lines represent $\frac{1}{2}$ of medium (0.426) and large (0.6) DIF thresholds. Transparent points exceeding those two thresholds are plotted.

Since our work of item profiling anchors on equity and fairness, we ultimately attempt to investigate item characteristics associated with item difficulty and DIF parameters due to gender, race, and other cultural variables. Although a study of a small sample (n = 48), there seems to be a significant relationship between the holistic grouping on cultural validity (the equitable and inequitable ECI binary) and DIF status, $p = 0.039$, 95% CI [0.002, 1.064], from the Fisher's exact test. The inequitable group significantly predicts the DIF items in binary logistic regression with a single

categorical predictor (β = 2. 639, *std* = 1.035, z = 2.550, $p < 0.05$). This means a holistic evaluation by coders shows 93.3% of the true positive rate of the ECI inequitable items truly are diagnosed as DIF items, whereas 32.3% of the false positive rate indicates the percentage of ECI equitable items that were in fact DIF items. Not surprisingly, the top ten testlets with the most frequent harder DIF for minority groups are correctly classified into the culturally unacceptable and biased cluster groups (Table 3-5).

Table 3-5

Top Ten Testlets with Number of DIF Items Aggregated

ECI Cluster Membership	General Context	Moderate and Large DIF			% Total DIF
		Gender	Language	Race	
Cluster 2 (Biased)	S114	0	1	7	38%
Cluster 2 (Biased)	S129	1	1	2	29%
Cluster 2 (Biased)	S213	0	1	3	29%
Cluster 2 (Biased)	S493	1	1	3	24%
Cluster 1 (unacceptable)	CS601	1	1	3	24%
Cluster 1 (unacceptable)	S195	0	2	4	21%
Cluster 1 (unacceptable)	S477	0	1	3	19%
Cluster 2 (Biased)	CS641	1	1	2	14%
Cluster 2 (Biased)	S485	0	0	3	14%
Cluster 2 (Biased)	S128	0	0	2	14%

Note. Only items with harder DIF for the focal groups are counted. Reference groups are gender (male), language (English), and race (White).

The two clusters that have equitable and culturally valid contexts, have higher average p-values (item's percentage correct) for the U.S. sample, though they are not statistically significant. The most important item features are entered into regression in predicting dependent variables of p-value and DIF status (Table 3-6). The variables, including inequitable items by ECI, linguistic accessibility, the open-ended and single MC item format, and whether an item has a social context, jointly explains the variance in a student's correct performance of p-values, with the adjusted R-squared around 0.3. They are also significant predictors of DIF items.

Table 3-6

Predicting Effects from OLS and Logistic Regression with Confidence Intervals

	Dependent variable	
	P-value (OLS)	**DIF prediction (logistic)**
Inequitable items by ECI	-7.109 (-15.999, 1.782)	1.967* (0.124, 3.811)
Linguistic accessibility	7.587** (2.356, 12.817)	
Open-ended format	-8.646 (-19.989, 2.697)	1.872* (0.083, 3.661)
Single MC format	5.237 (-5.219, 15.693)	1.987* (0.280, 3.694)
Have social context	-2.962 (-13.597, 7.673)	-1.980* (-3.907, -0.054)
Constant	36.845*** (18.977, 54.714)	0.565 (-0.738, 1.868)
Observations	48	48
R2	0.296	
Adjusted R2	0.212	
Akaike Inf. Crit.		54.541

Note: N = 48. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

In the next section, we will draw a closer look at the three PISA general contexts in the most concerning cluster: the culturally unacceptable. We will illustrate in-depth how qualitative analyses from the perspective of cultural validity is revealing.

3.7.4 *Discourse Analysis and Knowledge Construction*

Those inequitable items marked from the above procedures are then routed for a more in-depth round of item review. Here we will draw a closer look at the three PISA general contexts in the most concerning cluster-- the culturally unacceptable. Our qualitative codes from the coders echo a coherent story. For instance, an analysis of the "S477_Mary Montagu" context points to potential cultural biases embedded in science education and assessments (Figure 3-2). It shows that the historical narrative can be problematic if a text implies the cultural subtleties of power and the discourse embodies a Western-centric notion of the science history. Specifically, we will show two cases -- one where the cultural unbalance of the context is intersected with cultural assumptions, and another from the perspective of a critical race theory-- how inequitable contextualization can be interpreted by marginalized or minority groups for science

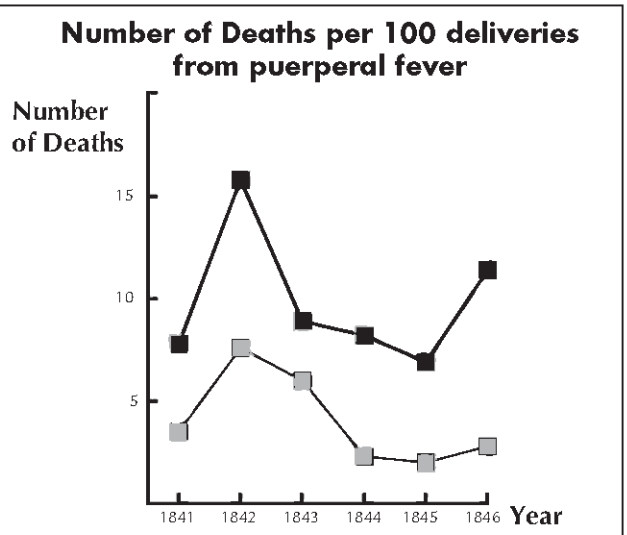
topics.

Case 1: S195_Semmelweis' Diary

Semmelweis' Diary Text 1

'July 1846. Next week I will take up a position as "Herr Doktor" at the First Ward of the maternity clinic of the Vienna General Hospital. I was frightened when I heard about the percentage of patients who die in this clinic. This month not less than 36 of the 208 mothers died there, all from puerperal fever. Giving birth to a child is as dangerous as first-degree pneumonia.'

These lines from the diary of Ignaz Semmelweis (1818-1865) illustrate the devastating effects of puerperal fever, a contagious disease that killed many women after childbirth. Semmelweis collected data about the number of deaths from puerperal fever in both the First and the Second Wards (see diagram).



Diagram

Physicians, among them Semmelweis, were completely in the dark about the cause of puerperal fever. Semmelweis' diary again:

'December 1846. Why do so many women die from this fever after giving birth without any problems? For centuries science has told us that it is an invisible epidemic that kills mothers. Causes may be changes in the air or some extraterrestrial influence or a movement of the earth itself, an earthquake.'

Nowadays not many people would consider extraterrestrial influence or an earthquake as possible causes of fever. We now know it has to do with hygienic conditions. But in the time Semmelweis lived, many people, even scientists, did! However, Semmelweis knew that it was unlikely that fever could be caused by extraterrestrial influence or an earthquake. He pointed at the data he collected (see diagram) and used this to try to persuade his colleagues.

QUESTION 1.1

Suppose you were Semmelweis. Give a reason (based on the data Semmelweis collected) why puerperal fever is unlikely to be caused by earthquakes.

.....

.....

Figure 3-5. PISA 2006 Science general context "Semmelweis' Diary".

This context of *Semmelweis' Diary* has the lowest rating on the "Equity" dimension, such as low applicability and bias prompting. Our qualitative analysis indicates why it could be problematic. For example, for one coder, it is noted:

> Firstly, for the ELL students and those from foreign countries, the word "ward" is critical in understanding the data to answer Question 2 correctly. Without much knowledge or experience of hospital structure, it would be difficult for a 15 year-old middle-schooler to understand what the data is trying to convey. Second, the extraterrestrial influence or a movement of the earth would sound like a reasonable scientific explanation from the Western perspective with its interest in astronomy. The setting and the Doctor's assessment about the epidemic's causes tend to reflect Western science's thinking and development.

It is not uncommon to expect a 'Western' explanation of epidemic around what *Semmelweis* jotted down as "changes in the air or some extraterrestrial influence or a movement of earth itself". However, possible explanations of other cultures may lay more emphasis on possible causes such as land, water, animals, environment, and the people's cultural or religious practices, which is more directly impacting on physical, mental, and spiritual health (Hidalgo et al., 1995). On the other hand, it may be true that some Eastern cultures have long used astrology to explain an extraordinary phenomenon instead of astronomical science. Had it used a different setting, different cause for speculations such as astrology, or a combination of earthquake, asteroid, or folklores, it would countervail the Western-centric science contextualization. The challenge is that it's often difficult for test item developers to think about who are the targeted test populations--their cultures, perceptions, and assumptions. Yet, identifying and being aware of the cultural issues, as well as being familiar with the targeted student populations and their cultural heritages, is helpful to enact equitable and effective knowledge, especially for the underserved student groups in the local setting.

Case 2: CS601 “Sustainable Fish”

SUSTAINABLE FISH FARMING

An increased demand for seafood is placing a greater burden on populations of wild fish. To reduce this burden, researchers are investigating ways to grow fish sustainably in fish farms.

Two challenges to creating a sustainable fish farm include (1) feeding the farmed fish and (2) maintaining water quality. Farmed fish require large amounts of food. A fish farm that is sustainable will grow the food needed to feed the farmed fish. Waste from the fish can build up in the farm to levels that are dangerous to the fish. In a sustainable fish farm, there is a constant flow of ocean water through the farm. Waste and excess nutrients (food that algae and plants need to grow) are removed from the water before it is returned to the ocean.

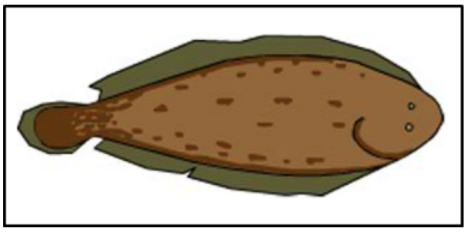


Figure 3-6. PISA 2015 Science general context “Sustainable fish”.

The context of *Sustainable fish*, which illustrates that the purpose of item contextualization, is also worth investigating. This context seems to portray a good purpose—growing fish sustainably for fish farms. It aims to solve the problem of environmental pollution caused by waste and chemicals, and that this facility is able to return clean water to the ocean. If we are to examine the context from the perspectives of critical race theory, an equitable context warrants critical lenses of whose knowledge of “sustainability” and why their knowledge is being presented, whose notion of science is, by whom, and for whose benefit. Our evaluative codes show the following:

> First, as the context illustrates, sustainable fish farms are there to reduce the burden of the demand for seafood. However, for many indigenous people such as Native Americans and First Nations, concerns over the impacts of fish farming on the wild fisheries, such as wild salmons, or the common soles, may outweigh any direct economic benefits from this fish farming. Many Native American children may be taught a different view to the idea of sustainability for a number of reasons, including potential bioaccumulation of chemicals or threats to their traditional ways of wild fishing. But more than that, it has to do with their cultural heritage,

> rights and titles, and general concerns over the health of their own communities and lands. For many, the notion of sustainability is not realized in the context, as their lives revolve around the ocean, iconic wild fish such as salmon. Sustainability is more about community and nature than making profits. It's kind like capital vs. spirituality.

Culturally, compared to the indigenous knowing of science and nature (Carjuzaa & Ruff, 2010), this context of doing sustainability for commercial fish farming satisfies a capitalist and materialist agenda, and is a different notion of sustainability from indigenous ways of balancing science-nature relations. With fish farms still physically located in waters connected to the ocean, there is a risk of marine pollution, the spread of disease, and the contamination of the genetics of other wild fish in the ocean. Therefore, the purpose of investigating science research or sustainable fish farming, has to take on a cultural perspective for the purpose of science, and its dynamic relationships with nature.

That said, a solution for an alternative representation of sustainable fish farming may, from an indigenous and environmentalist's perspective, calls for a completely closed and land-locked farm solution, which uses a recirculating system of waste and water. Examples eliciting more critical and creative thinking may be more preferable than a fixed solution. For instance, water could come from facilities like wells or the nearby river. Or perhaps, water can flow through the fish tanks and then is cleaned in a treatment plant and sent back to the fish. In short, a cultural perspective is needed to present the context in a more equitable way and with "strong-objectivity".

3.8 Conclusion

This study focusing on contextualized items shows how cultural validity is relevant in measurement and assessment. It is evident that culture permeates in different ways in

which students learn and process science concepts, and through which science assessment tasks are created and responded to. So far, we have explored the cultural validity of contexts and the utility of the ECI instrument in implementing the concepts on a sample of PISA science items. The results surface a descriptive 'master narrative' for science embedded in PISA item contexts, and a representation of scientific identities. We have also observed an intersectionality of the represented cultural identities in a skewed pattern. We found the significant factor of linguistic accessibility or complexity in predicting an item's difficulty is consistent with previous measurement research (e.g., Le Hebel et al., 2017). Our findings also show it is effective to interpret possible sources of DIF through a theory-oriented approach and a-priori cultural perspectives.

Apart from the exploratory model results, we have examined potential sources of bias in context subjects for items' general contexts. Common formation of bias can be from status differentiation, gendered behavior, or race-associated names (individuals) being positively or negatively oriented. Typical gender norms or stereotypes conveyed by item context could potentially reinforce the gender message that students receive. For linguistic agency, contexts in giving more proactive words like "leading" or "allowing" pertaining to male, White, middle-class 'scientists' and 'researchers' than local residents, natives, or other cultural groups, suggest that certain social identities are prescribed with more power and agency when representing science (Bramsen et al., 2011). They are exposed as problematic or implicit messages.

For item development and evaluation of item validity, we need to address the relationship between item characteristics and norms that could be the source of bias, inequality, or discrimination. Large-scale science assessments need to embrace the representation and inclusion of a variety of cultural groups in a multicultural society.

Minority group students may find a cultural name or example more relatable, and thereby make connections with science and identity belonging in the STEM fields. Lastly, we recommend practices such as the following in minimizing sources of construct-irrelevant variance for science contextualized assessment:

- acknowledge people of color and indigenous ways of knowing science, making sense of science, and connect with the student's indigenous funds of knowledge;
- flag and revise potential information about minority group bias and stereotypes;
- pursue a balanced and holistic approach overall for managing context representations, such as on a character's agentic representation in item's social context;
- increase context sensitivity and respect for a minority group's cultural history, by using realistic locations, practices, customs, resources, and ways of learning;
- annotate and be transparent about the existing variations of scientific knowledge and controversial practices.

3.9 Limitation and Future Direction

For the generalizability of the study, there are at least two concerns. The inferences are cautioned against the small-sample item-level modeling for item difficulty and DIF, but each has a plausible strength of interpretations for validity research. Evidently, we have to acknowledge the subjective nature of the ECI coding system to some extent due to the coder or researcher's positionality. Thus, continuing psychometric validation and other analysis is needed to evaluate the technical quality of the ECI instrument as well as the appropriateness and robustness of the cultural framework. Studies on a larger scale with the possible implementation of machine learning and natural language processing, will help to both explore and confirm relationships between item's contextual cultural features, student responses, and testing experiences. As such, more external and consequential construct validity research are needed.

Reference

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

Adams, R. and M. Wu (eds.) (2003), *Programme for International Student Assessment (PISA): PISA 2000 technical report.* PISA, OECD Publishing. https://doi.org/10.1787/9789264199521-en.

Ahmed, A., & Pollitt, A. (2007). Improving the quality of contextualized questions: An experimental investigation of focus. *Assessment in Education*, *14*(2), 201-232.

Bair, A. N., & Steele, J. R. (2010). Examining the consequences of exposure to racism for the executive functioning of Black students. *Journal of Experimental Social Psychology*, *46*(1), 127-132. https://doi.org/10.1016/j.jesp.2009.08.016

Berkhin, P. (2006). A survey of clustering data mining techniques. In J. Kogan & C. Nicholas (Eds.), *Grouping multidimensional data: Recent advances in clustering* (pp. 25-71). Springer.

Banks, J. A. (1993). The canon debate, knowledge construction, and multicultural education. *Educational researcher*, *22*(5), 4-14.

Bernstein, B. (1996). *Pedagogy, symbolic control and identity: Theory, research, critique,* Taylor & Francis.

Boaler, J. (1994). When do girls prefer football to fashion? An analysis of female underachievement in relation to 'realistic' mathematic contexts. *British Educational Research Journal*, *20*(5), 551-564. https://doi.org/10.1080/0141192940200504

Bramsen, P., Escobar-Molano, M., Patel, A., & Alonso, R. (2011, June). Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 773-782). Association for Computational Linguistics.

Carjuzaa, J., & Ruff, W. G. (2010). When Western epistemology and an Indigenous worldview meet: Culturally responsive assessment in practice. *Journal of Scholarship of Teaching and Learning*, *10*(1), 68-79.

Collins, P. H. (2002). *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. Routledge.

Cooper, B. & Dunne, M. (2000). Constructing the "legitimate' goal of a 'realistic' math item: A comparison of 10-11 and 13-14 year-olds, In A. Filer (Ed.). *Assessment: Social practice and social product* (pp. 87-109). Routledge.

Cooper, B., & Harries, T. (2002). Children's responses to contrasting realistic mathematics problems: Just how realistic are children ready to be?. *Educational Studies in Mathematics*, *49*(1), 1-23.

Cronbach, L. J. (1988). Five perspectives on validity arguments. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Lawrence Earlbaum.

del Rosario Basterra, M., Trumbull, E., & Solano-Flores, G. (Eds.). (2011). *Cultural validity in assessment: Addressing linguistic and cultural diversity*. Routledge.

Gay, G. (2010). *Culturally responsive teaching: Theory, research, and practice*. (2nd edition). Teachers College Press.

González, N., Moll, L. C., & Amanti, C. (Eds.). (2006). *Funds of knowledge: Theorizing practices in households, communities, and classrooms*. Routledge.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, *102*(1), 4.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Allyn and Bacon.

Hartigan, J. A., & Wong, M. A. (1979). A K-means clustering algorithm. *Applied Statistics*, *28*(1), 100-108. https://doi.org/10.2307/2346830

Harding, S. (1993) Rethinking Standpoint Epistemology: What is 'Strong Objectivity'?. In L. Alcoff and E. Potter (Eds.), *Feminist Epistemologies.* Routledge.

Hidalgo, N. M., Siu, S. E., Bright, J. A., Swap, S. M., & Epstein, J. L. (1995). Research on families, schools, and communities: A multicultural perspective. In J. A. Banks, & C. A. Banks (Eds.), *Handbook of research on multicultural education* (pp. 498-524). Jossey-Bass.

Huang, X., Wilson, M., & Wang, L. (2014). Exploring plausible causes of differential item functioning in the PISA science assessment: Language, curriculum or culture. *Educational Psychology*, *36*(2), 378-390. https://doi.org/10.1080/01443410.2014.946890

Jolliffe, I. (2011). Principal component analysis. *International Encyclopedia of Statistical Science*, 1094-1096. https://doi.org/10.1007/978-3-642-04898-2_455

Kane, M. T. (1992). An argument based approach to validity. *Psychological Bulletin*, *112*, 527-535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Praeger Publishers.

Kusimo, P., Ritter, M. G., Busick, K., Ferguson, C., Trumbull, E., & Solano-Flores, G. (2000). *Making Assessment Work for Everyone: How To Build on Student Strengths*. WestEd, 730 Harrison Street, San Francisco, CA 94107-1242.

Ladson-Billings, G., & Tate, W. (1995). Toward a critical race theory of education. *Teachers College Record*. *97*(1), 47-68.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Le Hebel, F., Montpied, P., Tiberghien, A., & Fontanieu, V. (2017). Sources of difficulty in assessment: Example of PISA science items. *International Journal of Science Education*, *39*(4), 468-487. https://doi.org/10.1080/09500693.2017.1294784

Little, C., & Jones, K. (2010). The effect of using real world contexts in post-16 mathematics questions. *Proceedings of the British Society for Research into Learning Mathematics, 30*(1), 137-144.

Messick, S. (1987). Validity. *ETS Research Report Series, 1987*(2), i-208. https://doi.org/10.1002/j.2330-8516.1987.tb00244.x

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge.

Noble, T., Suarez, C., Rosebery, A., O'Connor, M. C., Warren, B., & Hudicourt-Barnes, J. (2012). "I never thought of it as freezing": How students answer questions on large-scale science tests and what they know about science. *Journal of Research in Science Teaching*, *49*(6), 778-803.

OECD (2017). PISA 2015 Science Framework, In *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving.* OECD Publishing, Paris, https://doi.org/10.1787/9789264281820-3-en.

OECD (2009), *PISA 2006 Technical Report.* PISA, OECD Publishing. https://doi.org/10.1787/9789264048096-en.

Orner, M. (1996). Teaching for the moment: Intervention projects as situated pedagogy. *Theory into Practice*, *35*(2), 72-78.

Paek, I., & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel-Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement*, *71*(6), 1023-1046. https://doi.org/10.1177/0013164411400734

Robitzsch, A., Kiefer, T., & Wu, M. (2020). *TAM: Test analysis modules.* R package version 3.5-19, https://CRAN.R-project.org/package=TAM.

Ruiz-Primo, M. A., & Li, M. (2015). The Relationship between science assessment context characteristics and student performance: The Case of the 2006 and 2009 PISA science assessments. *Teachers College Record: The Voice of Scholarship in Education, 117*(1), 1-36. https://doi.org/10.1177/016146811511700118

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*(6), 613-629.

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.

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

Solano-Flores, G. (2011). Assessing the cultural validity of assessment practices: An introduction. In M. del Rosario Basterra, E. Trumbull, & G. Solano-Flores (Eds.). (2011). *Cultural validity in assessment: Addressing linguistic and cultural diversity*. Routledge.

Solano-Flores, G., Wang, C., & Shade, C. (2016). International semiotics: Item difficulty and the complexity of science item illustrations in the PISA-2009 international test comparison. *International Journal of Testing*, *16*(3), 205-219.

Stembridge, A. (2019). *Culturally responsive education in the classroom: An equity framework for pedagogy*. Routledge.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of data clusters via the Gap statistic. *Journal of the Royal Statistical Society B*, *63*, 411-423.,

Wu, M. L., Adams, R. J., & Wilson, M. (1998). *ACER ConQuest: Generalised item response modelling software*. ACER press.

Yildirim, H. H., & Berberoĝlu, G. (2009). Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. *International Journal of Testing*, *9*(2), 108-121. https://doi.org/10.1080/15305050902880736

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

Chapter 4. Modeling Sociocognitive and Sociocultural Context Effects and Student Performance for Contextualized Assessment: A Bayesian Hierarchical Explanatory Item Response Model

Nixi Wang [1], Klint Kanopka[2], Min Li [1], Dongsheng Dong [1], Philip Hernandez[2], Maria Araceli Ruiz-Primo[2], Jim Minstrell[3]

[1] *College of Education, University of Washington, USA*

[2] *Graduate School of Education, Standard University*

[3] *Facet Innovations*

Abstract

The design of contextualized assessment is commonly used in large-scale assessments with its apparent advantages. Yet, research on calibrating item contexts to model the probability of success on task-solving is minimal. There is a need for item response models to incorporate context frameworks to parse out construct-irrelevant variances. This paper addresses the sociocognitive and sociocultural characteristics of contextualized items in a testlet-based physics assessment for multiple levels of context, and examines corresponding effects on students' ability of solving tasks. Thus, we proposed and estimated a mixed-effect Bayesian item response theory model with student and item covariates, which incorporated context features and racial demographic groups as explanatory effects. Results showed context characteristics such as cognitive demand, cultural bias, and context familiarity are associated with student performance,

and the effects vary based on students' racial background. Implications for test item development and validity claims of contextualized assessments are discussed.

4.1 Introduction

Contextualized assessment tasks have been widely used in national and international large-scale assessments, such as in the Programme for International Student Assessment framework (OECD, 2017). Exemplified by the testing standards (AERA, APA, & NCME, 2014), there is a growing awareness in testing that advocates the evaluation and development designs that incorporate the underlying cognitive and sociocultural underpinnings of testing conditions, especially at the item level to address fairness and equity concerns. Toward this end, research on contextualized items aims to systematically evaluate and discover the socio-cognitive and cultural mechanisms underlying the testing process, with some efforts of combining sociocultural and sociocognitive theories of learning in measurement process and practices (Ruiz-Primo & Li, 2015, 2016; Wang et al., 2019; Wang, 2019).

The context in contextualized items allows measurement experts a window of leveraging the function and utility of items, which in turn offers opportunities on how contextual features may influence test respondents' performance. For instance, research efforts have been made in recent years to encode and decode context features, with the goal of engineering item contexts in more interpretable, sustainable, and fairer ways (Dong, 2020; Ruiz-Primo & Li, 2015; Wang et al., 2017). Previous work on cognitive interviews and assessment statistics indicates some promising sociocultural functionality of item contexts. For instance, the use and student's perception of realistic and authentic contextualized tasks are often considered in conjunction with student's cultural background (Cooper & Harries, 2002). Good and culturally valid contexts are

considered beneficial and motivating, as they help facilitate students' understanding of the nature of the problem, and thus promote students' constructive process of solving the tasks. However, questions remain unclear on the causal relations of the context features and student performance. We need to have a systematic approach to characterize and model interpretable categories from the complex and rich sources of information situated in item contexts.

In this current work, we seek to fill the gap of contextualized measurement research by analyzing a dataset from the DECISA science assessment, which designs principled feature building and experimentation of item contexts (Ruiz-Primo et al., 2019). The test, like other large-scale assessments, have a set of items nested at a secondary level of sub-clusters, called subtestlets, and a higher level of clusters called testlets. From a measurement perspective, error or nuisance variance can be due to *context* facet when the information presented in any of those three context levels is confusing or difficult. In this paper, we examine four context domains of cognitive demand, sociolinguistic familiarity, cultural bias, and physics nature and topics. We adopt a statistical framework of item response theory to formulate the relationship between student characteristics on the latent ability and observed item responses. Specifically, we seek to address the following research questions:

1. How do sociocognitive and sociocultural characteristics of the context in physics contextualized problems affect students' test performance with respect to their latent proficiency level? Specifically, how do cognitive, sociolinguistic familiarity, and cultural bias features of context predict students' success probability of solving an item?
2. What are group-specific context effects considering students' racial cultural background?

This study is organized as follows. First, the theoretical rationale of our context-coding framework is described and synthesized. The assessment of physics topics on forces and motion provides detailed coding examples at contexts' testlet, subtestlet, and (possible) item levels[1]. We then propose and formulate a Bayesian hierarchical explanatory IRT model, and apply the model to a physics assessment dataset. We will discuss findings on selecting the best performing model, followed by examining context effects estimated from the model, as well as accounting for individual cultural characteristics.

4.2 Background on Cognitive and Cultural Characteristics of Context

Contextualized items consist of sets of items that share a common stimulus (e.g., scenario or problem set-up). In contextualized assessment, students' knowledge and proficiencies toward a problem are inevitably mediated through a complex pool of factors, including relevant quality of items and student's personal and sociocultural background. We hypothesize that good qualities of sociocognitive and sociocultural contextual features in a contextualized item will enable student's transfer of learning and use of their funds of knowledge, thus facilitating the activation and access to the retention of information stored for solving a problem task (Ruiz-Primo & Li, 2015).

Sociocultural theories posit that individuals see the world through their respective cultural lenses, such as their language, ethnic and cultural communities, home practices, religion, gender, and racial identities. For instance, Carjuzaa and Ruff (2010) highlight the nature-culture relationships of science knowledge from the perspectives of American Indians. Students from Crow and Northern Cheyenne tribes

---

[1] To assess the robustness of the model estimates, we conduct an additional simulation study to evaluate how the qualities of model-based parameter estimates vary across different conditions of test design.

tend to think in “an indirect, circular, relationship-based, and contextualized manner”, which is representative of an Eastern cultural communication style, whereas most other white students conduct language in “a Socratic, direct, concise manner”, which exemplifies the Western linear communication style (p. 70).

On a cognitive lens, item context contains information and knowledge that are of high relevance for knowledge application and concretization, so that students can situate what knowledge they have learned in a context and make inferences as they respond to the related task. Scientific knowledge requires high relevance to context-based problems. The inherent level of cognitive process associated with reading and understanding the situated context necessitates students’ connection and exploration of the problem as a precondition for successful problem-solving.

Our cognitive properties of context are obtained based on educational-psychological and sociocognitive theories (Mislevy, 2018; Atkinson et al., 2007). According to sociocognitive theories, key cognitive features of contextualized problems include cognitive complexity (Schneider et al., 2013), Higher-order thinking (Greiff et al. 2013), reading or cognitive demand (Ruiz-Primo and Li, 2016; Crisp and Grayson, 2013; Morrison and Embretson, 2014), contextualization, concretization, and visualization of contextualization (Dong, 2020; Ruiz-Primo & Li, 2015; Vos, 2014). In more detail, examples of high cognitive demand may point to a contextualized problem that contains a number of technical terms, or difficult vocabulary that causes a heightened level of working memory, reasoning, and understanding of scientific problems. More notably, cognitively demanding situations or activities usually intersect with students’ sociolinguistic and immigration backgrounds (Atkinson et al., 2007).

Another important sociolinguistic feature is the *familiarity* of context (Lee, 2011), whether the language is familiar to all or most of the students, independently of

cultural ethnic group, geographic region, social class, or linguistic background. As it posits, sociolinguistic familiarity requires respondents less working memory and facilitates comprehension through automatic schema-driven processing (Song & Bruning, 2016; Li et al., 2012; Lee, 2011). Familiar contexts, such as daily life experiences and classroom experiences (Dong, 2020), assume general knowledge that is more common to all students from socioeconomic, language, and cultural backgrounds. Furthermore, familiar contexts are more likely to be realistic and authentic (Cooper and Dunne, 2000), by connecting students' funds of knowledge, emotion, experience, and cultural awareness with scientific applications, empirical phenomena, and practical settings.

Meanwhile, *cultural bias* implies that biased group-identity associated information may trigger psychological activation, stress, stereotype threats, and negative affections. The variable of *context length* is another indicator for linguistic information and is deemed essential based on extant theory and empirical studies. We hypothesize that longer contexts overall present more concrete and richer information for student's task in an efficient assessment. However, the appropriateness of longer context needs to be evaluated with caution because too long pieces of information may burden a respondent's information process. In addition to the core characteristics of cognitive and sociocultural demand of context, the process of *physics knowledge* is categorized into three knowledge types: epistemic, procedural, and factual scientific knowledge (Ruiz-Primo & Li, 2015). In terms of performance dimensions, those aforementioned context domains can be sources of extra difficulty. These potential significant context effects, if not correctly accounted for, can cause the person's latent trait estimates to be overestimated or item parameter estimates to be biased (Wainer et al., 2000).

4.3 The DECISA Context Experimentation and Context Coding System

The DECISA physics assessment is developed as a collaborative project effort by a team of university researchers, physics experts, and teachers. The assessment consists of 2 booklet designs with a total of 35 physics items, which intend to measure students' understanding and cognitive skills of force and motion topics.

Let's reiterate, a contextualized assessment item, from an item engineering perspective, can be viewed from two systematic components: subject content or concept, of which an item intends to measure students' knowledge; and item context, which facilitates the understanding of subject knowledge or concept of an item. Item and context programming have gained more attention for its usefulness in investigating, monitoring, and evaluating item psychometric conditions. While contents or concepts are usually fixed by test design, engineering context features and inspecting significant features is of increasing interest in recent measurement research. A break-down example of contextualized physics item on force and motion in the DECISA assessment is provided in Figure 4-1.

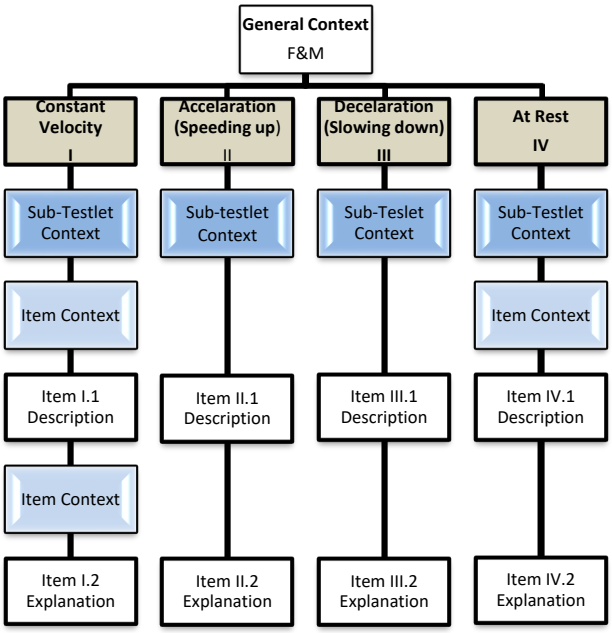


Figure 4-1. Prototype of contextualized items in a testlet.

Context Manipulation. The DECISA context experiment approach controls for similar item format and common physics construct across items, but only by manipulating context wording and presentations at the testlet and sub-testlet levels. Every two items belong to the same subtestlet where four types of force and motion fundamental ideas (at rest, speeding up, slowing down, or constant velocity) are mixed. Then every four items share a common testlet context, presented in scenarios of bus, sled, cart, and box. Across the four scenarios, testlet context pairs of bus and sled, and cart and box are considered parallel, with physics ideas and questions consistent with each other, but only wording and details have differed to different degrees.

Context coding. To systematically investigate the sociocognitive and sociocultural factors that may contribute to individual problem solving, we developed a coding system that applies expert coding at each of the testlet, sub-testlet, and item levels. At each level, context codes are generated to reflect a range of context features including cognitive comprehensibility, context functionality, cultural ethical considerations, and the nature of physics concepts and knowledge (Figure 4-2).

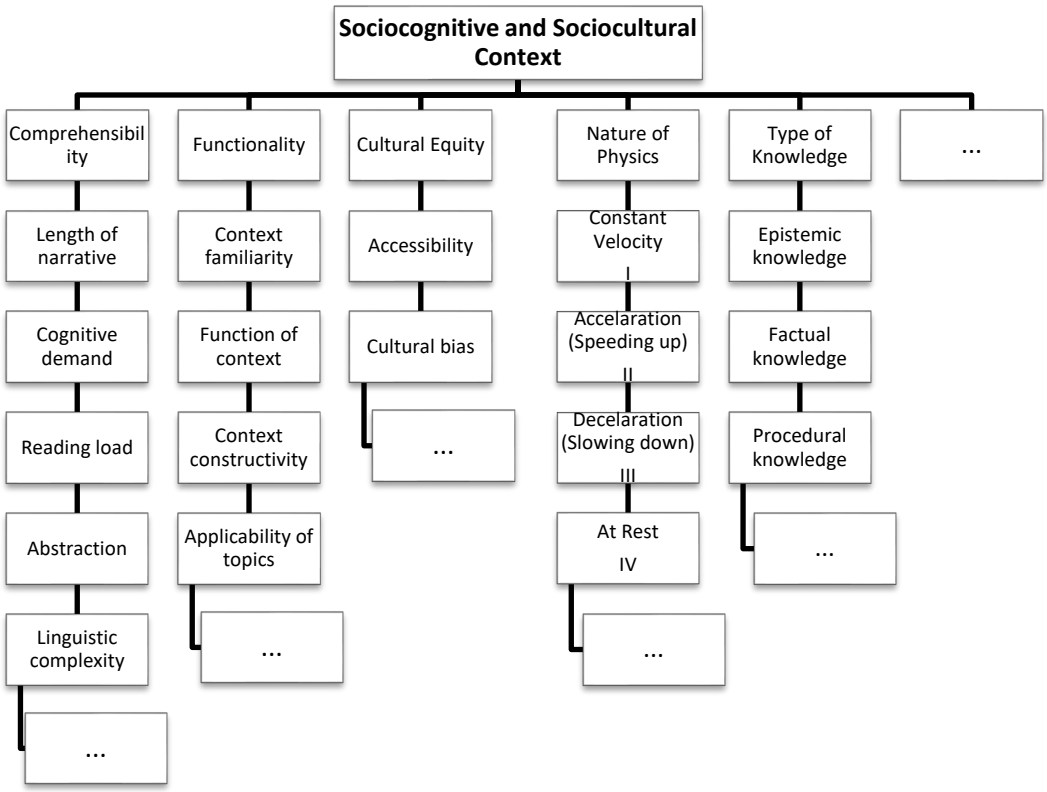


Figure 4-2. Sociocognitve and sociocultural context features engineered in the DECISA context coding system.

Performing context coding on a large scale with psychometric desirability is a challenging task. We have a team of seven researchers and experts conducting concurrent and separate coding to make sure the quality of codes are well retained. Our focal 32 physics items resulted in a context sample size of 58 across the three levels. While our instrument covered both the texts and visual components of the contexts, due to the limited sample sizes, we are only interested in investigating the direction and magnitude of textual context effects as item covariates. From this section on, an item context if mentioned will refer to our interest in subtestlet-level context, since we only have four testlet groups. Therefore, the testlet context effects are considered as fixed and relevant sociocognitive and sociocultural features not investigated in the modeling process due to the limited sample size.

4.4 Methods

4.4.1 *DECISA Data*

1631 middle school students in the 2017-18 school year from school districts in two states in the Pacific Northwest region participated in this study. Four testlets of items were grouped into two booklets with three anchor items for equating. Students from 67 classrooms were randomized to participate by periods of classes. Half of the students received booklet 1 (N = 825) and the remaining half received booklet 2 (N = 806). Each student took one booklet containing 20 multiple-choice physics questions and was given 50 minutes to complete the exam.

***Feature Selection*.** Since we have a long list of explanatory variables, we need to judiciously select the smallest number of predictors in order to control for their effects. For modeling purposes, we apply LASSO feature selection to reduce the chance of overfitting. While there are many ways to do this, LASSO applied on features and the estimated item difficulty parameter values as outcome variable uses a shrinking (regularization) process where it penalizes the coefficients of the regression variables shrinking some of them to zero (Muthukrishnan & Rohini, 2016). In this way, it can balance the variance-bias trade off, and can provide a good prediction accuracy when we have a small number of observation and a large number of features. The risk is that even small but undetected nonlinearities or indirect effects among predictors can seriously bias parameter estimates.

4.5 The Bayesian Hierarchical Explanatory Item Response Theory Model

Models such as the mixture linear logistic test model (LLTM; Mislevy & Verhelst, 1990) and the more generalized model belonging to the class of mixture IRT models, the mixture random weights LLTM (Choi & Wilson, 2014), were developed to examine key characteristics of item properties. In this study, context categories such as those at

the testlet and subtestlet levels can be indicated through the LLTM framework as item properties.

One approach for item analysis is to descriptively quantify item characteristics and model respondents' performance together with the observed covariates. Alternatively, a multiple-step approach can be used in which respondent's latent proficiency is modeled first based on their item response (e.g., through maximum a posteriori estimates), and subsequently with estimated item parameters, a regression model is placed on the context's nominal categorical response variables and examine context characteristics as explanatory effect. Currently, with our item pool being small, we adopt a concomitant approach of Bayesian hierarchical modeling, which models context variables, item responses, and student trait concurrently in a measurement item response theory model. The advantages of choosing the Bayesian concomitant modeling approach over the multiple-step frequentist approach is that 1) Bayesian methods do not assume strictly distributional assumptions such as normality, 2) it can model curvilinear or conditional relations with more ease than the strictly linear relationships, 3) the posterior distribution of model parameters can be estimated for each observation with the probabilities carried through the modeling process, and 4) the Bayesian approach is more advantageous to our task as it can scale better to more complex models as more robust and flexible. Compared to point estimates, model parameters are viewed as a sample from a population distribution of parameters, thus providing posterior distribution values as rich information about the parameters. It is also more straightforward to make group comparisons in a Bayesian fashion (Kruschke, 2014). Lastly, with the hierarchical structure, information across levels and nominal factors can be shared through partial pooling as appropriate, where parameter estimates can be

more robust as well as less influenced by extreme patterns and noise in the data (Gelman & Hill, 2006).

4.5.1 *Model Specification*

Here we use a formulation of IRT parameterization incorporating partially pooled person and item parameters similar to those in generalized linear multilevel models (GLMMs) (De Boeck et al., 2011). Regarding the specific natures of covariates, we consider item context effects may vary across persons and items and be correlated with one another, and each racial subgroup is described as having an underlying ability distribution, with each person as a random variant of the distribution. We take the step to further jointly modeling persons' responses together with context and student characteristics as covariates in a joint measurement model. In this way, the standard deviation of the posterior distribution is obtained as the conditional standard error of measurement (CSEM). Incorporating prior knowledge in psychometric testing as prior distributions, we aim our model will potentially foster a theoretical understanding of the context function and model the effects of context characteristics in students' underlying cognitive process.

In our modeling approach of the Bayesian Contextualized Item Model (BCIM), the parameters in conjunction with the IRT one-parameter framework can be described as (1) item difficulty ξ_i that describes the threshold of answering an item correct as item property, (2) θ_p as the ability of person p in the sense that higher values of θ_p imply higher success probabilities regardless of the administered item. (3) Additionally, we want to estimate the effects of person or item *covariates* varying between persons and items. In this case, the context features are considered as internal item covariates (De Boeck et al., 2011) and thus without measurement error. We think the effect of contexts are constant within each item.

A simple BCIM specification in IRT 1PL framework is specified as follows:

$$y \sim Bernoulli(\eta_{ip})$$

$$\log(\eta_{ip}/(1-\eta_{ip})) = logit(\eta_{ip}) = \theta_p + \xi_i +$$

$$\sum_{j=1}^{J} b(item\ covariates)_j x_{ji} + \sum_{k=1}^{K} b(person\ covariates)_k x_{kp}$$

$$\theta_p \sim Normal(0, \sigma_\theta)$$

$$\xi_i \sim Normal\ (0, 3)$$

$$\sigma_\theta \sim HalfCauchy(0, 5)$$

where θ_p and ξ_i are person and item parameters, b_j and b_k are the regression coefficients, and x_{ji} is the value of the *j*th predictor for item *i*. When items are considered independent from one another, item prior distribution is usually specified as a normal distribution with mean 0 and standard deviation σ_ξ. As observed in psychometric practices, the item easiness parameters of a Rasch model is usually within *Normal*(0, 3), and can be used as a weakly informative prior on the logistic scale of responses. Note under the above parameterization, we specifically use the item easiness formulation ($\psi = \theta_{ip} + \xi_i$) instead of the Rasch framework $\theta_{ip} - \xi_i$ for the linear model. Both formulations are equivalent but the former is in alignment with the formulation in the R package *brms* (Bürkner, 2017; Bürkner, 2018) we used. Therefore, we can interpret ξ_i as the easiness of item, with higher values of ξ_i implying higher success probabilities regardless of person. We also notice the assumption of independence of items are violated in testlet-based assessments. In our parameterization, we hypothesize the dependent structure is due to item clusters sharing the same context, therefore the formulated item contextual effects. Potentially, we can also assume a hierarchical multivariate normal distribution on the correlated item parameters for the prior in the form of

$$(\xi_{1i}, \ldots, \xi_{Ki}) \sim Multinormal(0, \Sigma_{\xi})$$

where ξ_{Ki} is the item parameter of item *i* for a distributional parameter ψ_K in generalized form, and Σ_{ξ} is the item covariance matrix.

We place weakly informative priors on variance components. For the hyperparameters, i.e., the standard deviations and correlation matrices, we use the half-Cauchy and the LKJ prior (Lewandowski et al., 2009), LKJ (2), for the correlation matrix in *brms* and *Stan* (Stan Development Team 2019). If necessary, we can also specify the context covariate prior. Here the flat "uninformative" prior (the default in *brms* is used) since the variables listed have already gone through feature selection before entering the model.

As mentioned, we use Bayesian methods for estimation not only because we can consider prior knowledge into the model, but also its flexibility and feasibility of computing complex models with our interest in estimating the full distribution of parameters than single estimates. Estimation is performed in *Stan* using Markov-Chain Monte-Carlo (MCMC) sampling via adaptive Hamiltonian Monte Carlo (HMC) sampler (Hoffman and Gelman 2014; Stan Development Team 2019), which is frequently used in Bayesian IRT modeling and other high-dimensional models for sampling from posterior distribution (see for example, Fox, 2010; Levy & Mislevy,2017; Rupp, Dey, & Zumbo, 2004). HMC sampler compared to the conditional sampling of the Gibbs algorithm generates more efficient random walks and mix at higher speeds by adding a momentum variable through exploring the parameter space. For each of the sequence of MCMC chains, the first 2000 warm-up iterations out of 4000 were discarded for parameters and hyperparameters.

4.5.2 *Analytic Strategy*

Model Selection. With different constraints on person and item fixed and random effects, we evaluate models separately with convergence diagnostics, and compare models based on model fits. We use a Bayesian 1PL model as the baseline model, where item parameters are considered as random effects. Additionally, considering the assumption that local item independence is not hold with the testlet structure, we also fit a Testlet 1PL model to the data to examine the testlet effects. For the BCIM predictors, we have $\eta_{ip} = b_{00} + b_1 \,\text{context length}_{\text{i}} + b_2 \,\text{cog demand}_{\text{ip}} + b_3 \,\text{familiarity}_{\text{ip}} + b_4 \,\text{bias}_{\text{ip}} + b_5 \,\text{item}_{\text{ip}} + u_{0p} + e_{ip}$ as BCIM context-only model. Lastly, we have the BCIM full conditional model incorporating student characteristics and can be written as: $\eta_{ipg} = b_{00} + b_1 \,\text{context length}_{\text{i}} + b_{2g} \,\text{cog demand}_{\text{ipg}} + b_{3g} \,\text{familiarity}_{\text{ipg}} + b_{4g} \,\text{bias}_{\text{ipg}} + b_{5g} \,\text{Race}_{\text{pg}} + b_{6g} \,\text{item}_{\text{ipg}} + u_{0pg} + e_{ipg}$. For both BCIM specifications, we consider a maximal multilevel model (Barr et al., 2013) for our parameters. Under such design, the context-only model has within-testlet contexts as fixed effects for the intercept and the slope, and correlated random intercepts and slopes for respondents (individual students' perceived understanding of context's cognitive demand, familiarity, and bias may vary). For the full model, where the person and item-characteristic variables are simultaneously modeled, we sequentially examine from free to more constrained conditions, such as the correlated random intercepts and slopes of context and student variables for items.

There are two common types of model performance indices: information-based criteria with bias correction, and predictive performance criteria with many indices amidst them. We will use the leave-one-out information criterion (LOOIC) and the widely applicable information criterion (WAIC), which consider the pointwise log-likelihood of the full Bayesian posterior distribution (Vehtari et al., 2017). Both LOOIC

and WAIC provide estimates of the out-of-sample predictive accuracy, which samples new student's latent proficiency from the racial subgroup, generates new response data from a new student sample, and evaluates how well a model makes predictions about the new dataset.

To avoid overfitting, we implement the approximate LOO-CV laid out in Vehtari et al. (2017), which is a relatively new and more favorable procedure than AIC and DIC criteria for stabilizing and diagnosing importance weights. From existing posterior simulation draws, we compute LOO-CV using Pareto smoothed importance sampling (PSIS; Vehtari et al., 2019), with the higher Pareto *k* diagnostic >1 indicating bad reliability of the estimates. PSIS LOO-CV is more preferred as it provides useful diagnostics and effective sample size as well as Monte Carlo standard error estimates. When we had a well-specified model, we expect the estimated effective number of parameters (*p_loo*) to be smaller than or similar to the total number of parameters in the model. We used the functions in the *loo* package (Vehtari et al., 2017) to compare the LOOIC and WAIC values.

Model Diagnostics. We assess posterior convergence by examining trace pots and diagnostic statistics. For instance, the Geweke Diagnostic shows the z-scores for a test of equality of means between the first and last parts of each chain, which should be < 1.96. In this way, it checks whether a chain has stabilized. In Stan and brms output, a lack of convergence for each individual posterior fit can be seen for Rhats (the recommended cut-off is 1.05). However, these kinds of criteria cannot always guarantee convergence. Even though they can certainly indicate problems with convergence, other

problems might still be present[2]. To assess model adequacy, we take the observed data and perform the posterior predictive checks, where we simulate data based on the posterior samples of model parameters. The results allow us to investigate the data generating process and get an intuitive sense that the distribution of the generated data resembles the distribution of the actual data.

Person, Item, and Covariate Parameters. With the prior confidence established on the chosen model, we can make posterior inferences and compute summary statistics. For the context parameters, we will plot the posterior distributions of parameter estimates and test the relative evidence for two competing hypotheses, i.e., the estimated relative evidence for the existence of the sociocultural context effects against its non-existence by comparing a model with the intercept to a model without the intercept. The Bayes Factor (BF) quantifies a ratio of marginal likelihoods, which relies on an approximation of the marginal likelihood through bridgesampling (Gronau et al., 2017). For example, the BF of model 1 indicates the likelihood of the observed data under a given hypothesis (e.g. the overall context effect differs from 0) relative to another hypothesis (e.g. the effect is equal to 0). The relative strength of evidence for a hypothesis can be anecdotal (1:3), moderate (3:10), and strong (10:30), and very strong (30:100). We will also report the 95% credible intervals (CrI), which are a Bayesian equivalent to confidence intervals, except that they have a 95% probability of containing the population value of the parameter.

---

[2] For analysis of model sensitivity, we conducted prior predictive checks to investigate how sensitive estimated posterior model parameters are to the true simulating parameters for the given set of design and prior parameters.

Additional analyses. A simulation design is proposed to oversee the quality of model-based parameter estimates and model sensitivity (See the design and analytic plan in supplemental materials). In addition, analyses using the two-step GLMM through the maximum likelihood estimation will be reported in Appendix B to compare results from the frequentist's methods. Item difficulty estimates from Rasch models will be obtained first, followed by a two-level GLMM incorporating subtestlet contextual effects.

4.6 Results

The results for the empirical dataset are presented in three steps. To investigate context effects, we start a descriptive analysis at the testlet level context. At the subtestlet level, we present the selection and evaluation of Bayesian models fitted to the data and corresponding psychometric desirability, the visualization and interpretation of parameter estimates from the best performing model, and further comparison of the context effects (the parameter of interest) across racial subgroups.

4.6.1 *Testlet Effect Analysis*

Since rich DECISA contexts are structured mostly at testlet and subtestlet levels, we first observe student distributions under the four testlets: Bus, Sled, Box, and Cart. Note that items paralleled in sequence within the testlets measure the same force or motion concepts, and only their situated contexts differ in their construction of context features (See item descriptive statistics in Appendix 2.A). For an intuitive observation, we normalize student scores to compare subgroup score distributions. Figure 4-3 reveals the average standardized student scores were lower than the grand mean for racial minority groups. The testlets with socioculturally unbiased contexts tend to have the slightly higher density for Black, Hispanic, and other student minority groups near or

lower than the mean score range. Overall, we find no obvious group differences based on testlet contexts.

We compute the intra-class correlation (ICC) to estimate the relative variability associated with item and person groups. For model specification, if between-person or between-item ICC is high, there may be a need to adjust the model and allocate a unique intercept for that level. The ICC for 1PL model is 0.25, indicating the proportion of variance between item and between people are 25% out of the total variance.

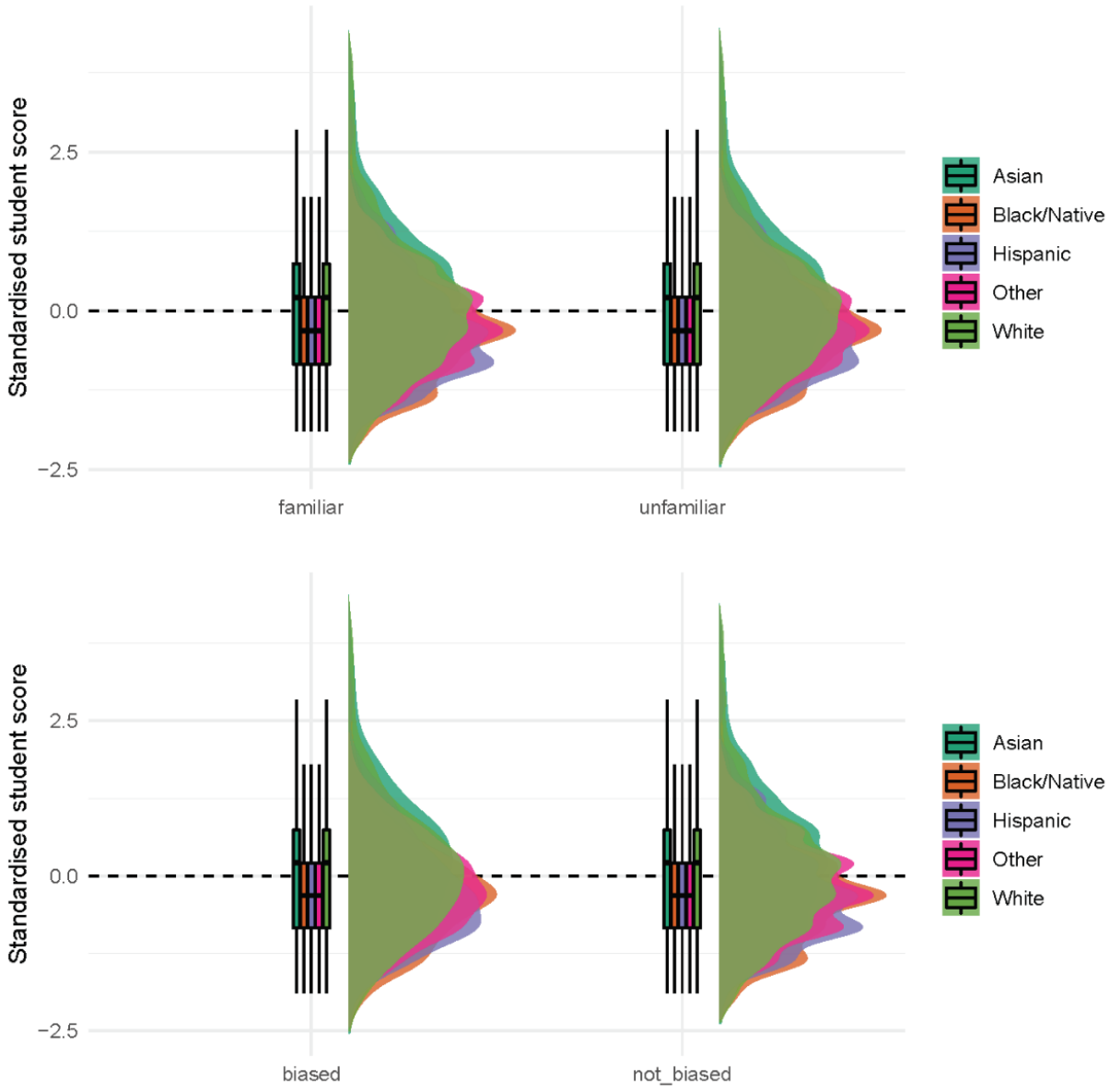


Figure 4-3. Standardized student's test scores at testlet level context characteristics.

Note. The upper and lower panel represent student score distribution by sociolinguistic familiarity and sociocultural bias at testlet level, respectively. Colors represent students' racial groups. The boxplot represents the median, first, and third quartiles.

With items nested within a testlet share a dependence structure, it is logical to examine the testlet effect by applying our second model: the Testlet 1PL model. The four testlets are considered random but not limited testlet examples. We draw random samples from the posterior random effect parameter estimates and plot the differences between paired testlets (measuring the same concept). Figure 4-4 shows the posterior means of testlet effects fall around or below 0.5. The difference between Box and Cart testlet is further away from zero than the difference between Bus and Sled test effects. While Bus and Sled testlet have larger variances, it is interesting to note that the Box and Cart testlets share similar context characteristics in terms of the context length, cognitive demand, sociocultural bias, and sociolinguistic familiarity (See Table 4-1). The difference in the two testlet effects could be due to other testlet context differences, or the variability of contexts at the subtestlet levels.

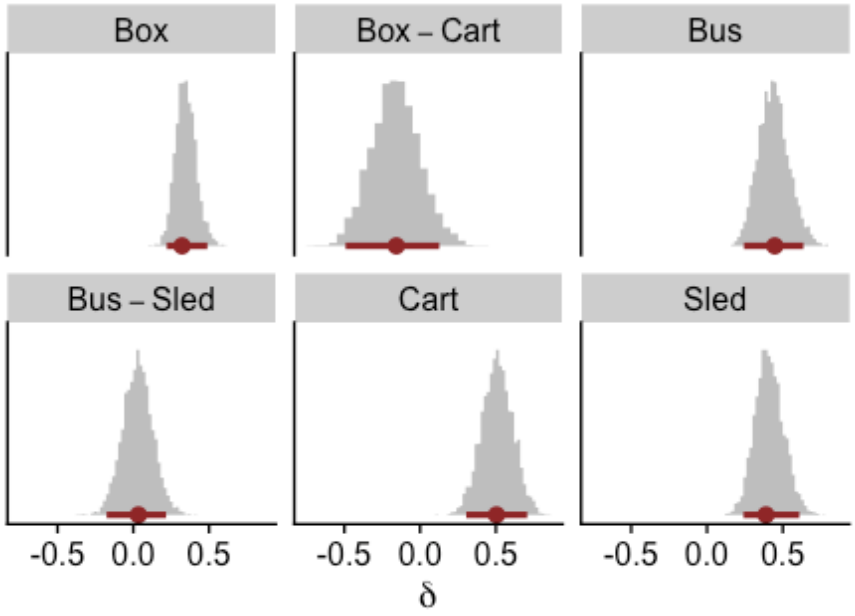


Figure 4-4. Posterior testlet effects and difference in effects for response probability.

Table 4-1

Testlet Effect Variances for the DECISA testlets

Context category															
Testlet context	Box				Cart				Bus				Sled			
Variance τ	0.10				0.10				0.21				0.31			

Note. Orange, purple, blue, and grey colors represent four context categories of context length, cognitive demand, sociocultural bias, and sociolinguistic familiarity in sequence. Red colors each represent longer length, high-demand, biased, and unfamiliar context in sequence.

4.6.2 *Fit of the Models*

For the four types of models, we consider two approaches to assess the performance of a given model or for model selections. The first is information criteria including the pointwise log likelihood measure for each data point. The WAIC has several properties that make it a better measure of within-sample predictive accuracy than the AIC and DIC (Gelman, Hwang, & Vehtari, 2013). The log pointwise predictive density (LPPD), a measure of within-sample predictive accuracy (Gelman et al., 2014), along with ELPD-WAIC indicate the BCIM full model has the best fit overall (Table 4-2). The second approach, Leave-One-Out Cross-Validation (LOO-CV), which focuses on predicting new data given new parameter values, i.e., the *out-of-sample* predictive accuracy, has asymptotically equivalent results.

Table 4-2

Model Comparison Based on Measures of Predictive Accuracy

	1PL Mean (s.e.)	**Testlet 1PL** Mean (s.e.)	**BCIM$_{item}$** Mean (s.e.)	**BCIM$_{full}$** Mean (s.e.)
WAIC	34724.9 (118.8)	34350.7 (160.9)	29416.6 (141.1)	26473.8 (132.5)
PWAIC	1852.4 (16.0)	1790.7 (15.6)	1584.0 (16.3)	1509.3 (14.9)
ELPD-WAIC	-18531.5 (59.4)	-17175.3 (80.4)	-14740.8 (70.6)	-13236.9 (66.2)
ELPD	-18532.6 (89.4)	-17200.3 (80.6)	-14775.0 (71.1)	-13279.5 (66.6)
LOOIC	34130.3 (168.8)	34400.5 (161.3)	29549.9 (142.3)	26559.02 (133.1)
R^2/R^2_{Mar}	0.157 / 0	0.192 / < 0.01	0.206 / 0.069	0.206 / 0.080

	1PL	Testlet 1PL	$\mathbf{BCIM_{item}}$	$\mathbf{BCIM_{full}}$
RMSE	0.587	0.586	0.588	0.588

Note. Deviance (-2 times log predictive density) and correctives for parameter fitting using WAIC (using the corrective PWAIC), and leave-one-out cross-validation (LOO-CV) for each of the models fitted to the data. Lower values of WAIC and LOOIC imply higher predictive performance.

The difference in ELPD is much larger than the estimated standard error of the difference, indicating that all the models have relatively good predictive performance, and still BCIM models perform the best. With the 1PL model serving as the baseline model, we see that both BCIM models increase the conditional and marginal R^2(explained variance by fixed effects) than the other two models**.** When we tried to model testlet effects along with context covariates, however, according to the LOO-PIT (Leave-One-Out probability integral transformation) checks there is still some misspecification, and a reasonable guess is that context effects already explain testlet variances enough. Still, applying skew-normal priors or zero-inflated models to the data would be a next-step improvement (we leave that for another case study).

We expect the best selected model has a good model convergence for the MCMC performance. Therefore, we inspect both BCIM models to visually and quantitatively diagnose MCMC performance (i.e., visually check whether the MCMC samples are well mixed and converge to stationary distributions). In Figure 4-5, the trace plots for the group-level parameters have MCMC samples well mixed and converge to target distributions. In the meanwhile, the $\hat{R}$ values from the Gelman-Rubin index (Gelman & Rubin, 1992) mostly close to 1.00 (none was greater than 1.05) for each parameter estimation indicate no problem for BCIM models. The effective sample size statistic $\widehat{n_{\text{eff}}}$ and the Monte Carlo Standard Error (MCSE) values indicated there didn't seem to be a problem with *with-in* sequence correlation. Other convergence diagnostic plots such as the correlation plots, and Geweke Diagnostic plots, provide

further supporting evidence of the approximate convergence of the chains. To check the validity of the model, we have also conducted posterior predictive checking (PPC) to compare our observed data to data generated from the posterior distributions.

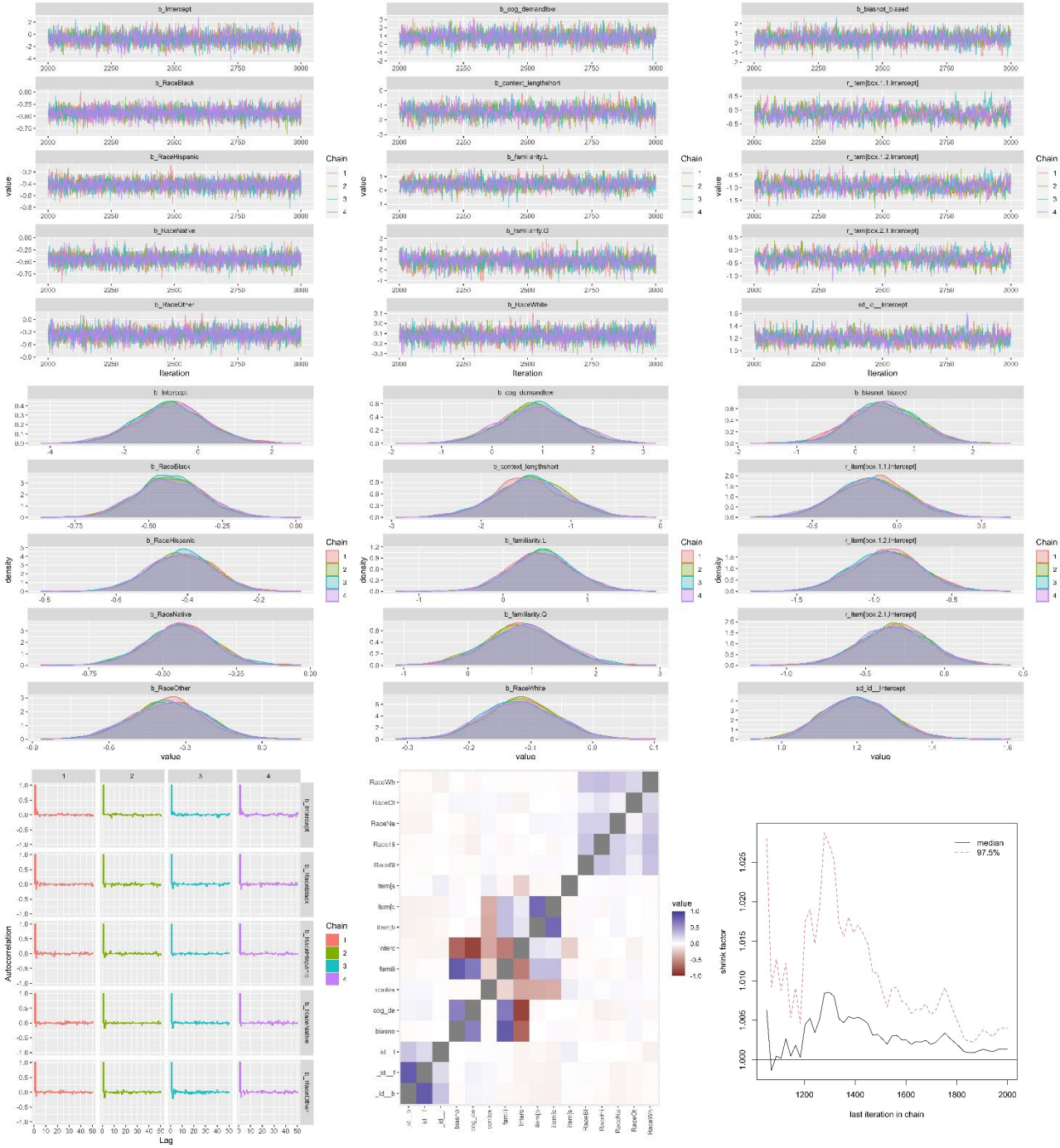


Figure 4-5. Trace plots (upper panels) and density plots (middle panels) for the intercept (b_) parameters of $BCIM_{full}$, and (bottom panels) the autocorrelation, crosscorrelation, and geweke diagnostic plots.

4.6.3 *Race and Context Effects*

The hierarchical structure of BCIM models shares the advantage of "shrinkage" effects (Gelman et al., 2013) in the individual estimates. Shrinkage effects, or partial pooling, allows each individual's estimates to inform the group's estimates, which in turn inform the estimates of all individuals. Estimates from the BCIM models are reported in Table 4-3. For each contextual parameter, we examine the posterior distributions and the uncertainty in the estimates characterized by the highest density interval (HDI). The 95% HDI refers to the span of values that are most credible and cover 95% of the posterior distribution (Kruschke, 2014). For instance, we observe the hypothesized context effect, context length (short), has the coefficient (β = -1.36, *95% CrI* [-2.15, 0.60], *BF* = 399) significantly different from zero at 95% HDI across student groups. Bayes factor (BF) analyzed that with a constructed prior on the (negative) context length effect in reference to the null, the obtained data was almost 399 times more likely to occur given a very strong effect of short context length than a null effect indexed by the reference level. This result echoes findings in Dong (2020) about a concrete context effect.

While the significance of context length holds for when controlling for student racial groups, the associations of other context effects with student success response probability have more uncertainty as we observe larger standard errors relatively, except for the quadratic term of sociolinguistic familiarity added as the incremental effect. From the left column in Table 4-3, we can interpret the ratio (*OR* = 1.40, 95% *HDI* [0.47, 4.33]) of the sociocultural bias effect, for example, that the odds of having a successful response for students is 1.4 times greater with items having an unbiased context than a biased context.

The BCIM$_{full}$ results. Accounting for student characteristics, we summarize the marginal posterior distribution of each parameter obtained with BCIM$_{full}$ in Table 4-3 and Figure 4-6. Each parameter again has the Rhat values close to 1 indicating good convergence. Controlling for the students' racial background, we see that most of the racial groups other than the Asian group have averaged success response probability estimated to be significantly lower than the Asian students. Altogether, the student race and context variables entered in the model roughly explain a 5% additional variance out of the total variances ($\Delta R^2 = .05$). Context effects alone increase explained variance in fixed effects by 7%.

Table 4-3

Odds Ratios and 95% Credible Intervals of Model Coefficient Estimates

	BCIM$_{context}$		**BCIM$_{full}$**	
Predictors	***Odds Ratios***	***CI (95%)***	***Odds Ratios***	***CI (95%)***
Intercept	0.50	0.08 - 2.87	0.48	0.07 - 3.08
Context length: short	0.26	0.12 - 0.55	0.65	0.52 - 0.81
Cognitive demand: low	1.87	0.47 - 7.60	0.66	0.55 - 0.78
Familiarity (Linear)	1.41	0.67 - 3.01	0.64	0.51 - 0.80
Familiarity (Quadr)	3.14	1.11 - 8.78	0.69	0.52 - 0.91
Bias: not-biased	1.40	0.47 - 4.33	0.89	0.79 - 1.00
Race: Black			0.23	0.10 - 0.52
Race: Hispanic			2.41	0.59 - 9.66
Race: Native			1.58	0.78 - 3.34
Race: Other			2.37	0.75 - 7.09
Race: White			1.6	0.54 - 5.12
Random Effects				
σ^2	3.29		3.29	
τ_{00} id	1.03		1.45	
τ_{00} item	0.49		0.53	
τ_{11} id. Not-biased	4.26		1.48	
τ_{11} id. Familiarity-L	2.18		0.78	
τ_{11} id. Familiarity-Q	3.10		2.38	
τ_{11} item. Race-Black			0.09	
τ_{11} item. Race-Hispanic			0.08	
τ_{11} item. Race-Native			0.05	
τ_{11} item. Race-Other			0.03	
τ_{11} item. Race-White			0.01	
ICC	0.32		0.31	

Note. $N_{item} = 32$, $N_{id} = 1451$. N of total observations is 25931. In $BCIM_{full}$ specification, the reference group is Asian students.

Since our interest is on item-level variances, we plot the direction of probabilities in the estimated coefficients and observe they are close to the context-only model. The overall associations of sociocognitive and sociocultural context effects and student performance is positive. Bayesian hypothesis testing shows that students tend to a have higher success probability with items that have longer contexts than shorter contexts (probability exceeds 95%). We have a moderately strong evidence (*Evi. Ratio* = 8.28, *P* = 89%) that a low cognitive demand context tends to support slightly more successful answers to an item than a high demand context, controlling for other factors. There is also moderate evidence that contexts with sociolinguistic familiarity tend to support student's success in responses than ones with sociolinguistic unfamiliarity (*Evi. Ratio* = 8.41, *P* = 89%). Similarly, contexts without a cultural bias are more attributable to success response probabilities than the opposite (*Evi. Ratio* = 3.91, *P* = 80%), with evidence still being moderately strong.

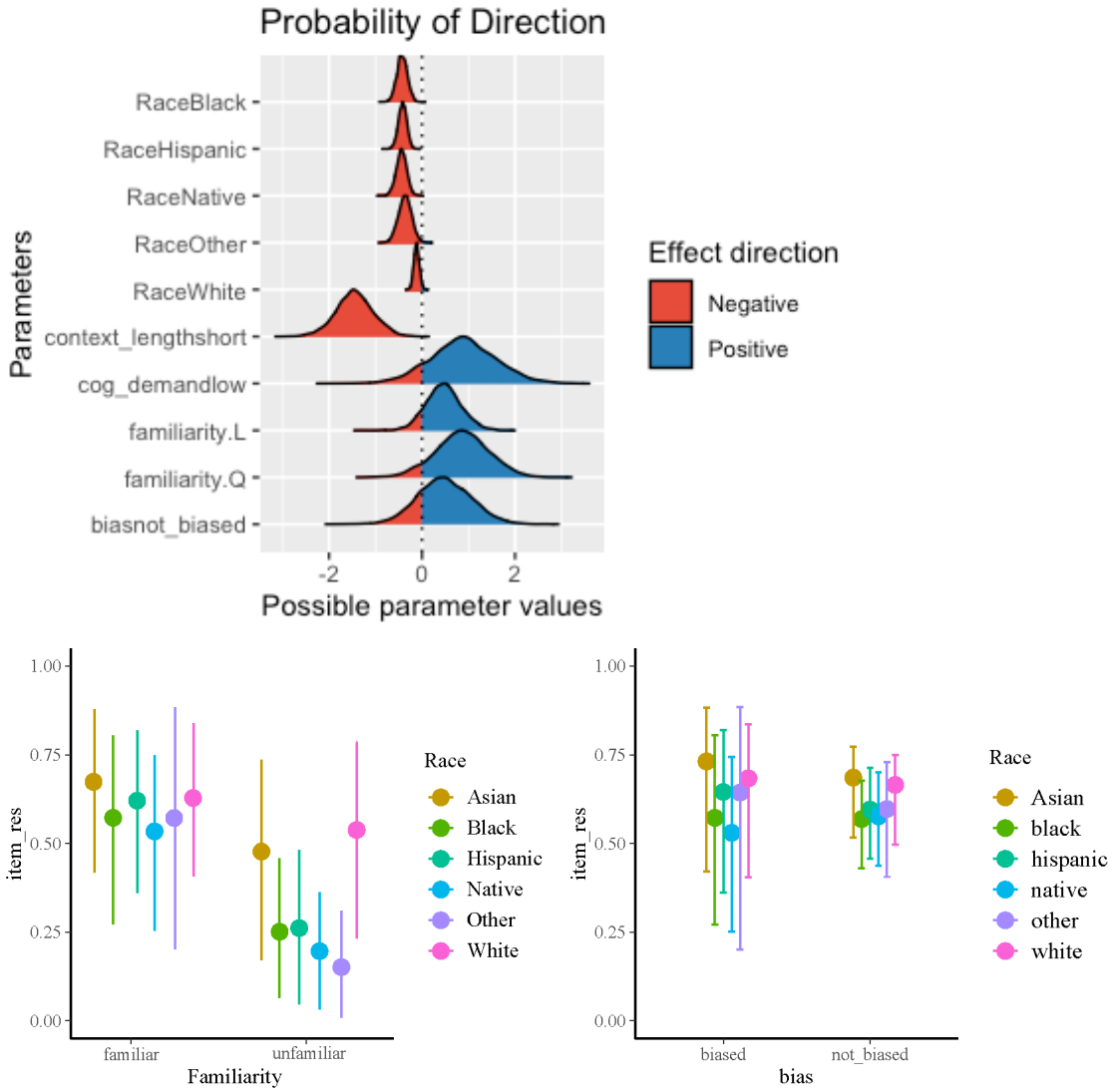


Figure 4-6. Plot of probability of direction (upper panel) and plots of conditional context effects on predicted success response probability (lower panel).

To examine the difference in a particular context parameter between racial subgroups, we visualize the predicted response probability with 95% Credible Intervals conditional on student groups. Across the levels of the sociolinguistic familiarity context category, the posterior mean values seem to be lower for unfamiliar contexts across groups, with the difference more distinctly for Black, Hispanic, Native, and other racial minority groups. Lower predicted success probability in smaller variations is more associated with unfamiliar context for other racial minority groups than for the Asian and White groups. There also seems to be a weak interaction effect between

sociocultural bias of context and student's racial background. Black and Native students tend to perform better with unbiased contexts, while the other groups may not show increased effects if not the opposite effects. The predicted probabilities of success tend to have lower variation in uncertainty (standard errors) for unbiased sociocultural context than for biased sociocultural context.

Though the above conditional effects are straightforward to visualize, it cannot prove the causal link between the represented context effects and the propensity of giving a successful response. For example, we observe that testlet Bus and testlet Sled share similar patterns of context codes about the four variables of interest at the subtestlet level. Thus, the discrepancy of item parameters between Box 2.2 and Cart 2.2 might be due to other factors. For another useful example, more variations of subtestlet context codes are present within the Bus and Sled testlets. Here we observe for the second subtestlet that the Sled subtestlet has more problematic codes in the two categories of sociolinguistic familiarity and sociocultural bias than the Bus subtestlet. Unsurprisingly, we see that in Figure 4-7, among the plotted posterior item parameters (random effects with 95% CrI) between the Bus and Sled, more *item difficulty* is generated in Sled 2.1 and 2.2 for the posterior mean difficulty than Bus 2.1 and 2.2.

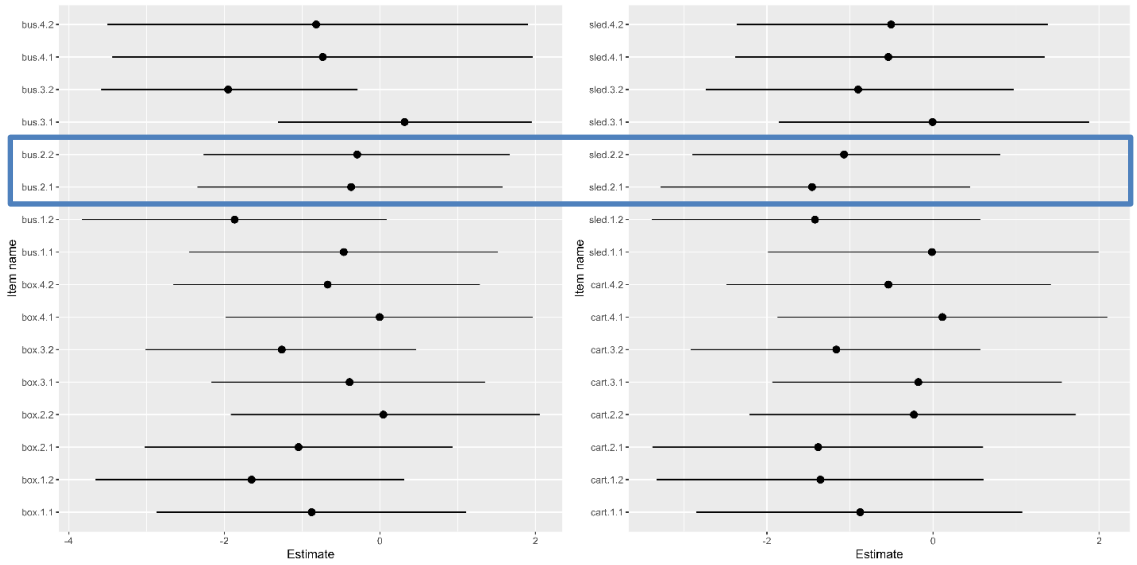


Figure 4-7. Item easiness parameters as random effects with 95% CrI for parallel items within the testlets.

4.6.4 *Context Predictors Using Factor Scores*

To capture all other contextual features in our context system, we further model predictors from loadings of factor analysis of context codes. Dropping codes exhibiting no variation, we retain a final total of 15 testlet context features, 26 subtestlet context features, and 21 item context features (feature categories may overlap) with codes developed in the system to capture variations of contexts in our sample of 32 physics and motion items. For each observation, item-level dimension loadings from Multiple Correspondence Analysis (MCA) scores are obtained. Figure 4-8 shows the first two dimensions explain most the variance, but we retain the first five (explaining 56.4% of the total variance) to capture other sources of physics-related information.

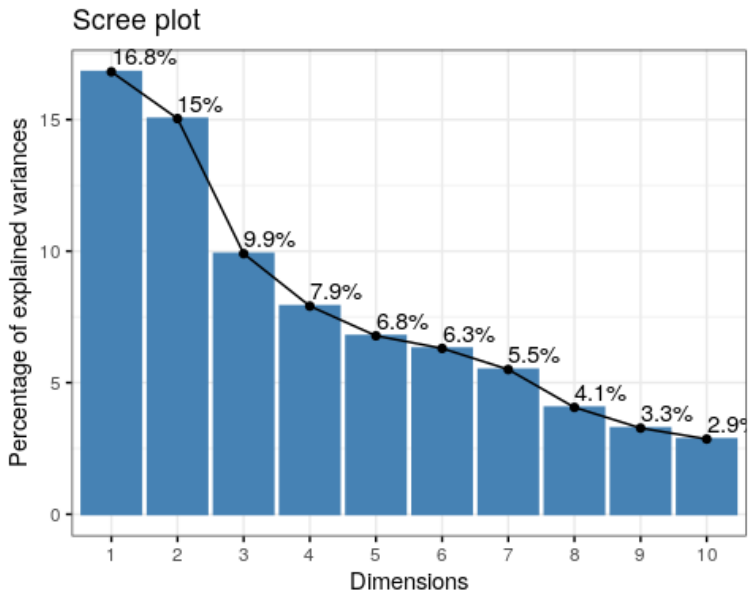


Figure 4-8. Scree plot for multiple correspondence analysis.

In summary, the analysis in Table 4-4 reveals a very strong negative association between both Dimension 1 and Dimension 2 and students' success response probability. Both BFs tend to surge to infinitely large numbers. Compared to the previous $BCIM_{full}$ model, using factor scores from the item level explains an additional 3.7% variance

from the entire model, while the dimensional predictor approach only increases 0.8% variances by the fixed effects. The BFs reveal no evidence for an association between either of Dimension 3, 4, and 5 and student performance (*Evi. Ratio* in the range of [1.47, 2.38], *P* from [0.6-0.7]).

Table 4-4

Estimates from the BCIM Model Using Factor Scores

Parameter	Estimate	SE	Lower B.	Upper B.	$\widehat{R}$
Intercept	-1.11	0.37	-1.88	-0.4	1
Race: Black & Native	-0.39	0.09	-0.58	-0.21	1
Race: Hispanic	-0.33	0.08	-0.5	-0.18	1
Race: Other	-0.36	0.09	-0.54	-0.19	1
Race: White	-0.09	0.06	-0.21	0.05	1
Dimension 1	-2.4	0.33	-3.06	-1.79	1.01
Dimension 2	-2.62	0.66	-3.98	-1.35	1.01
Dimension 3	0.75	1.66	-2.69	3.89	1
Dimension 4	0.44	0.83	-1.15	2.11	1
Dimension 5	0.24	1.16	-2.12	2.44	1

Note. Posterior mean, standard error, 95% credible interval and $\widehat{R}$ statistic for fixed effects are reported. ICC = 0.61, R^2/R^2_{Mar}= 0.237/ 0.088, N_{item} = 32, $N_{student}$ = 1451.

The interpretability of Dimension 1 and 2 are well aligned with our hypothesis of sociocognitive and sociocultural context effects as well. We found that items that have high factor scores on Dimension 1 (i.e., problematic items) tend to have more technical language to reflect sociolinguistic unfamiliarity at the item level. At the subtestlet and item levels, those contexts also have higher numbers of pieces of information to process physics concepts, which is an indirect link of a higher level of cognitive demand. Thus, we can say that Dimension 1 reflects items that are likely to be culturally or linguistically biased, and requires a student's high cognitive demand. Interestingly, those contexts describe the motion topic without explicitly describing any forces. On the contrary, (unproblematic) items that are low on Dimension 1 tend to have formal or common, rather than technical, language. They also have longer context length, but unlikely to activate irrelevant information.

In a similar fashion, problematic items that have high loadings on Dimension 2 tend to have shorter item-level contexts with only one-piece of information. Typically, in such a way they tend not to have unfamiliar language, nor biased context (even no context). Those contexts also tend to be very focused, with usually all forces specified but with unspecified motion. Yet, they lose the advantage of contextualization, as we can see from the model results, which is a highly significant factor for student performance.

4.7 Discussion

Research has shown that variance in item difficulty estimates belonging to the same content construct leads to construct irrelevant variance and results in a less accurate estimation of a student's latent proficiency (e.g. Ruiz-Primo and Li, 2016). Our experimental design of booklet items and the Bayesian approach to modeling item context's sociocognitive and sociocultural effects provides an important investigation into such issues. The BCIM results illustrate the effects of sociocognitive and sociocultural context in explaining the construct-irrelevant variances in the observed item responses. This approach provides comparable scores for respondents regardless of the differentiated construction of content contextualization. With the sampling design of booklets and respondents in large-scale educational assessment, we expect that our contextualized item response model can mitigate the constraint of matrix sampling of testlets in booklet formation by applying a common context categorization framework. Moreover, we can compare and evaluate conditional standard errors (CSEMs) across similarly designed items, and with a Bayesian approach, the CSEMs are usually reduced for respondents with lower proficiency on the scale, than the traditional standard error of measurement point estimate. Consequently, the reliability of respondent scores will be improved. Significant context effects are indirect evidence for score validity and

inferences on fairness and equity claims. Therefore, items of problematic contexts may pose a threat to their test validity.

In this study, we have replicated and extended previous findings showing that contexts with its presence as a beneficial effect, along with sociocognitive and sociocultural attributes are associated with students' cognitive performance on test items. More precisely, we disentangled contextual correlates of sociocognitive and sociocultural dimensions by testlet, sub-testlet, and item levels, and examined their effects on student performance. At the testlet level, we described variances associated with testlets and how the four testlets differ in student distribution. Followed by the Bayesian contextualized item response modeling, we demonstrate the direct positive effects of the key variables of context length (an indicator of context's constructive role), cognitive demand, sociolinguistic familiarity, and cultural bias. The results confirm the output from a two-step regression analysis on item difficulties (See Appendix 2.B), and our selection of those four variables is ranked high on their regulated feature importance. Both modeling results tend to be more sensitive to problematic labeled context codes, and detect negative context factors that are detrimental to students' cognitive performance.

While our definition of context familiarity pertains to the sociolinguistic hurdles, there are other approaches to understanding context familiarity that can be possibly modeled. Our definition of sociolinguistic familiarity has already included a condition where cognitive and linguistic familiarity may happen with the constant engagement of instructional efforts and differentiates itself from other technical or unfamiliar language coupled with far proximity of application. However, context familiarity due to the nature of context may have mixed results. Previous studies in Dong (2020) have looked at context familiarity based on the ones derived from a classroom setting and

experiences, and from other daily life experiences. Their results tend to indicate a hypothesis that familiar language might be mediated by students' close proximity to applying the context (possibly an indicator of context applicability). Future modeling for disentanglement of the two concepts is needed for item and context measurement.

At the same time, this study reveals that the relationship between students' success in item response and context familiarity or bias varies by racial subgroups. With a breakdown of context features, we have gained detailed sources of construct-irrelevant variance that are related to sociocultural and sociocognitive features, and the effects are conditional on students' cultural backgrounds to a variety of degrees. We have some plausible evidence that Black and Native American student groups may benefit more from culturally unbiased contexts represented during a test than other racial subgroups. It is possible that Asian and White student groups tend to be less discriminable by certain categories of sociocognitive and sociocultural construction of contexts than other minority groups. However, it is by no means that we forsake or ignore those elements of context features, as the benefits of context should play an important role in promoting students' testing experiences and cognitive performance, and maybe more significant than flagging the problematic elements of contexts. Since our context system generates an improvement in adjusted R^2, there could possibly be other variables explaining and mediating through each other and have interaction effects with student groups. Future studies could add more predictors describing physics knowledge and the nature of topics.

We have so far investigated models only under the Rasch framework. Further studies can explore other models, such as the two-parameter models by incorporating the item discrimination parameter. We have also investigated context effects by considering there is no measurement error in context codes. This assumption can be

easily violated in reality. Bayesian models accommodating measurement errors in variables can be implemented. It is possible that the estimated relationships between context effects and response performance may be spurious if there's a neglect of other non-context effects and interaction effects with context effects in our specification.

This study serves as an exemplary analysis of contextualized item response models. However, further inclusion of more context variables is still welcome and might be much needed. We notice the computational cost of using Bayesian modeling is sometimes high. But overall, the Bayesian approach sophistically handles the high dimensional models when multilevel models estimated in frequentist's methods encounter convergence problems or have model singular fit. In fact, fitting the most complex models using the maximal effects turns out to be more robust in the Bayesian approach. In addition, adding other auxiliary process information in future studies, such as response times, would possibly improve the quality of item and respondent parameter estimates as well.

Reference

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

Atkinson, D., Churchill, E., Nishino, T. & Okada, H. (2007). Alignment and Interaction in a Sociocognitive Approach to Second Language Acquisition. *The Modern Language Journal*, 91: 169-188. https://doi.org/10.1111/j.1540-4781.2007.00539.x

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00328

Bürkner, P. C. (2018). Advanced Bayesian multilevel modeling with the R package brms." *The R Journal*, 10(1), 395-411. doi:10.32614/RJ-2018-017.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan." *Journal of* Statistical *Software*, 80(1), 1-28. doi:10.18637/jss.v080.i01.

Carjuzaa, J., & Ruff, W. G. (2010). When Western epistemology and an Indigenous worldview meet: Culturally responsive assessment in practice. *Journal of Scholarship of Teaching and Learning*, *10*(1), 68-79.

Choi, I., & Wilson, M. (2014). Multidimensional classification of examinees using the mixture random weights linear logistic test model. *Educational and Psychological Measurement*, *75*(1), 78-101. https://doi.org/10.1177/0013164414522124

Cooper, B., & Dunne, M. (2000). Constructing the "legitimate' goal of a 'realistic' math item: A comparison of 10-11 and 13-14 year-olds. In A. Filer (Ed.). *Assessment: Social practice and social product* (pp. 87-109). New York, NY: Routledge.

Cooper, B., & Harries, T. (2002). Children's responses to contrasting realistic mathematics problems: Just how realistic are children ready to be?. *Educational Studies in Mathematics*, *49*(1), 1-23.

Crisp, V., & Grayson, R. (2013). Modelling question difficulty in an A level physics examination. *Research Papers in Education*, *28* (3), 346-372.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer

function from the lme4 package in R. *Journal of Statistical Software*, 39(12), 1-28. doi:10.18637/jss.v039.i12.

Dong, D. (2020). *What do we know about context: An integrated analysis of context characteristics of science assessment item*, (Doctoral dissertation, University of Washington, Seattle, WA). ProQuest Dissertations and Theses. Retrieved from http://hdl.handle.net/1773/45482.

Fox, J. (2010). *Bayesian item response modeling: Theory and applications*. NY: Springer Science & Business Media.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). FL: Chapman & Hall/CRC Press.

Gelman, A., Hwang, J., & Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997-1016. https://doi.org/10.1007/s11222-013-9416-2

Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/Hierarchical models. https://doi.org/10.1017/cbo9780511790942

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4). https://doi.org/10.1214/ss/1177011136

Greiff, S., Wüstenberg, S., Molnar, G., Fischer, A., Funke, J., & Csapo, B. (2013). Complex problem solving in educational settings - something beyond g: concept, assessment, measurement invariance, and construct validity. *J. Edu. Psychol.* 105, 364-379. doi: 10.1037/a0031856

Gronau, Q. F., Singmann, H., & Wagenmakers, E. (2017). Bridgesampling: An R package for estimating normalizing constants. *arXiv preprint arXiv:1710.08162*.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *The Journal of Machine Learning Research*, 15(1), 1593-1623.

Kruschke, J. K. (*2014*). *Doing Bayesian data analysis*: *A tutorial with R, JAGS, and. Stan*. 2nd Edition. Academic Press.

Lee, J. (2011). *Second language reading topic familiarity and test score: Test-taking strategies for multiple-choice comprehension questions*, ProQuest Dissertations and Theses.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989-2001. https://doi.org/10.1016/j.jmva.2009.04.008

Levy, R., & Mislevy, R. J. (2017). *Bayesian psychometric modeling*. FL: CRC Press.

Li, M., Ruiz-Primo, M., Wills, K., & Giamellaro M. (2012b). *Instructionally sensitive assessments across science modules*. Paper Presented at the American Educational Research Association (AERA) Annual Meeting, Vancouver, Canada

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Mislevy, R. J., & Verhelst, N. (1990). Modelling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*(2), 195-215. https://doi.org/10.1007/bf02295283

Morrison, K. M., & Embretson, S. E. (2014). Using cognitive complexity to measure the psychometric properties of mathematics assessment items. *Multivariate Behavioral Research, 49*(3), 292-293.

Muthukrishnan, R., & Rohini, R. (2016, October). LASSO: a feature selection technique in predictive modeling for machine learning. In *2016 IEEE international conference on advances in computer applications (ICACA)* (pp. 18-20). IEEE.

OECD (2017). PISA 2015 Science Framework, in *PISA 2015 assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving*, OECD Publishing, Paris, https://doi.org/10.1787/9789264281820-3-en.

Ruiz-Primo, M. A., & Li, M. (2015). The Relationship between Item Context Characteristics and Student Performance: The Case of the 2006 and 2009 PISA Science Items. *Teachers College Record: The Voice of Scholarship in Education*, *117*(1), 1–36. https://doi.org/10.1177/016146811511700118

Ruiz-Primo, M.A., & Li, M. (2016). PISA science contextualized items: the link between the cognitive demands and context characteristics of the items. *RELIEVE*, *22*(1), art. M11.

Ruiz-Primo, M., Li, M., Minstrell, J., Kanopka K., Hernandez, P., Dong, D., & Zhai, X. (2019). *Testing the generalization to the domain inference: The use of*

contextualized clusters of items. Paper Presented at the National Council on Measurement in Education (NCME) Annual Meeting, Toronto, Canada.

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 424-451. https://doi.org/10.1207/s15328007sem1103_7

Schneider, M. C., Huff, K. L., Egan, K. L., Gaines, M. L., & Ferrara, S. (2013). Relationships among item cognitive complexity, contextual demands, and item difficulty: Implications for achievement-level descriptors. *Educational Assessment*, *18*(2), 99-121.

Song, M., & Bruning, R. (2016). Exploring effects of background context familiarity and signaling on comprehension, recall, and cognitive Load. *Educational Psychology, 36*(4), 691-718.

Stan Development Team (2019). *Stan modeling language: User's guide and reference Manual*. URL http://mc-stan.org/manual.html.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing. 27(5)*, 1413-1432. doi:10.1007/s11222-016-9696-4 (journal version, preprint arXiv:1507.04544).

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., & Gabry, J. (2019). Pareto smoothed importance sampling. *preprint arXiv*:1507.02646

Vos, V. (2014). *The use of context in Science education*. Retrieved from http://dspace.library.uu.nl/bitstream/handle/1874/297294/The%20Use%20of%20Context%20in%20Science%20Education.pdf?sequence=2

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. *Computerized Adaptive Testing: Theory and Practice*, 245-269. https://doi.org/10.1007/0-306-47531-6_13

Wang, N., Li, M., & Dong, D. (2019, April). *Measuring equitable contextualized items in science assessment.* Paper presented at the Annual Meeting of the American Educational Research Association (AERA). Toronto, Canada.

Wang, N., (2019, April). *A culturally responsive approach to international science assessment.* Paper presented at the 63rd Annual Conference of the Comparative and International Education Society (CIES), San Francisco, CA.

Wang, T., Li, M., Thummaphan, P., & Ruiz-Primo, M.A. (2017). The effect of sequential cues of item contexts in science assessment. *International Journal of Testing*, 17:4, 322-350, DOI: 10.1080/15305058.2017.1297818

Chapter 5. Exploring Differential Item Functioning in Cognitive Diagnostic Computer Adaptive Testing: A Simulation Study

Nixi Wang, Chun Wang

College of Education, University of Washington, USA

Abstract

In innovative designs of modern assessment architecture, item features can be incorporated in cognitive diagnostic purposes. In this study, we study item bias through the lens of applying the Cognitive Diagnostic Model (CDM) in the Computer Adaptive Testing setting. Differential item functioning (DIF) in the context of cognitive diagnostic computer adaptive testing (CD-CAT) has not yet been systematically studied. To examine a test's validity argument and test fairness, DIF under various conditions of CD-CAT is important to investigate. We evaluate the performance of Wald test for both uniform and nonuniform DIF recovery under various conditions and procedures of CD-CAT through a simulation study. Type 1 error rates tend to be below and close to the nominal level across conditions. For empirical power, the performance is differentiated from low to high mainly based on the choice of item selection algorithm, test-, and item- condition. Furthermore, classification accuracy and test security under those conditions are evaluated. Discussion includes practical implications, limitations of the study, and future directions.

5.1 Introduction

In recent years, educational and psychometric measurement has gained great advances in computer-based technology and automated algorithms. One direction is the application of cognitive diagnostic models (CDM) being incorporated in the computer adaptive testing (CAT) setting. As evinced in many of its strengths, the administration and design of CAT embrace more responsiveness and flexibility than the paper-pencil test with a prefixed set of items. Integrated into its architecture is an optimal selecting scheme of administering an item into a test, on the fly, that quickly locates the area and range of an examinee's latent ability rather than a whole latent continuum. As such, adaptive testing can operationalize tailoring to the trait of each individual test-taker, and deliver computationally efficient and accurate results (Weiss, 1982).

On the other hand, the CDM-based assessments aim to measure respondents with fine-grained information on knowledge constructs or components. Based on the combination of knowledge components as required in each item, CDM estimates a set of skills inherently designed in each item and the mastery status of an examinee on each of the skill attributes. Compared to a single score in traditional testing, a diagnostic test provides a profile of mastery of skills, which gives students personalized evaluation or recommendations for improvement. Student test-takers can benefit from such assessments as they receive feedback on the concepts of subject matter and areas that he/she needs to work on (Cheng, 2009). Thus, both the CDM framework and CAT-enabled technology have great potential for creating learner-centered assessments that promote learning.

Cognitive diagnostic computer adaptive testing (CD-CAT) advances educational measurement by building on the combination of strengths from CAT and CDM. It not only provides fine-grained student-diagnostic information from a design matrix (Q-

matrix) as item attributes, but also through an adaptive algorithm, simultaneously performs quick and efficient estimations of each examinee's latent trait. One practical application, for example, is that after a lesson in the class, students complete a short formative assessment using classroom computers while they are still fresh in the memory for curriculum materials. Then, diagnostic scores could be immediately generated as student profiles highlighting persons' strengths and weaknesses of each learning concept. Figure 5-1 illustrates a case of Q-matrix, which is a 1-0 matrix mapping skills required by each item, whereas the alpha-matrix gives an evaluation of attribute mastery in a vector by each person. In recent years, the CD-CAT is becoming a practical, powerful, and convenient measurement tool, which yields substantial advantages regarding information efficiency (Huebner, 2015).

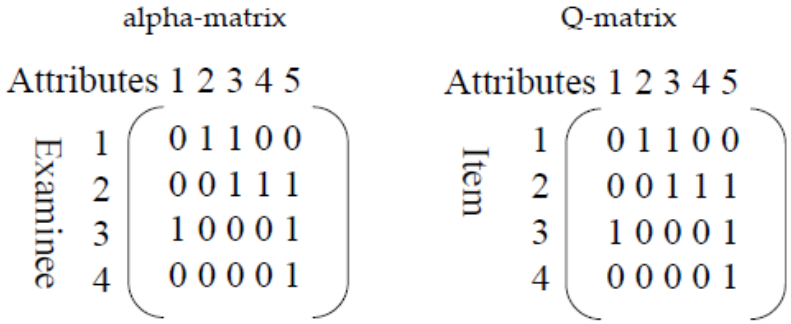


Figure 5-1. An illustration of person attributes (Alpha-Matrix) and item attributes (Q-Matrix).

CD-CAT research focuses much attention on optimizing item selection methods for different considerations (Cheng, 2009, 2010; Kaplan, de la Torre, & Barrada, 2015; Wang, 2013; Xu, Chang, & Douglas, 2016). Studies are fast developing for more comprehensive and systematic reviews on its psychometric techniques. Issues such as DIF could lead to increased measurement error and invalid estimates for examinees from different groups (Lord, 1980). While the discussion of DIF issues in CD-CAT is forthcoming, there are several studies that have explored the application of DIF methods

in either CDMs (Li, 2008; Hou, la Torre, & Nandakumar, 2014; Zhang, 2006) or in CAT setting (Lei et al., 2006). For instance, Feng (2004) and Piromsombat (2014) found that CAT could effectively recover from biased estimates obtained in the early stages of CAT. Most of the DIF studies focus on the calibration process and the IRT parameterization of items, whereas in the CD-CAT, the latent space is a vector of discrete classes.

Items in the CD-CAT exhibiting DIF are conditioned on the attribute mastery profiles for different groups. Little is known about how and to what extent CD-CAT can self-adjust to DIF presence and types, and what DIF detection methods can be effective in stages of CD-CAT. Extant research has not explored aspects of DIF such as the DIF magnitude, impact, or location in CD-CAT. To achieve that, we review relevant literature in the sections below, and conduct a simulation study that seeks to answer the following research questions:

(1) What is the performance of the Wald test for DIF recovery under the simulated conditions of a CD-CAT setting?

(2) How do factors of CD-CAT including sample size, DIF magnitude, and item-selection algorithm affect the performance of detecting DIF items?

(3) How do DIF conditions affect item usage and estimation in terms of attribute and pattern classification accuracy?

5.2 Background

The popular deterministic input, noisy, “and” gate model (DINA; Junker and Sijtsma, 2001), along with the noisy inputs, deterministic “and” gate (NIDA) models (de la Torre & Douglas, 2004; Junker & Sijtsma, 2001), the reparameterized unified/fusion model (RUM: Hartz, 2002), the compensatory deterministic input, noisy ‘or’ gate (DINO) and noisy inputs, deterministic ‘or’ gate (NIDO) models (Templin & Henson, 2006), and

generalized DINA (G-DINA) models (de la Torre, 2011) are in the CDM family for estimating examinees' discrete latent attributes based on their item responses. Item-attribute relationships can be described by a Q-matrix. Each examinee's attribute profile is defined by a vector α = $(a_1, \ldots, a_K)$ of *K* skills. If an item pool has *J* items, the Q-matrix (a J x K matrix) with binary entries zero and one can specify whether the item structure measures each attribute *k* or not. Given an examinee's attribute profile *α*, the response X_j to item *j* under the corresponding model falls under a Bernoulli distribution. For the DINA model, an examinee needs to master all attributes required by an item in order to get a correct response. The probability of getting an item correct follows a conjunctive rule as below:

$$P(X_j = x|\alpha) = \{P(X_j = 1|\alpha)\}^x \{P(X_j = 0|\alpha)\}^{1-x}, \; x = 0,1. \tag{1}$$

where we have,

$$P(X_j = 1|\alpha) = \left(1 - S_j\right)^{\delta_{j;a}} {g_j}^{1-\delta_{j;a}}, \tag{2}$$

$$\delta_{j;a} = \prod_{k=1}^{k} (a_k)^{q_{jk}} = 1 = (a_k \geq q_k \; for \; all \; k). \tag{3}$$

For each item, there are two parameters S_j and g_j that are known as the slipping and guessing parameters in CDM. $\delta_{j;a}$ = 1, meaning the respondent is capable of mastering all the required attributes in solving a problem, is the only condition to have a positive response while the guessing parameter vanishes to zero, when in that case the probability $P(X_j = 1|\alpha)$ is 1 - S_j; and vice versa the $P(X_j = 0|\alpha) = g_j$. Using a Bayesian framework, the RUM model parameterizes the item parameters at the person attribute level just as the NIDA model. It defines the probability of a correct response as follows:

$$P(X_{ij} = 1|a_k, S, g) = \prod_{k=1}^{k} [\left(1 - S_j\right)^{\delta_{j;a}} {g_j}^{1-\delta_{j;a}}]^{q_{jk}}. \tag{4}$$

To increase the identifiability of the unified model, Hartz (2002) reparametrized Equation 4 to add two newly transformed parameters that combine the guessing and slipping: 1) π_j^* as the probability of answering item *j* correctly, given that the examinee has mastered all the attributes under the conjunctive rule; and 2) r_{jk}^* as a "penalty" parameter for not mastering an attribute and is expressed as the ratio of the probability of answering item *j* correctly given non-mastery to the probability of answering item *j* correctly given mastery: $r_{jk}^* = \frac{P(X_{ijk}=1|a_{ik}=0)}{P(X_{ijk}=1|a_{ik}=1)} = \frac{g_{jk}}{1-s_{jk}}$. Therefore, the reduced version of RUM (Roussos et al., 2007) can be rewritten as the following and omitting the supplemental residual ability:

$$P(X_{ij} = 1|a_k) = \pi_j^* \sum_{k=1}^{K} r_{jk}^{*(1-a_{ik})q_{jk}}. \quad (5)$$

Estimation methods such as EM algorithm in implementing marginalized maximum likelihood estimation (MMLE) for the DINA model, or MCMC algorithm for the reduced RUM model can be used to obtain response scores and person attribute estimates. Once the precalibrated item bank is set, criteria concerning the next item selection methods in relation to its optimized psychometric properties have been proposed as many. Due to the discreteness of the attribute space of CDMs, the standard CAT information selection method developed for IRT, like the Fisher information cannot be directly applied to the person's discrete latent trait spaces. Proposed new item selection indexes in the literature include but are not limited to the Kullback-Leibler (KL)-based global discrimination index (GDI), the Shannon entropy procedure (Xu, Chang, & Douglas, 2003), the posterior-weighted KL information (PWKL) (Cheng, 2009), the mutual information method (Wang, 2013), the restrictive stochastic item selection (Wang et al., 2011), and the progressive control algorithm (Lin and Chang, 2018).

The KL information (Cover & Thomas, 1991) measures the divergence between two probability distributions, in the CD-CAT application, the conditional probability distribution of a person *i*'s attribute patterns on item *j*, $f(X_{ij}|\widehat{\boldsymbol{a}}_l)$, given the current estimated latent class, and $g(X_{ij}|a_c)$, which is the probability distribution of that person *i*'s attribute patterns on item *j*, given the true theta. The log likelihood ratio form became:

$$KL_j(\widehat{\boldsymbol{a}}_l||\boldsymbol{a}_c) = \sum_{X=0}^{1} \log\left(\frac{P(X_{ij}=x|\widehat{a}_l)}{P(X_{ij}=x|a_c)}\right) P(X_{ij}=x|\widehat{a}_l). \quad (6)$$

Under Cheng's (2009) specification, the posterior probability of person *i* given the responses to the *t* items is denoted as $\pi^{(t)}(\boldsymbol{a}_c)$. The posterior distribution after the *t*th response can be written as $\pi^{(t)}(\boldsymbol{a}_c) \propto \pi^{(0)}(a_c)L\left(\boldsymbol{X}_i^{(t)}\middle|\boldsymbol{a}_c\right)$, where $\boldsymbol{X}_i^{(t)}$is the response vector, $\pi^{(0)}(\boldsymbol{a}_c)$ is the prior and $L\left(\boldsymbol{X}_i^{(t)}\middle|\boldsymbol{a}_c\right)$ is the likelihood of the observed responses given the attribute pattern vector. The PWKL maximizes the index by using the posterior distribution of the attribute vectors as weights:

$$PWKL_j(\widehat{\boldsymbol{a}}_i^{(t)}) = \sum_{c=1}^{2^K}\left[\sum_{X=0}^{1} \log\left(\frac{P(X_j=x\middle|\widehat{\boldsymbol{a}}_i^{(t)})}{P(X_j=x|\boldsymbol{a}_c)}\right) P\left(X_j=x\middle|\widehat{\boldsymbol{a}}_i^{(t)}\right)\pi^{(t)}(\boldsymbol{a}_c)\right]. \quad (7)$$

In this study, Kaplan et al.'s (2015) KL-based global-discrimination index (GDI) can be a useful rule for selecting the next item, due to its reduced structure and relatively sound efficiency. The GDI for item j can be calculated as:

$$\xi_j^2 = \sum_{c=1}^{2^{k_j^*}} \pi^{(t)}\left(a_{cj}^*\right)\{P(X_j=1\middle|a_{cj}^*) - \bar{P}_j^{(t)}\}^2 . \quad (8)$$

where $\bar{P}_j^{(t)} = \sum_{c=1}^{2^{k_j^*}} \pi^{(t)}\left(a_{cj}^*\right)\{P(X_j=1\middle|a_{cj}^*)$, and k_j^* is the first k_j^* attributes required in item *j*. So the reduced attribute profile a_{cj}^* has latent attribute patterns c = 1… $2^{k_j^*}$.

Also, for each step t for item j, $\pi^{(t)}$ is the posterior distribution on a reduced attribute pattern where item j has k_j^* attributes required.

5.2.1 *Differential Item Functioning*

There are a few studies that have investigated DIF in the context of CDMs (Hou et al., 2014; Li, 2008; Li & Wang, 2015; Svetina et al., 2018; Paulsen et al., 2020) that focus on either item-level or attribute-level DIF. In the DINA framework, uniform DIF is introduced if the probabilities of correct responses are consistently higher/lower for one group across all required attribute profiles. Nonuniform DIF is usually defined if the probabilities of correctly answering the item are lower for one group on some latent attribute profiles but higher for the same group on some other latent attribute profiles. Paulsen and his colleagues (2020) investigated the impact of various DIF conditions in CDM on attribute and profile classification accuracy. They found that attribute-level classification accuracy is mostly robust to DIF of large magnitudes in various conditions, while profile classification accuracy is negatively influenced by the presence of DIF.

Some studies investigated various DIF estimation methods using conditional variables such as overall score and latent attribute profiles, with the latter yielding better type I error and power rates. Hou et al., (2014) examined Wald test (Wald, 1943; applied to the DINA setting based on the information matrix estimation method developed by de la Torre and Lee (2013). The proposed Wald test performed well in detecting DIF when items are relatively discriminating and DIF size is large, based on parameter estimations from item-wise information matrix, but it introduced inflated Type I error rates when items poorly discriminate. In their study, Li and Wang (2015) used a multiple group CDM approach to model group differences through log-linear cognitive diagnosis model (LCDM) DIF method. It was compared with the Wald

method for two and three groups using the MCMC algorithm, and concluded LCDM-DIF was comparable to Wald method in most of the conditions. For the three-group conditions, the power of the Wald method was slightly better than that of the LCDM-DIF, but the latter produced lower Type I error rates. LCDM-DIF also outperformed Wald when there's a high chance to slip or guess items. Svetina et al. (2018) evaluated the impact of Q-matrix misspecification on the performance of logistic regression, Mantel-Haenszel, and Wald for detecting DIF in CDMs. They found that all were affected by Q-matrix misspecification, specifically, the Type I error rate control of LR and MH was better than that of Wald; but LR and Wald had greater power than MH. Alternatively, in Li (2008)'s study, it is found a higher order (HO) DINA model could simultaneously account for DIF and DAF through the lower and higher level models, respectively, producing better Type I error and power rates than the MH method. However, overall those studies observed a large range of power rates, ranging from extremely (unacceptably) low rates of 0.10s to high power rates of 0.90s.

In Hou et al. (2014) and Li and Wang (2015), item parameter estimates are derived from the CDM model separately, and they are used to calculate the Wald statistic, which tests the null hypothesis that group parameters are equal. The statistic is distributed asymptotically under a Chi-square distribution with 2*(G-1) degrees of freedom, which is also the rank of variance-covariance matrix. G is the number of comparison groups. The Q statistics (Kim et al., 1995) can be conceived as the Wald statistic under multiple groups. The homogeneity of item parameters across groups under the null hypothesis is stated as:

$$H_o: \delta_{i1} = \delta_{i2} = \cdots = \delta_{iG}, \tag{9}$$

The item parameter vector under DINA model can be $\hat{\delta}_i^{\mathrm{T}} = (\hat{g}_{i1,}\hat{s}_{i1}, \ldots, \hat{g}_{iG}, \hat{s}_{iG})$. Rewriting hypothesis in a matrix form is H_o: $C_i\delta_i = 0$, where C is a contrast matrix that can be written as:

$$C = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & -1 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ 1 & 0 & 0 & 0 & -1 & 0 & \cdots \\ 0 & 1 & 0 & 0 & 0 & -1 & \ddots \end{bmatrix} \tag{10}$$

Following the Wald statistic, the Q statistic is then computed as:

$$Q_i = [C_i\hat{\delta}_{i,}]'\{C_i\hat{\Sigma}_i C_i'\}^{-1}[C_i\hat{\delta}_{i,}], \tag{11}$$

where $\hat{\Sigma}_i$ is the asymptotic variance-covariance matrix specified as follows:

$$\hat{\Sigma}_i = \begin{bmatrix} \sigma^2_{\hat{g}_{i1}} & \sigma_{\hat{g}_{i1}\hat{s}_{i1}} & \cdots & \sigma_{\hat{g}_{i1}\hat{g}_{iG}} & \sigma_{\hat{g}_{i1}\hat{s}_{iG}} \\ \sigma_{\hat{s}_{i1}\hat{g}_{i1}} & \sigma^2_{\hat{s}_{i1}} & \cdots & \sigma_{\hat{s}_{i1}\hat{g}_{iG}} & \sigma_{\hat{s}_{i1}\hat{s}_{iG}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_{\hat{g}_{i1}\hat{g}_{iG}} & \sigma_{\hat{s}_{i1}\hat{g}_{iG}} & \cdots & \sigma^2_{\hat{g}_{iG}} & \sigma_{\hat{s}_{i1}\hat{g}_{iG}} \\ \sigma_{\hat{s}_{i1}\hat{g}_{iG}} & \sigma_{\hat{s}_{i1}\hat{s}_{iG}} & \cdots & \sigma_{\hat{s}_{i1}\hat{g}_{iG}} & \sigma^2_{\hat{s}_{iG}} \end{bmatrix} \tag{12}$$

It is important to recognize that even though the Wald test for comparing a constrained and unconstrained model only requires the estimates from the unconstrained model, technically, just like testing DIF items under IRT (e.g. Woods et al., 2013), by doing separate calibrations for multiple groups, we do need to caution about item contaminated during the calibration stage, and perform item purification procedure if possible. Preceding the multiple group Wald test in the CDM context for each item, it would be desirable to perform an iterative procedure to filter out non-anchor items and estimate students' attribute profiles as fixed, and re-estimate DIF for all items.

In large sample sizes, to get an effect size measure would be beneficial for items exhibiting DIF. There are several measures under the IRT setting, such as Cohen's *d* (Cohen, 1988). George and Robitzsch (2014) adapted the unsigned area (UA) measure (Raju, 1990) to the CDM setting for an item indicator between focal and reference group:

$$UA_i = \sum_{l=1}^{L} w(\alpha_l) \left| P(X_i = 1 | \alpha_l, G_{focal}) - P(X_i = 1 | \alpha_l, G_{reference}) \right|. \quad (13)$$

where L is the attribute skill space for a set of attribute classes, and $w(\alpha_l) = \frac{1}{2}[P(\alpha_l | G_{focal}) + P(\alpha_l | G_{reference})]$.

5.3 Methods

5.3.1 *Data Generation*

For the simulation of response data, we used the sample size, Q-matrix, and item parameters for generating binary responses across two groups by assuming there is no impact between group latent ability distributions. Like Hou et al. (2014), only three fixed sets of guessing and slip parameter values in Table 5-1 were considered in the data generation. The smaller guessing and slip parameter values indicate higher discrimination. The DIF condition was incorporated by setting the item parameter values equal across reference and focal groups, except for the true DIF items where the item parameter for the focal group was manipulated by the DIF design, as illustrated in Table 5-2. For the Q-matrix, we assumed independence among the items and no correlation among the K attributes. Students' attribute profiles were generated with equal probabilities from a discrete uniform distribution for $2^2 = 4$ and $2^3 = 8$ latent classes. Once the response data was simulated for all items in the item bank, the CAT response data was generated in accordance with the key components of adaptive testing.

Note that when selecting items, the CD-CAT algorithm treats items without the knowledge of true DIF items. That is, the interim latent profile estimates were generated by maximizing the joint likelihood function of the DINA model (Equation 14) using the original item parameters, which are without DIF manipulation for both groups. This treatment reflects a more realistic practice of CD-CAT as the DIF parameters are

unknown in the item selection process. Original item parameters were also used for estimating interim during CAT, but the observed responses for the foal group were generated using DIF-manipulated item parameters. This is because when DIF exists but is unknown during the test administration, students from the focal group will be affected as reflected in their item responses. Lastly, the estimated item parameters were used in all computations of Wald test procedures.

$$L(s, g;\ \alpha) = \prod_{i=1}^{N} \prod_{j=1}^{J} \left[(1 - s_j)^{y_{ij}} {s_j}^{1-y_{ij}}\right]^{\eta_{ij}} \left[{g_j}^{y_{ij}} (1 - g_j)^{1-y_{ij}}\right]^{1-\eta_{ij}}, \tag{14}$$

5.3.2 *Simulation Conditions*

Previous research has shown the performance of traditional DIF methods is influenced by factors such as sample size, proportions of DIF items, the magnitude of DIF, test length, and impact between group distributions (e.g. Rogers & Swaminathan, 1993). For this simulation design, the six factors (shown in Table 5-1) are manipulated for the exploration of DIF detection in the CD-CAT setting:

Magnitude of DIF. The magnitude of DIF refers to the absolute difference in success probabilities between the reference (male) and focal (female) groups. Three levels of the DIF magnitude (small, medium, and large) were considered, with the small (ΔS_j or Δg_j = .05) and medium sizes large (ΔS_j or Δg_j = .10) in line with the former literature (Hou et al., 2014; Zhang, 2006). Furthermore, three levels of guessing g_j and slipping S_j parameters $(.1, .2, .3)$ were selected for the DIF items in the measurement model. Note the data for DIF conditions was simulated without the influence of the impact of latent ability distributions between focal and reference groups.

Percentage of DIF items. Previous studies suggested a possible wide range of proportions of DIF items (Paulsen et al., 2020; Qiu et al., 2019). Based on our item bank

size, we considered 0% (no DIF), 20%, 40% of items in the item bank exhibiting DIF, which are aligned to the conditions of 20% and 40% in previous studies.

Item-selection algorithms. In the CD-CAT setting, we investigate the three most commonly used item selection algorithms that are appropriate in a computer adaptive fashion. The first one is a baseline algorithm that randomly selects an item from the item bank and administers items to calculate the examinee's posterior distribution. The GDI and PWKL information algorithms are also used to compute the item selection indices. When most informative items in the pool are administered, the posterior distribution is then updated. Though there is more than one stopping rule to end the cycle, we use a fixed test length to satisfy the test termination rule.

Test and item design. Test design elements including K item attributes, L test length, and J items in the item bank were considered. Based on the past literature for DIF simulation (e.g., Hou et al, 2014; Piromsombat, 2014; Nandakumar & Roussos, 2004; Lei et al., 2006), a test length of 20 items in combination with 2 attributes per item, and a test length of 40 items with 3 attributes; a fixed sample size of 1500 per group; a fixed item bank size = 100; $S_j \sim U$ (-3.6, 3.6); and $g_j \sim U$ (0, .3) were used for the initial analysis to generate a distribution similar to those of previous ones. The conditional ability variables are comparable in patterns across the percentage correct (PCC), IRT-based ability estimate ($\hat{\theta}$), and person's attribute mastery profiles $\widehat{(a)}$.

Table 5-1.

Summary of DIF Conditions for The Simulation (N = 3000)

Item Selection	DIF Percentage	DIF Size	Guessing and Slipping	(K,	L,	J,	N)
Random	No DIF	0.05	S = G = .1	(2	20	100)	3000
PWKL	Low (20%)	0.15	S = G = .2	(3	40	100)	6000
GDI	High (40%)		S = G = .3				

Type of DIF. While some previous studies investigating DIF only explore the uniform DIF situation, both uniform and nonuniform DIF were considered in this design. In this simulated example, an item is manipulated to have uniform DIF when the values of slip and guessing parameters are changed in a direction more toward each other or away from each other. Generally speaking, higher guessing or lower slipping parameters contribute to higher success probability for one group. We allow the difference in direction between the groups to vary across the uniform DIF conditions. An item is assumed to exhibit nonuniform DIF when the probability of success between the focal (female) and the reference (male) groups is dependent on the latent trait of proficiency level. When the slip and guessing parameters are consistently higher, an item is less capable of distinguishing the high-mastery respondents from nonmastery respondents and the data becomes noisier. Overall, a summary of DIF parameters in CD-CAT is presented in Table 5-2.

Table 5-2

Illustration of Uniform and Non-Uniform DIF For Item Parameters

DIF condition	Uniform DIF parameter	Non-uniform DIF parameter
No DIF	$s = 0.1, g = 0.1$; $s = 0.2, g = 0.2$; $s = 0.3, g = 0.3$;	
Low Proportion of DIF:		
Small (Δ = .05)	(g: M = 0.2, F = 0.35)	(g: M = 0.2, F = 0.25)
	(s: M = 0.2, F = 0.05)	(s: M = 0.2, F = 0.25)
Large (Δ = .15)	(g: M = 0.2, F = 0.15)	(g: M = 0.2, F = 0.35)
	(s: M = 0.2, F = 0.25)	(s: M = 0.2, F = 0.35)
High Proportion of DIF:		
Small (Δ = .05)	(g: M = 0.2, F = 0.35)	(g: M = 0.2, F = 0.25)

DIF condition	Uniform DIF parameter	Non-uniform DIF parameter
	(*s*: M = 0.2, F = 0.05)	(*s*: M = 0.2, F = 0.25)
Large (Δ = .15)	(*g*: M = 0.2, F = 0.15)	(*g*: M = 0.2, F = 0.35)
	(*s*: M = 0.2, F = 0.25)	(*s*: M = 0.2, F = 0.35)

Note: For the low proportions of DIF, the first 1-12 items out of 100 items in the item bank were manipulated; for the high proportion of DIF, the first 1-40 items were designated as DIF items. For both low and high DIF, data was simulated across the three levels of guessing (G) and slipping (S) parameter values.

Previously, the Wald test performed an inflated Type 1 error rate under the Q-matrix and other specifications in Hou et al (2014), which followed the structure of 10 items each for a balanced category of 1-, 2-, and 3- attributes, 1000 students and 30 replication trials were generated (n=1000). For our exploratory study, we conducted over a sum of 360 conditions, with each condition then replicated over 1000 trials. Here, only the Wald test without specified anchor items for the DIF detection was used. Statistical properties of different DIF estimation methods, such as Mantel-Haenszel (MH), logistic regression (LR), have been studied for their relative performance in previous DIF research. The simulation study was implemented in R (R Core Team, 2020), and the Wald test was conducted using the GDINA package in R (Ma & de la Torre, 2018).

5.3.3 *Data Analysis*

Type I error and Empirical Power. To assess the performance of different procedures in detecting DIF, we used the criteria of Type 1 Error and Empirical Power for data analysis. The Wald test was performed at the .05 alpha level to investigate the error and empirical power across different test conditions containing DIF and the two item selection algorithms. Here, Type 1 error rates were calculated as the proportion of DIF-free items that were falsely flagged as DIF items (false positives). Among the no-DIF conditions, the Type 1 error rates were calculated for each of the no-DIF items and

averaged across all no-DIF items over the same number of attributes. The statistical power indicates the performance of a hypothesis test, in this case the Wald test, in rejecting a false null hypothesis when all procedures have comparable observed Type 1 error rates. The empirical power rate from the percentages of obtained test statistics that were greater than the empirical cutoff under the same condition, where the 95th percentile of the test statistic under the empirical distributions under the null hypothesis was used. According to previous research (e.g. Cohen, 1992; de la Torre and Lee, 2013; Ma et al., 2021), a test power of 0.8 or above is considered adequate.

Classification Accuracy and Item Usage. First of all, the efficiency of different CD-CAT administrations with the presence of DIF is compared. For each simulation condition, the means of the average attribute recovery rate and the average attribute pattern recovery rate are computed. For the kth attribute, let α_{ikl} and $\hat{\alpha}_{ikl,}$be the true and estimated attribute in attribute vector l, l = 0,1…l, for examinee i, respectively. The calculation is written as

$$P(Attribute)_l = AVE_n(\sum_{i=1}^{n}\sum_{k=1}^{l} I\left[\alpha_{ikl} = \hat{\alpha}_{ikl,}\right]$$
$$P(Pattern)_l = AVE_n(\sum_{i=1}^{n}\prod_{1}^{l} I\left[\alpha_{ikl} = \hat{\alpha}_{ikl,}\right]. \quad (15)$$

where I is the indicator function. The classification accuracy assumes attributes are uniformly distributed for a fixed-length test condition. With other distributions of the attribute vectors, weights can be used from different sampling designs (for details about weights assignment, see Kaplan et al., 2015; de la Torre & Douglas, 2004).

In addition, different indicators of item bank usage are checked. The overlap rate has been used as an indicator of item bank security (Chen et al., 2003). It can be seen as an estimate of the proportion of administered items that are shared by two examinees, and is calculated as $T = \frac{n}{q}s_{er}^2 + \frac{q}{n}$, where s_{er}^2 is the variance of the exposure rates of the

items. The convenience and flexibility of a computer test may be severely compromised if item exposure is not well controlled for test security. Thus, both item usage and test overlap rate were computed to evaluate the item pool usage.

5.4 Results

5.4.1 *Type I Error Study*

Type I error rates are calculated both when all items are DIF-free and when some DIF items are present. The observed Type I error rates for the DIF-free condition are provided in Table 5-3 and 5-4, accounting for each condition across 1000 replications. Here, we evaluate the pattern of the false positive rates with the nominal level of 0.05 (95% of the time falling between 0.4 and 0.6 based on the exact binomial distribution). First, we found that under all test conditions with the random item selection algorithm, the Type I error rates were consistently underestimated and below the level of 0.001. This is the case for both uniform and nonuniform DIF conditions, indicating the risk of rejecting the null hypothesis incorrectly under the random selection algorithm is extremely low.

With longer test length and attributes ($L = 40$, $K = 3$), the PWKL has an averaged Type I error rate at the nominal level. Different item properties yielded in either inflated or deflated false positive rates. With reference items having a good level of guessing and slipping parameters ($g = s = 0.1$), the PWKL converges to around 0.05 level. Same patterns are observed for nonuniform DIF presence as well. Across the conditions of test design, guessing and slipping parameters at 0.2 tend to overestimate Type I error, while a larger probability for guessing and slipping ($g = s = 0.3$) tends to underestimate the error. The results are severely underestimated, though, when it comes to shorter test length and the number of attributes. The trend is that increasing sample

size and item properties move the Type I error rates closer to the nominal level, given the magnitude of all differences is small.

Compared to the PWKL, the GDI has underestimated Type I error rates across conditions, except with longer test (L = 40) and acceptable item properties at ($g = s$ = 0.1, 0.2). In such situations, the averaged error of the GDI algorithm fluctuates from 0.4 to 0.65. Although its performance seems to be influenced by a combination of conditions, the error rates for the two item selection algorithms are similar. For both, the type I error rates are low under 0.04 with shorter tests and attributes. With lower guessing and slipping scenarios (more discriminating), the performance with the GDI tends to pick up more false-positive DIF items, while the PWKL tends to detect more false positive DIF items with noisier data to a certain extent. Sample size has a negligible influence on increasing Type 1 error, except with the GDI it tends to decrease as the sample size increases. In general, both the GDI and the PWKL perform quite consistently across uniform and nonuniform DIF conditions, with the GDI having more deflated Type I error rates than the PWKL.

In summary, aligned with previous simulation studies, we have observed the Type I error rates closer to the nominal level with more attributes per item combined with a longer test length. We have also investigated the results under the DINA model without a computer adaptive selection algorithm, which is equivalent to the calibration of the whole test without any item selection from the item bank. The Type I error rates evaluated under no adaptive algorithms are consistent with the results reported in previous studies (e.g., Hou et al., 2014; Svetina et al., 2018), which is expected as the reference level of performance.

5.4.2 *Empirical Power Study*

The empirical power results for detecting uniform and nonuniform DIF under different simulated conditions are analyzed in this section. Rather than the traditional testing situations, the selective nature of item administration makes analyzing the power of detecting DIF items in the CD-CAT item bank a quite interesting and challenging task, but more importantly, a valuable reference for practical implications. As a confirmative check, we observed the DIF detection method of Wald test performed quite comparably with the empirical power results reported in Hou et al. (2014), under a full DINA context, which is without adaptively selecting items in the item bank through CAT.

There are several results of empirical power that can be noted. First, using different item-selecting algorithms in CD-CAT resulted in sometimes fairly different levels of power based on the factors, and some of the differences were substantial ($\Delta \leq 0.5$).

Table 5-3

Type I Error Rates for Different Non-DIF Conditions (α = .05) For Uniform DIF

	Item Bank (J = 100)									
	GDI		Total	PWKL		Total	Random		Total	Grand Total
Test Condition	L = 20, K = 2	L = 40, K = 3		L = 20, K = 2	L = 40, K = 3		L = 20, K = 2	L = 40, K = 3		
Sample size (n = 3000): N_R = 1500, N_F = 1500										
Total	**0.012**	**0.040**	**0.026**	**0.016**	**0.049**	**0.033**	**< .001**	**< .001**	**< .001**	**0.020**
$g_{Rj} = s_{Rj}$ = 0.1	0.031	0.065	0.048	0.031	0.043	0.037	< .001	< .001	< .001	0.028
$g_{Rj} = s_{Rj}$ = 0.2	0.003	0.049	0.026	0.013	0.074	0.044	0.001	< .001	< .001	0.023
$g_{Rj} = s_{Rj}$ = 0.3	0.002	0.007	0.004	0.004	0.031	0.018	0.001	< .001	< .001	0.007
Sample size (n = 6000): N_R = 3000, N_F = 3000										
Total	**0.011**	**0.034**	**0.023**	**0.018**	**0.054**	**0.036**	**< .001**	**< .001**	**< .001**	**0.020**
$g_{Rj} = s_{Rj}$ = 0.1	0.029	0.056	0.043	0.034	0.049	0.041	0.001	< .001	0.001	0.028
$g_{Rj} = s_{Rj}$ = 0.2	0.002	0.042	0.022	0.015	0.079	0.047	< .001	0.001	0.001	0.023
$g_{Rj} = s_{Rj}$ = 0.3	0.003	0.003	0.003	0.004	0.035	0.019	< .001	< .001	< .001	0.007
Grand Total	**0.011**	**0.037**	**0.024**	**0.017**	**0.052**	**0.034**	**< .001**	**< .001**	**< .001**	**0.020**

Note. g_{Rj} = guessing parameter for the reference group; s_{Rj}= slipping parameter for the reference group;
K = number of attributes required for success on the item, L = test length.

Table 5-4

Type I Error Rates for Different Non-DIF Conditions (α = .05) for Nonuniform DIF

	Item Bank (J = 100)									
	GDI		Total	PWKL		Total	Random		Total	Grand Total
Test Condition	L = 20, K = 2	L = 40, K = 3		L = 20, K = 2	L = 40, K = 3		L = 20, K = 2	L = 40, K = 3		
Sample size (n = 3000): N_R = 1500, N_F = 1500										
Total	**0.011**	**0.039**	**0.025**	**0.015**	**0.049**	**0.032**	**< .001**	**< .001**	**< .001**	**0.019**
$g_{Rj} = s_{Rj}$ = 0.1	0.029	0.065	0.047	0.029	0.044	0.037	< .001	< .001	< .001	0.028
$g_{Rj} = s_{Rj}$ = 0.2	0.002	0.045	0.023	0.013	0.071	0.042	< .001	< .001	< .001	0.022
$g_{Rj} = s_{Rj}$ = 0.3	0.002	0.006	0.004	0.003	0.031	0.017	< .001	< .001	< .001	0.007
Sample size (n = 6000): N_R = 3000, N_F = 3000										
Total	**0.010**	**0.036**	**0.023**	**0.018**	**0.055**	**0.037**	**< .001**	**< .001**	**< .001**	**0.020**
$g_{Rj} = s_{Rj}$ = 0.1	0.026	0.061	0.044	0.034	0.049	0.041	< .001	< .001	< .001	0.028
$g_{Rj} = s_{Rj}$ = 0.2	0.002	0.044	0.023	0.015	0.081	0.048	0.001	< .001	< .001	0.024
$g_{Rj} = s_{Rj}$ = 0.3	0.002	0.004	0.003	0.005	0.036	0.020	< .001	< .001	< .001	0.008
Grand Total	**0.010**	**0.037**	**0.024**	**0.016**	**0.052**	**0.034**	**< .001**	**< .001**	**< .001**	**0.019**

Note. g_{Rj} = guessing parameter for the reference group; s_{Rj}= slipping parameter for the reference group;
K = number of attributes required for success on the item, L = test length.

Table 5-5

Empirical Power Rates for Uniform DIF (Proportion DIF = 40%)

	Item Bank (J = 100)									
	GDI		Total	PWKL		Total	Random		Total	Grand Total
Test Condition	L = 20	L = 40		L = 20	L = 40		L = 20	L = 40		
N = 3000	**0.421**	**0.378**	**0.400**	**0.450**	**0.454**	**0.452**	**0.270**	**0.429**	**0.350**	**0.400**
Large DIF	**0.695**	**0.650**	**0.672**	**0.713**	**0.738**	**0.725**	**0.533**	**0.832**	**0.683**	**0.693**
$g_{Rj} = s_{Rj} = 0.1$	0.826	0.638	0.732	0.756	0.722	0.739	0.303	0.591	0.447	0.639
$g_{Rj} = s_{Rj} = 0.2$	0.677	0.737	0.707	0.718	0.803	0.761	0.796	0.996	0.896	0.788
$g_{Rj} = s_{Rj} = 0.3$	0.581	0.575	0.578	0.665	0.688	0.677	0.500	0.909	0.705	0.653
Small DIF	**0.148**	**0.106**	**0.127**	**0.186**	**0.171**	**0.179**	**0.008**	**0.026**	**0.017**	**0.107**
$g_{Rj} = s_{Rj} = 0.1$	0.312	0.202	0.257	0.340	0.295	0.318	0.013	0.057	0.035	0.203
$g_{Rj} = s_{Rj} = 0.2$	0.083	0.094	0.089	0.147	0.147	0.147	0.006	0.013	0.010	0.082
$g_{Rj} = s_{Rj} = 0.3$	0.049	0.021	0.035	0.072	0.071	0.072	0.004	0.007	0.006	0.037
N = 6000	**0.579**	**0.555**	**0.567**	**0.590**	**0.619**	**0.604**	**0.499**	**0.601**	**0.550**	**0.574**
Large DIF	**0.817**	**0.834**	**0.825**	**0.784**	**0.848**	**0.816**	**0.934**	**0.989**	**0.962**	**0.868**
$g_{Rj} = s_{Rj} = 0.1$	0.860	0.873	0.867	0.807	0.856	0.832	0.842	0.966	0.904	0.867
$g_{Rj} = s_{Rj} = 0.2$	0.827	0.868	0.848	0.796	0.880	0.838	0.999	1.000	1.000	0.895
$g_{Rj} = s_{Rj} = 0.3$	0.763	0.760	0.762	0.750	0.809	0.780	0.962	1.000	0.981	0.841
Small DIF	**0.342**	**0.276**	**0.309**	**0.395**	**0.390**	**0.393**	**0.064**	**0.214**	**0.139**	**0.280**
$g_{Rj} = s_{Rj} = 0.1$	0.577	0.470	0.524	0.587	0.568	0.578	0.134	0.446	0.290	0.464
$g_{Rj} = s_{Rj} = 0.2$	0.263	0.236	0.250	0.353	0.373	0.363	0.038	0.128	0.083	0.232
$g_{Rj} = s_{Rj} = 0.3$	0.185	0.122	0.154	0.246	0.228	0.237	0.021	0.068	0.045	0.145
Grand Total	**0.500**	**0.466**	**0.483**	**0.520**	**0.537**	**0.528**	**0.385**	**0.515**	**0.450**	**0.487**

Note. L = 20 is test length of 20 with 2 attributes; L = 40 is test length of 40 with 3 attributes; Large DIF size = 0.15; Small DIF size = 0.05.

Table 5-6

Empirical Power Rates for Uniform DIF (Proportion DIF = 20%)

	Item Bank (J = 100)									
	GDI		Total	PWKL		Total	Random		Total	Grand Total
Test Condition	L = 20	L = 40		L = 20	L = 40		L = 20	L = 40		
N = 3000	**0.430**	**0.372**	**0.401**	**0.446**	**0.448**	**0.447**	**0.262**	**0.423**	**0.343**	**0.397**
Large DIF	**0.710**	**0.645**	**0.678**	**0.714**	**0.728**	**0.721**	**0.516**	**0.823**	**0.670**	**0.689**
$g_{Rj} = s_{Rj} = 0.1$	0.826	0.618	0.722	0.756	0.705	0.731	0.296	0.582	0.439	0.631
$g_{Rj} = s_{Rj} = 0.2$	0.703	0.732	0.718	0.708	0.797	0.753	0.789	0.997	0.893	0.788
$g_{Rj} = s_{Rj} = 0.3$	0.602	0.584	0.593	0.679	0.682	0.681	0.462	0.891	0.677	0.650
Small DIF	**0.150**	**0.100**	**0.125**	**0.178**	**0.168**	**0.173**	**0.008**	**0.023**	**0.016**	**0.105**
$g_{Rj} = s_{Rj} = 0.1$	0.327	0.192	0.260	0.322	0.281	0.302	0.015	0.052	0.034	0.198
$g_{Rj} = s_{Rj} = 0.2$	0.077	0.084	0.081	0.139	0.153	0.146	0.006	0.013	0.010	0.079
$g_{Rj} = s_{Rj} = 0.3$	0.045	0.024	0.035	0.074	0.070	0.072	0.004	0.005	0.005	0.037
N = 6000	**0.583**	**0.555**	**0.569**	**0.583**	**0.618**	**0.600**	**0.495**	**0.595**	**0.545**	**0.571**
Large DIF	**0.824**	**0.833**	**0.828**	**0.784**	**0.854**	**0.819**	**0.927**	**0.986**	**0.957**	**0.868**
$g_{Rj} = s_{Rj} = 0.1$	0.874	0.853	0.864	0.816	0.851	0.834	0.825	0.959	0.892	0.863
$g_{Rj} = s_{Rj} = 0.2$	0.830	0.867	0.849	0.780	0.892	0.836	0.999	1.000	1.000	0.895
$g_{Rj} = s_{Rj} = 0.3$	0.768	0.778	0.773	0.755	0.820	0.788	0.958	1.000	0.979	0.847
Small DIF	**0.342**	**0.276**	**0.309**	**0.382**	**0.381**	**0.382**	**0.063**	**0.203**	**0.133**	**0.275**
$g_{Rj} = s_{Rj} = 0.1$	0.579	0.463	0.521	0.564	0.557	0.561	0.139	0.418	0.279	0.453
$g_{Rj} = s_{Rj} = 0.2$	0.264	0.236	0.250	0.351	0.358	0.355	0.030	0.126	0.078	0.228
$g_{Rj} = s_{Rj} = 0.3$	0.182	0.130	0.156	0.230	0.229	0.230	0.021	0.066	0.044	0.143
Grand Total	**0.506**	**0.463**	**0.485**	**0.515**	**0.533**	**0.524**	**0.379**	**0.509**	**0.444**	**0.484**

Note. L = 20 is test length of 20 with 2 attributes; L = 40 is test length of 40 with 3 attributes; Large DIF size = 0.15; Small DIF size = 0.05.

Table 5-7

Empirical Power Rates for Nonuniform DIF (Proportion DIF = 40%)

	Item Bank (J = 100)									
	GDI		Total	PWKL		Total	Random		Total	Grand Total
Test Condition	L = 20	L = 40		L = 20	L = 40		L = 20	L = 40		
N = 3000	**0.425**	**0.360**	**0.393**	**0.468**	**0.445**	**0.457**	**0.223**	**0.408**	**0.315**	**0.388**
Large DIF	**0.668**	**0.612**	**0.640**	**0.703**	**0.712**	**0.708**	**0.440**	**0.799**	**0.619**	**0.656**
$g_{Rj} = s_{Rj} = 0.1$	0.830	0.726	0.778	0.792	0.806	0.799	0.759	0.985	0.872	0.816
$g_{Rj} = s_{Rj} = 0.2$	0.655	0.633	0.644	0.710	0.734	0.722	0.400	0.856	0.628	0.665
$g_{Rj} = s_{Rj} = 0.3$	0.520	0.476	0.498	0.608	0.597	0.603	0.160	0.555	0.358	0.486
Small DIF	**0.182**	**0.108**	**0.145**	**0.233**	**0.178**	**0.206**	**0.006**	**0.017**	**0.011**	**0.121**
$g_{Rj} = s_{Rj} = 0.1$	0.347	0.205	0.276	0.397	0.296	0.347	0.010	0.031	0.021	0.214
$g_{Rj} = s_{Rj} = 0.2$	0.130	0.093	0.112	0.202	0.174	0.188	0.004	0.012	0.008	0.103
$g_{Rj} = s_{Rj} = 0.3$	0.068	0.027	0.048	0.100	0.065	0.083	0.004	0.007	0.006	0.045
N = 6000	**0.568**	**0.542**	**0.555**	**0.584**	**0.609**	**0.596**	**0.447**	**0.569**	**0.508**	**0.553**
Large DIF	**0.802**	**0.815**	**0.809**	**0.762**	**0.842**	**0.802**	**0.854**	**0.995**	**0.925**	**0.845**
$g_{Rj} = s_{Rj} = 0.1$	0.885	0.898	0.892	0.811	0.891	0.851	0.998	1.000	0.999	0.914
$g_{Rj} = s_{Rj} = 0.2$	0.812	0.832	0.822	0.768	0.876	0.822	0.940	1.000	0.970	0.871
$g_{Rj} = s_{Rj} = 0.3$	0.710	0.714	0.712	0.706	0.760	0.733	0.624	0.986	0.805	0.750
Small DIF	**0.334**	**0.269**	**0.302**	**0.407**	**0.375**	**0.391**	**0.040**	**0.142**	**0.091**	**0.261**
$g_{Rj} = s_{Rj} = 0.1$	0.528	0.426	0.477	0.576	0.536	0.556	0.082	0.303	0.193	0.409
$g_{Rj} = s_{Rj} = 0.2$	0.306	0.255	0.281	0.404	0.384	0.394	0.025	0.088	0.057	0.244
$g_{Rj} = s_{Rj} = 0.3$	0.169	0.127	0.148	0.240	0.204	0.222	0.012	0.034	0.023	0.131
Grand Total	**0.497**	**0.451**	**0.474**	**0.526**	**0.527**	**0.527**	**0.335**	**0.488**	**0.411**	**0.471**

Note. L = 20 is test length of 20 with 2 attributes; L = 40 is test length of 40 with 3 attributes; Large DIF size = 0.15; Small DIF size = 0.05.

Table 5-8

Empirical Power Rates for Nonuniform DIF (Proportion DIF = 20%)

	Item Bank (J = 100)									
	GDI		Total	PWKL		Total	Random		Total	Grand Total
Test Condition	L = 20	L = 40		L = 20	L = 40		L = 20	L = 40		
N = 3000	**0.425**	**0.364**	**0.394**	**0.473**	**0.443**	**0.458**	**0.224**	**0.410**	**0.317**	**0.390**
Large DIF	**0.680**	**0.615**	**0.648**	**0.704**	**0.712**	**0.708**	**0.443**	**0.806**	**0.625**	**0.660**
$g_{Rj} = s_{Rj} = 0.1$	0.832	0.718	0.775	0.768	0.781	0.775	0.750	0.983	0.867	0.805
$g_{Rj} = s_{Rj} = 0.2$	0.657	0.623	0.640	0.696	0.741	0.719	0.392	0.860	0.626	0.662
$g_{Rj} = s_{Rj} = 0.3$	0.552	0.504	0.528	0.649	0.615	0.632	0.188	0.575	0.382	0.514
Small DIF	**0.169**	**0.113**	**0.141**	**0.242**	**0.173**	**0.207**	**0.005**	**0.014**	**0.010**	**0.119**
$g_{Rj} = s_{Rj} = 0.1$	0.339	0.209	0.274	0.404	0.286	0.345	0.010	0.030	0.020	0.213
$g_{Rj} = s_{Rj} = 0.2$	0.106	0.101	0.104	0.235	0.166	0.201	0.002	0.007	0.005	0.103
$g_{Rj} = s_{Rj} = 0.3$	0.062	0.030	0.046	0.087	0.066	0.077	0.003	0.006	0.005	0.042
N = 6000	**0.579**	**0.540**	**0.560**	**0.581**	**0.598**	**0.589**	**0.452**	**0.571**	**0.511**	**0.553**
Large DIF	**0.806**	**0.812**	**0.809**	**0.770**	**0.836**	**0.803**	**0.864**	**0.997**	**0.931**	**0.848**
$g_{Rj} = s_{Rj} = 0.1$	0.876	0.870	0.873	0.822	0.860	0.841	0.998	1.000	0.999	0.904
$g_{Rj} = s_{Rj} = 0.2$	0.812	0.833	0.823	0.767	0.878	0.823	0.937	1.000	0.969	0.871
$g_{Rj} = s_{Rj} = 0.3$	0.730	0.734	0.732	0.722	0.769	0.746	0.658	0.990	0.824	0.767
Small DIF	**0.352**	**0.268**	**0.310**	**0.391**	**0.360**	**0.376**	**0.039**	**0.145**	**0.092**	**0.259**
$g_{Rj} = s_{Rj} = 0.1$	0.550	0.425	0.488	0.558	0.492	0.525	0.079	0.301	0.190	0.401
$g_{Rj} = s_{Rj} = 0.2$	0.311	0.257	0.284	0.398	0.384	0.391	0.025	0.095	0.060	0.245
$g_{Rj} = s_{Rj} = 0.3$	0.195	0.122	0.159	0.218	0.203	0.211	0.013	0.039	0.026	0.132
Grand Total	**0.502**	**0.452**	**0.477**	**0.527**	**0.520**	**0.524**	**0.338**	**0.491**	**0.414**	**0.472**

Note. L = 20 is test length of 20 with 2 attributes; L = 40 is test length of 40 with 3 attributes; Large DIF size = 0.15; Small DIF size = 0.05.

Second, the proportion of DIF items in the item bank seems to have little to negligible influence on the recovery of DIF items in terms of patterns and magnitude of empirical power rates (see Tables 3.5- 3.8). Third, the item parameters along with different DIF sizes tend to have a huge impact on the DIF detection rate, with the interaction of others. Across the item selection algorithms and the simulated CD-CAT conditions, when there is sufficiently large sample size ($N = 6000$), good-quality item parameter ($g = s = 0.1$) and presence of a large DIF size ($\Delta_i = .15$), DIF detection of the Wald test consistently satisfies an excellent power rate above 0.8. According to Cohen's (1992) cutoff values, power rates above 0.70 and 0.80 are considered moderate and excellent, respectively.

The empirical powers for recovering uniform DIF are shown in Table 5-5 and Table 5-6. Across the three item-selecting algorithms, the random selecting algorithm serves as a baseline performance for power rate. It has very limited power when DIF presence has a small size ($\Delta_{Reference - Focal} = .05$), not exceeding 0.45. Its power for detecting large-size DIF items ($\Delta_{Reference - Focal} = .15$) is somewhat unstable with shorter tests and smaller sample size ($N = 3000$, $L = 20$). That is, with the random algorithm, detecting DIF loses its power for less discriminating items ($g = s = 0.2, 0.3$) for nonuniform DIF given the above condition; but for the uniform DIF, the power drops for high and low guessing and slipping parameters ($g = s = 0.1, 0.3$). Such a distinction does not exist for the other two algorithms. However, when the sample size is large ($N = 6000$), the random selecting algorithm has the highest power among the three for detecting the large-size DIF items across the conditions. The fluctuating rates from the random algorithm as the power of detecting DIF is somewhat understandable, since the algorithm doesn't differentiate any information associated with item properties but assigns a uniform probability distribution of item sampling. In a way, the interim

estimation of a respondent's latent ability using bad items tends to be less efficient and may lead to inaccurate estimates of group-level item parameters for conducting the statistical test. As the sample size and test length increase, such shortcomings eventually is reduced.

The PWKL has the highest power rates for 70% of the occasions among the three item selection methods. It has the most advantage for detecting DIF items with small DIF sizes in terms of the power rates, albeit no greater than 0.6; and with the largest difference in power rates 50% higher than those of the GDI algorithm. When there is a large size of DIF ($\Delta_{Reference\ -\ Focal}$ = .15) among item parameters, the PWKL tends to give in its advantage to the GDI or the random algorithm based on the combined factors of test length, DIF size, and item properties. On large-size DIF detection, when there is a short test, it is expected for the GDI to have excellent power rates (over 0.8), and the best performance among the three, given items are relatively high-discriminating ($g = s = 0.1$). As the sample size increases, power under the GDI still performs better under the large-size DIF and a short test than the PWKL, while the random algorithm has the highest power rates on average. There is an exception for the GDI to top the power, when the test is short and the item qualities are good ($g = s = 0.1$). Both the GDI and the PWKL tend to reduce the usage of items among the DIF contaminated sets from item administration. For the GDI situation under the short test, the power rates slightly decrease when DIF proportions are higher, whereas the PWKL and random algorithm tend to increase power with higher proportions of DIF. In general, the GDI tends to differentiate between item properties more sensitively, and the power rates are overall more comparable with the PWKL than the rates from the random selection.

Results for the nonuniform DIF are reported in Table 5-7 (40% proportion) and Table 5-8 (20% proportion). They display the power rates of the three item selection algorithms with equivalent performance across high or low proportions of DIF conditions. On the other hand, the averaged power rates increase by about 3.4% from the nonuniform DIF detection to the uniform DIF across the CD-CAT settings. The same patterns existing in the uniform DIF across the test conditions can also be observed for the nonuniform DIF performances, with the difference between types of DIF being of almost negligible magnitude for the GDI and the PWKL. The Wald method is able to detect DIF items of large DIF magnitude with adequate power rates when items are of high quality (more discriminating), and more so with the nonuniform DIF than the uniform DIF. In contrast, the procedure is less likely to correctly detect DIF items of a small DIF magnitude ($\Delta_{Reference - Focal} = .05$), yet the power rates for the nonuniform type are slightly higher than the uniform type around 70% of the time for the GDI and the PWKL. As the baseline performance, the power under the random selection algorithm increases about 10% from nonuniform to uniform DIF conditions, but sporadically; it increases substantially when the sample sizes increase from 3000 to 6000, up to 13 times higher. Noticeably, under the conditions of larger DIF size ($\Delta_{Reference - Focal} = .15$) and longer test ($L = 40$), the random algorithm tends to yield the best power for nonuniform DIF among the three algorithms, except when item quality is bad, indicating a possible interaction effect between test length, DIF size, item quality, and item selection methods.

The GDI has the highest power rate for nonuniform DIF in certain conditions with large DIF sizes. For example, the algorithm may have some advantage over the PWKL when the test length is short (with a small number of attributes per item) and item quality is good. The PWKL still outperforms in other cases, consistently yielding

the best power for the small DIF size ($\Delta_{Reference - Focal} = .05$). An inspection of the PWKL reveals the algorithm tends to administer slightly less DIF-contaminated items into the tests than the GDI, with negligible differences. This means for the two algorithms, some DIF items are never exposed throughout one condition; whereas for the random algorithm, every item in the item bank has an equal chance of being used and exposed. For the former, exclusions of the DIF items may lead to the compromise of the power rates to some extent. For the latter, it may explain the almost perfect power when there is a sufficient large sample size and DIF magnitude for the random algorithm in the Wald test, except when the bad item quality starts to detriment the estimation accuracy.

In summary, we see that for the DIF recovery, the GDI and the PWKL have different advantages under different conditions of CD-CAT setting. The PWKL seems to be more consistent and outperforms overall in terms of power, while the GDI seems to be more sensitive to item properties and the number of attributes required by items. It is understandable since the GDI algorithm takes both the item discrimination and the posterior distribution into account. Lastly, the random algorithm can sometimes severely restrict the power, and is more unstable overall. Thus, to answer the first research question, we understand now that the DIF recovery in the CD-CAT setting can be impacted by which item-selecting algorithm we choose for the test specification. The interaction of conditions also creates complications in discussing power rates. To understand what really happened, we examined factors of CD-CAT separately, including sample sizes, test length and the number of item attributes, DIF condition, and item parameters (Figure 5-2 and 5-3).

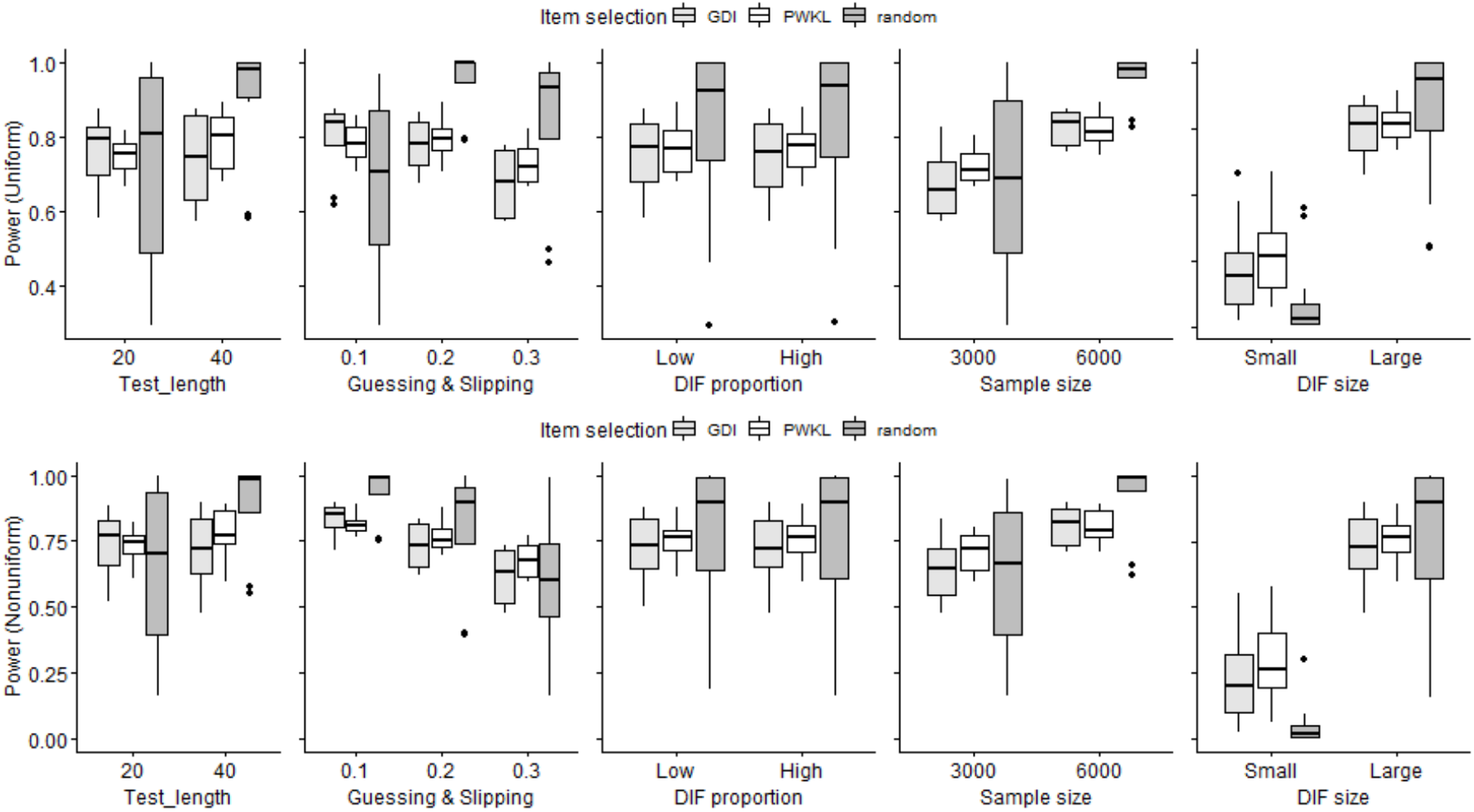


Figure 5-2. Boxplots from 25th to 75th percentile for empirical power across conditions.

Note. The first four columns are the power rates from the large size DIF condition.

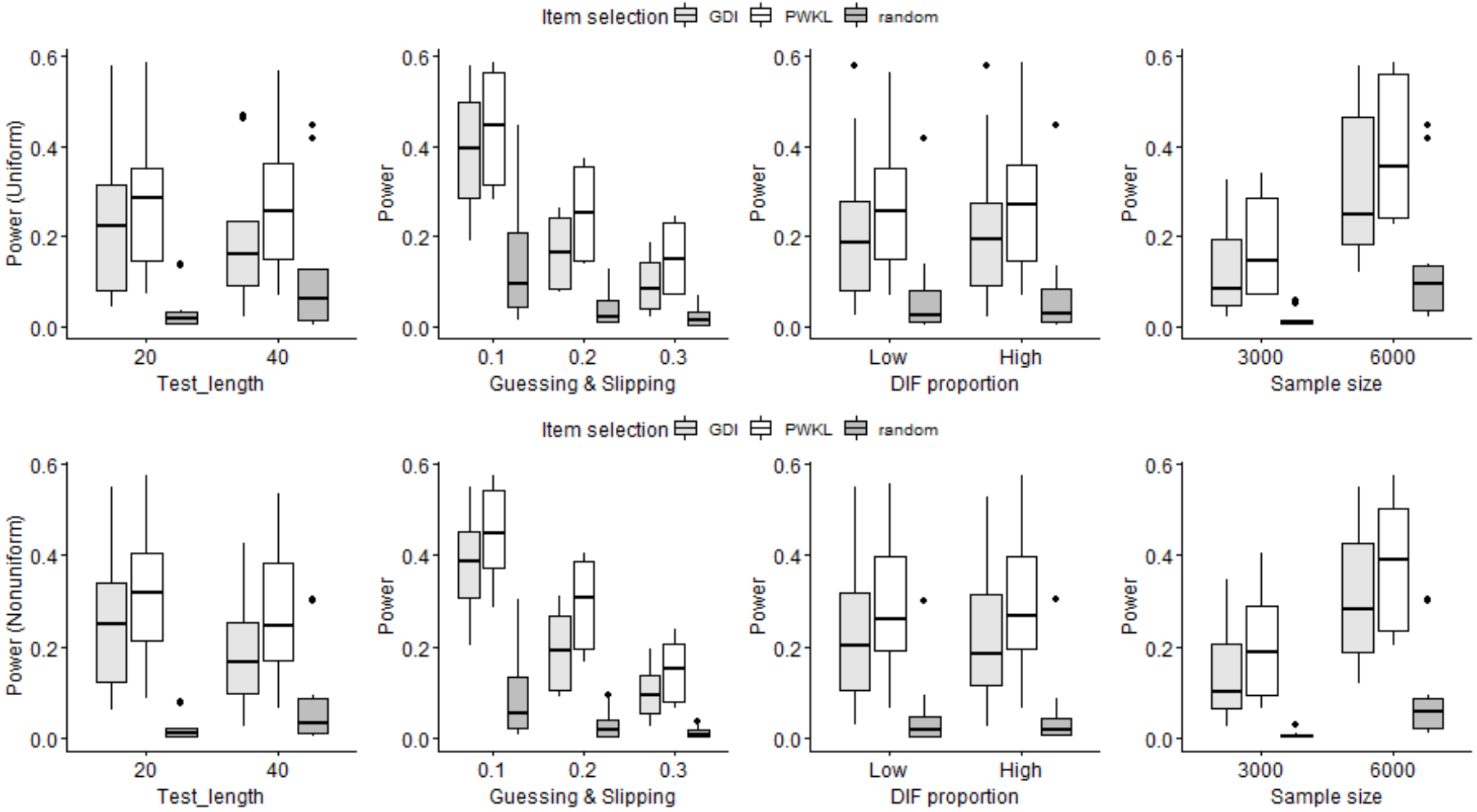


Figure 5-3. Boxplots from 25th to 75th Percentile for the power rates in the small size DIF conditions.

The impact of sample size. Regardless of the item selection algorithm and test conditions, increasing the sample size contributes to the higher power rates unanimously. The difference is most substantial when detecting DIF for a small DIF size, where power rates are restricted in the smaller sample size ($N_{Reference} = N_{focal} =$ 1500). For conditions with a large DIF size, we expect the larger sample sizes to increase power rates by 25%-60% overall. It is also observed that for conditions of the large DIF size, increasing the sample size can reduce the variability in power rates while boosting the power substantially. For situations under the small DIF size, there is usually a greater variability in power rates along with the increased power.

The impact of item quality. According to the previous literature (e.g. Hou et al., 2014), the less discriminating quality of items based on larger slip and guessing values negatively affects the ability of the Wald test to detect DIF. As expected, item quality impacts the item-selecting algorithms through information index, and in turn, how the algorithm perceives and administers items will impact the power for DIF detection. In most cases, increasing guessing and slipping parameters (lower item quality) will decrease power rates for DIF detection, except for the uniform DIF scenarios. There are several exceptions when the DIF size is large, the algorithms have better power rates with medium quality items ($g = s = 0.2$) than the good or bad quality items ($g = s = 0.1$, 0.3). That is no longer the case with the GDI and the PWKL on short tests (with fewer attributes per item).

Impact of test length. Test length conditions are created with the combined factor of the number of attributes per item. The shorter test has two attributes per item, and the longer test has three attributes required per item. Thus, it may be harder to isolate the effect of each in explaining the decreasing patterns of power rates under the GDI algorithm. For the uniform DIF, increasing the test length generally increases the

power, except for some conditions with small DIF size and/or high-quality items with guessing and slipping parameters equal to 0.1, where it affects both the GDI and the PWKL algorithm. For the nonuniform DIF, increasing the test length results in lower power rates for the small DIF size conditions for both algorithms. In addition, under the GDI, the power rates drop by a small degree (from 3% to 14%) for the test-length increase in large DIF conditions ($N = 3000$). This behavior could be due to the complication created by item property and DIF sizes in item parameters.

The Impact of DIF Proportion and DIF Size. As expected, the larger DIF size ($\Delta_{R-F} = .15$) results in much higher power rates than the small size ($\Delta_{R-F} = .05$), regardless of the item or test conditions. The proportion of DIF items has a trivial impact on the empirical power. Overall, the power rates are similar across different proportions of DIF items. For the random item selection, the power rates are mostly equivalent across high or low proportions of DIF items. For the other two, the higher proportion (40%) condition has a slightly higher average power rate, except when the DIF size is large or the test length is short; where it has a trivial decrease under the GDI algorithm.

5.4.3 *Classification Accuracy and Item Exposure Rate*

We calculated the relative efficiency in attribute and pattern classification accuracy for each iteration. For all simulated conditions, we observe the average attribute recovery rates are slightly higher than the pattern recovery rates, which is expected in the results of previous CD-CAT studies (e.g. Kaplan et al., 2015). Since the attribute recovery rates share similar patterns with the pattern recovery rates, and the differences of the accuracy rates between different sample sizes are trivial (< 0.01), only the results of the average pattern recovery rates are reported in Table 5-9 and Table 5-10 for uniform and nonuniform DIF presence.

Based on the type of DIF presence, the classification accuracy differs for different item selection algorithms. The GDI has the highest classification accuracy overall, with an averaged pattern recovery rate of 0.96 across the conditions. The PWKL has a sufficient averaged pattern recovery rate of 0.94 for the uniform DIF and 0.92 for the nonuniform DIF, slightly lower than the GDI across all conditions. The random selection algorithm has an averaged recovery rate of 0.9 and 0.89 respectively for the two types of DIF. There are some conditions under the random selecting algorithm, where the recovery rates are the lowest of all (0.67 for pattern recovery and 0.86 for attribute recovery). This is mostly due to the effect of large DIF size in conjunction with poor item properties (large guessing and slip values). It seems such a combination will cause more bias in the presence of longer tests for both the PWKL and the random selection algorithm. Inefficient item parameters impact the interim estimates as the algorithms administer more items, and slightly decrease the accuracy over time. On the other hand, the accuracy rates of the GDI improve with the longer test length.

To understand the impact of DIF presence on the classification accuracy of CD-CAT, we calculate the difference between pattern recovery rates from the different DIF scenarios to the DIF-free conditions. Based on the previous research (Kaplan et al., 2015), the differences in the recovery rates larger than the cut-off point of 0.10 is considered substantial, between 0.01 and 0.10 is considered slight, and below 0.01 is negligible. Under most conditions of the attribute recovery, the differences are negligible to slight based on item parameter levels (Figure 5-5). For the pattern recovery, the DIF presence creates a substantial decrease in the GDI recovery rates, when there is a large DIF size, a high proportion of DIF items, and the items have high guessing and slip values ($g = s = 0.3$). Figure 5-6 shows that DIF presence reduces pattern accuracy to a negligible degree with high quality items ($g = s = 0.1$), a slight

degree with median quality items ($g = s = 0.2$), and slight to substantial degrees with low quality items ($g = s = 0.3$).

Regarding the overlap rate results, it is observed that the random selection algorithm shows the best test security, with an overlap rate no larger than .40 (for the test length of 40 out of 100 items in the item bank). The averaged overlap rate for the PWKL (0.71) is slightly higher than those of the GDI (0.60). For both, the cost of the overlap rate increases as the item quality decreases with higher guessing and slip values and larger DIF size. On the other hand, item exposure rate refers to the relative frequency with which an item is administered across all CAT administrations, that is, the proportion of all tests in which an item is administered in one condition. Figure 5-7 shows an example of item exposure rate in the item bank under the GDI algorithm, along with attribute and person's attribute pattern recovery rate by item sequence. This is done under the low DIF proportion and large DIF size for the uniform DIF. As we can see, the GDI is fairly efficient with administering only four to five items, and about only half of the DIF items (item 1 to item 20) are selected or exposed throughout the test.

We specifically calculated the proportion of exposure rate being zero among the DIF-contaminated items (DIF item exclusion) and the average item usage rate across the DIF contaminated item set for each iteration. The GDI and the PWKL turned to only select an averaged portion of 51% and 46% of the DIF contaminated items for conditions under the test length of 20, and an averaged 99.7% and 87.6% of the DIF items for conditions of test length of 40. On the other hand, the random selection algorithm uses every item in the item bank (100%). Averaging the item exposure rates over the DIF items, all algorithms show a controlled rate about the rate of the number of

items in the test over the item bank size. The differences between the baseline of the random selection algorithm and the GDI or the PWKL are minimal (Figure 5-6).

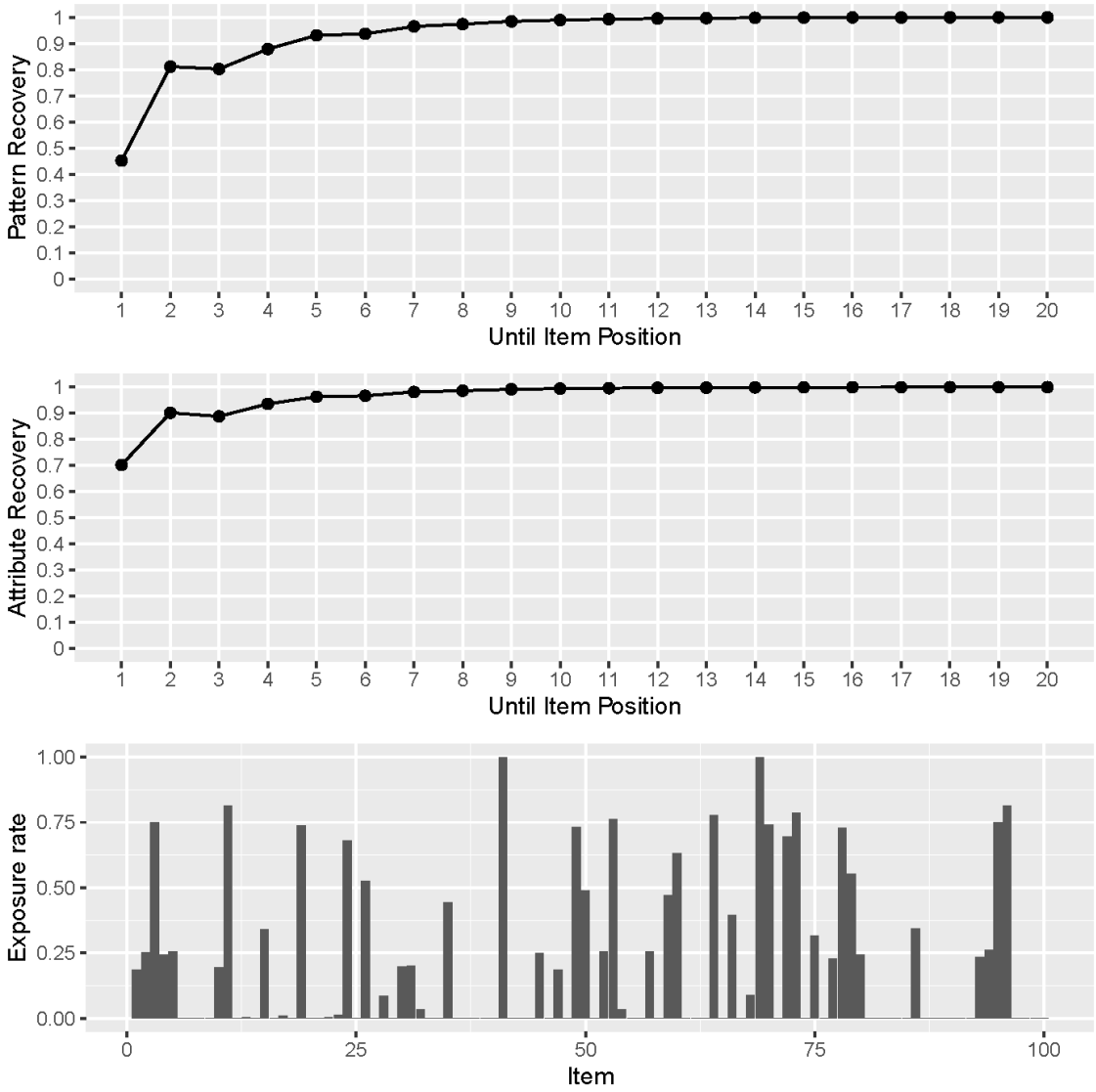


Figure 5-4. Pattern recovery, attribute recovery, and exposure rate of GDI algorithm under one simulated condition with test length of 20 items, 2 attributes per item.

Note. The first 20 items are the true DIF items with large DIF size.

Table 5-9

Pattern Recovery Rates for Uniform DIF

Item Quality	DIF Condition	Item Bank (J = 100)								
		GDI		Overlap rate	PWKL		Overlap rate	Random		Overlap rate
		L = 20	L = 40		L = 20	L = 40		L = 20	L = 40	
$g_{Rj} = s_{Rj} = 0.1$	Hi-L	0.999	1.000	0.594	0.994	0.995	0.694	0.988	0.985	0.300
	Hi-S	1.000	1.000	0.599	0.998	0.998	0.697	0.991	0.990	0.300
	Lo-L	1.000	1.000	0.594	0.997	0.997	0.697	0.990	0.988	0.300
	Lo-S	1.000	1.000	0.598	0.998	0.998	0.697	0.992	0.990	0.300
	No-DIF	1.000	1.000	0.597	0.998	0.998	0.695	0.991	0.990	0.300
$g_{Rj} = s_{Rj} = 0.2$	Hi-L	0.980	0.989	0.594	0.958	0.956	0.704	0.935	0.924	0.300
	Hi-S	0.988	0.994	0.595	0.972	0.972	0.700	0.942	0.935	0.300
	Lo-L	0.986	0.993	0.595	0.967	0.967	0.702	0.941	0.932	0.300
	Lo-S	0.988	0.995	0.596	0.971	0.972	0.702	0.943	0.935	0.300
	No-DIF	0.988	0.995	0.596	0.971	0.972	0.700	0.943	0.937	0.300
$g_{Rj} = s_{Rj} = 0.3$	Hi-L	0.863	0.874	0.608	0.828	0.798	0.714	0.785	0.741	0.300
	Hi-S	0.881	0.902	0.600	0.845	0.823	0.706	0.796	0.756	0.300
	Lo-L	0.877	0.895	0.606	0.840	0.817	0.709	0.795	0.755	0.300
	Lo-S	0.883	0.904	0.601	0.845	0.824	0.708	0.798	0.758	0.300
	No-DIF	0.883	0.905	0.603	0.845	0.825	0.709	0.797	0.758	0.300

Note. Pattern recovery rate is averaged across the sample size. Overlap rate is the averaged overlap rate across test lengths and sample sizes; Hi-L is high proportion of DIF with large size; Hi-S is high proportion of DIF with small size; Lo-L is low proportion of DIF with large size; Lo-S is low proportion of DIF with small size; L = 20 is test length of 20 with 2 attributes; L = 40 is test length of 40 with 3 attributes.

Table 5-10

Pattern Recovery Rates for Nonuniform DIF

Item Quality	DIF Condition	Item Bank (J = 100)								
		GDI		Overlap rate	PWKL		Overlap rate	Random		Overlap rate
		L = 20	L = 40		L = 20	L = 40		L = 20	L = 40	
$g_{Rj} = s_{Rj} = 0.1$	Hi-L	0.998	0.999	0.596	0.993	0.993	0.695	0.981	0.978	0.300
	Hi-S	1.000	1.000	0.597	0.997	0.997	0.697	0.989	0.987	0.300
	Lo-L	1.000	1.000	0.597	0.996	0.997	0.696	0.987	0.986	0.300
	Lo-S	1.000	1.000	0.596	0.998	0.998	0.697	0.991	0.989	0.300
	No-DIF	1.000	1.000	0.597	0.998	0.998	0.696	0.992	0.990	0.300
$g_{Rj} = s_{Rj} = 0.2$	Hi-L	0.965	0.979	0.596	0.941	0.939	0.700	0.906	0.892	0.300
	Hi-S	0.983	0.992	0.596	0.963	0.963	0.703	0.934	0.923	0.300
	Lo-L	0.980	0.990	0.595	0.958	0.959	0.700	0.927	0.917	0.300
	Lo-S	0.986	0.993	0.595	0.967	0.968	0.703	0.939	0.931	0.300
	No-DIF	0.989	0.995	0.596	0.971	0.971	0.703	0.943	0.935	0.300
$g_{Rj} = s_{Rj} = 0.3$	Hi-L	0.806	0.807	0.605	0.769	0.725	0.711	0.724	0.665	0.300
	Hi-S	0.860	0.879	0.603	0.822	0.797	0.710	0.774	0.731	0.300
	Lo-L	0.848	0.863	0.603	0.810	0.780	0.709	0.762	0.718	0.300
	Lo-S	0.873	0.892	0.602	0.834	0.811	0.709	0.787	0.745	0.300
	No-DIF	0.883	0.904	0.603	0.846	0.825	0.709	0.798	0.759	0.300

Note. Pattern recovery rate is averaged across the sample size. Overlap rate is the averaged overlap rate across test lengths and sample sizes; Hi-L is high proportion of DIF with large size; Hi-S is high proportion of DIF with small size; Lo-L is low proportion of DIF with large size; Lo-S is low proportion of DIF with small size; L = 20 is test length of 20 with 2 attributes; L = 40 is test length of 40 with 3 attributes.

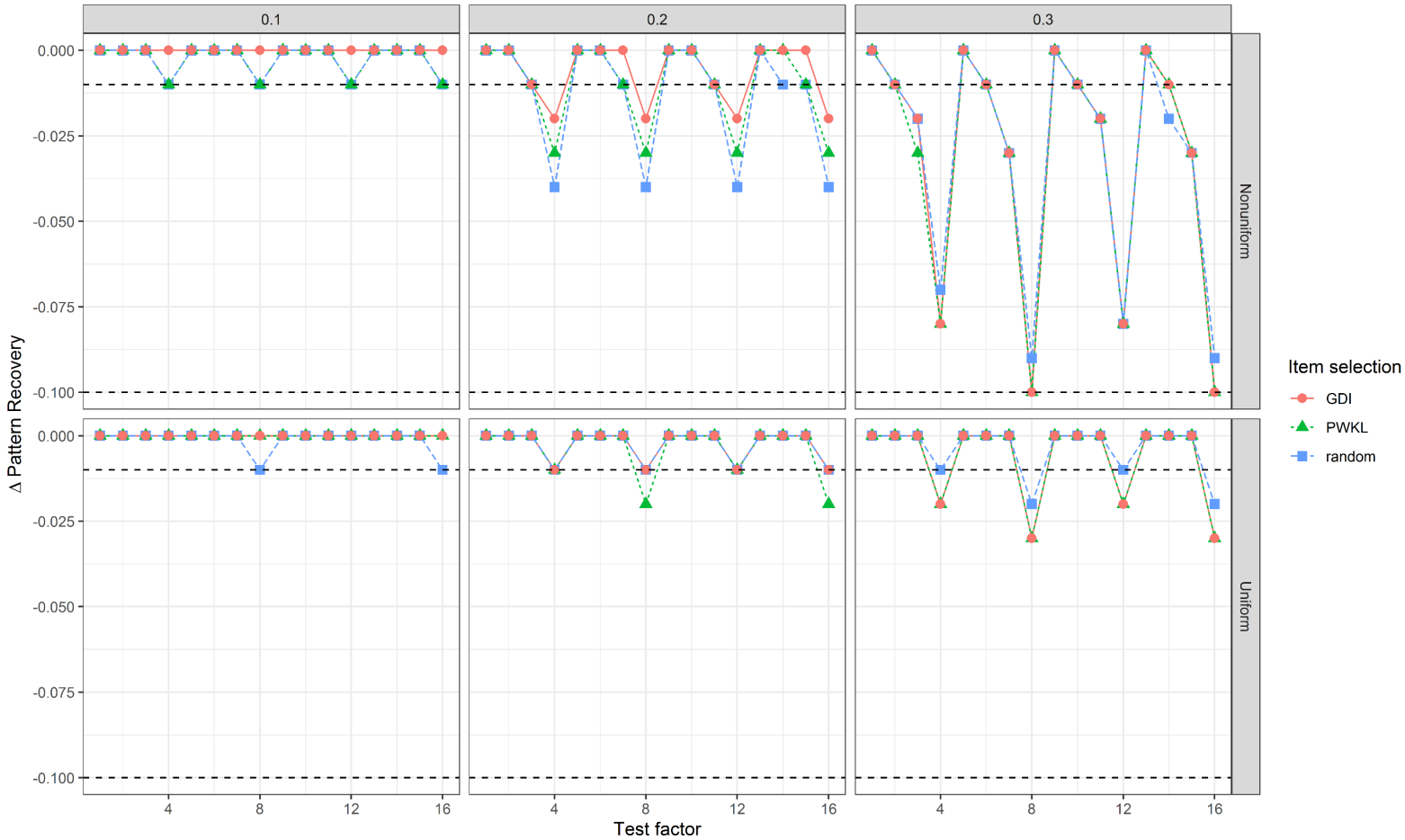


Figure 5-5. Difference in pattern recovery rate between non-DIF and DIF conditions.

Note. Dotted lines are two cut points with 0.01 and 0.1. Note about the x-axis for 1-16 test factors: 1-4, no-DIF, low proportion with small size, high proportion with small size, high proportion with large size for test length of 20 and sample size of 3000; 5-8, same four conditions with test length of 40 and sample size of 3000; 9-12, same four conditions with test length of 20 and sample size of 6000; 13-16, same four conditions with test length of 40 and sample size of 6000.

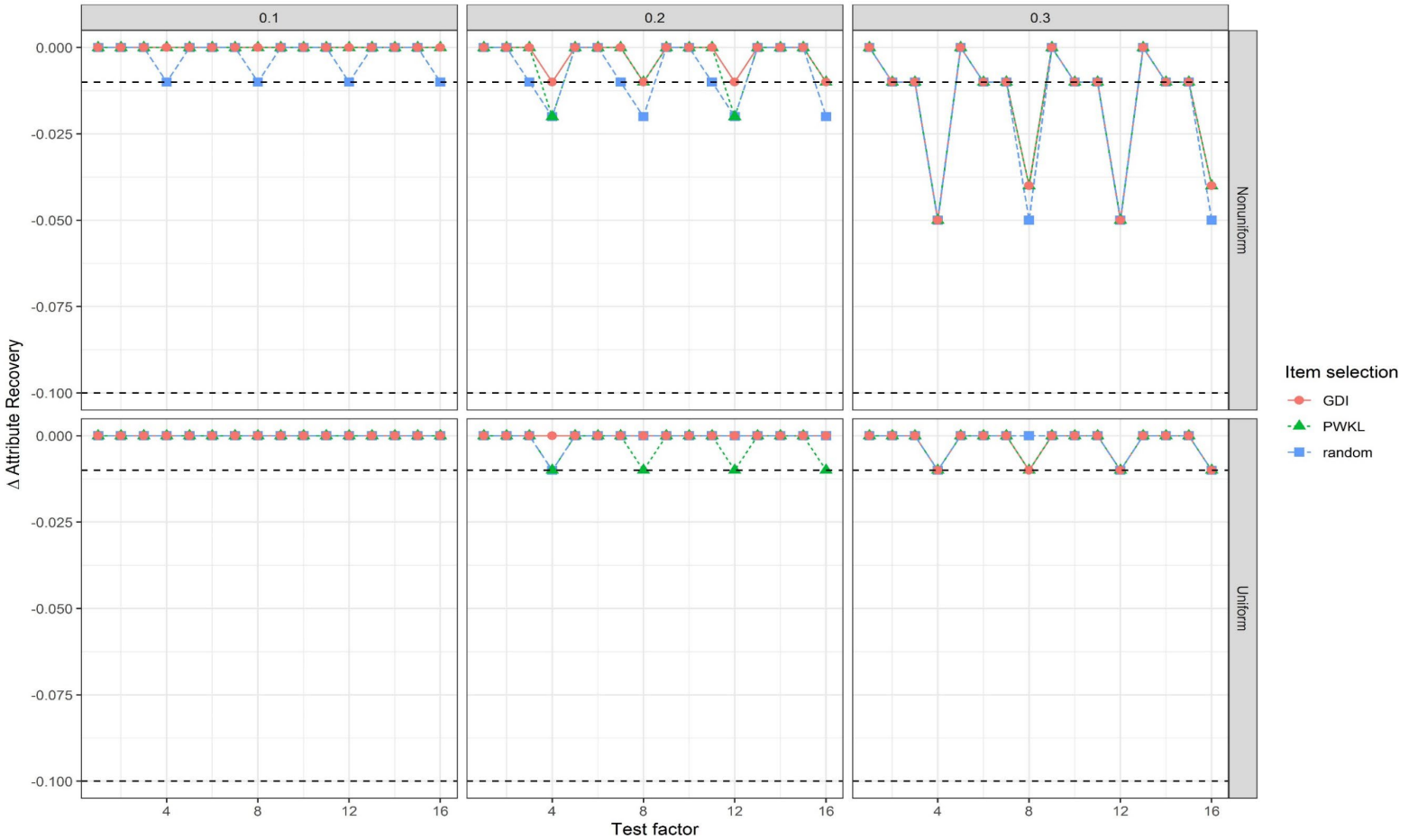


Figure 5-6. Difference in attribute recovery rate between non-DIF and DIF conditions.

Note. Dotted lines are two cut points with 0.01 and 0.1. Note about the x-axis for 1-16 test factors: 1-4, no-DIF, low proportion with small size, high proportion with small size, high proportion with large size for test length of 20 and sample size of 3000; 5-8, same four conditions with test length of 40 and sample size of 3000; 9-12, same four conditions with test length of 20 and sample size of 6000; 13-16, same four conditions with test length of 40 and sample size of 6000.

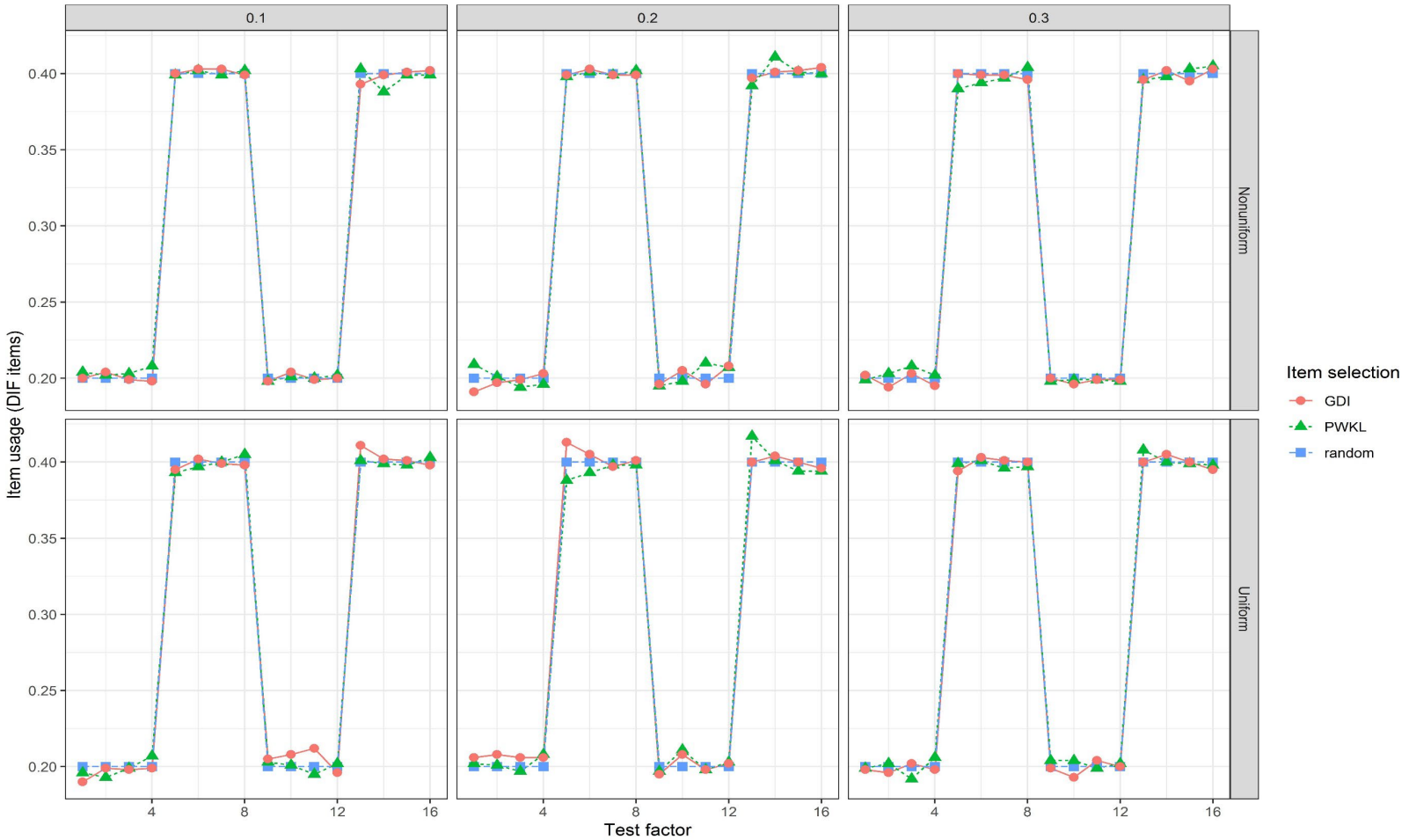


Figure 5-7. Averaged item usage rate for the DIF-contaminated item set across simulated conditions.

Note. about the x-axis for 1-16 test factors: 1-4, no-DIF, low proportion with small size, high proportion with small size, high proportion with large size for test length of 20 and sample size of 3000; 5-8, same four conditions with test length of 40 and sample size of 3000; 9-12, same four conditions with test length of 20 and sample size of 6000; 13-16, same four conditions with test length of 40 and sample size of 6000.

5.5 Discussion

The results of this study showed that in detecting DIF under a variety of conditions in the CD-CAT environment, the range of power and Type-I error rates, and DIF item usage have differed. As a baseline method, random item selection uniformly picks an item to the test despite the item or student's latent attribute distribution. As it turns out, the random selection algorithm is fairly unstable with the occurrence of extremely high or low power rates under some conditions, and extremely low Type I error rates overall. The information-based item selection algorithms have more expected and consistent results in terms of the empirical power of DIF detection and classification accuracy performance.

There are trade-offs as one adopts an information-based item selection algorithm for DIF detection. The PWKL renders better power rates overall, whereas the GDI has slightly more advantage on the accuracy of estimation of attributes and patterns as expected. We found the PWKL item selection algorithm perform consistently better in longer test and the larger number of item attributes in terms of the recovery rate of DIF detection. For longer tests and larger DIF sizes, power rates can be improved by the larger sample size of respondents. The GDI performs the best with short tests and highly discriminating items for the large-size DIF recovery. In other situations, with lower quality items and small DIF size, the PWKL still outperforms the GDI, with a classification accuracy of at least 0.73. While both algorithms have controlled the exposure of DIF-contaminated items on average, they may overuse or underuse some of the DIF items.

The studies of DIF in the CD-CAT setting are critical for test fairness and validity. This study provides an exploratory base for future studies to investigate an appropriate cut-off threshold for the power and Type I error rates for DIF detection in

the CD-CAT setting. Researchers will need to tactfully balance the estimation accuracy, computational efficiency, exposure control of the DIF contaminated set and the concerns of sufficient power in detecting DIF. As Zwick et al (1993) noted, when examinees are given fewer items in an adaptive testing context, every item has more contribution to calculating and estimating the latent ability. Hence, item bias could exert a stronger effect on the ability estimates of the examinees. Our results imply that item bias (DIF) could influence the accuracy of person and item parameter estimates through the item information index. It may create complications that examinees in one group with better performances on previous items receive fewer and less difficult items whereas examinees who have poorer performances on previous items will receive more and more difficult items.

There are a few limitations to this study. Since the Wald detection method assumes examinees' responses to the items presented in the CD-CAT are DIF-free when calculating group population distributions. To robustly detect DIF, we need some item purification methods to be specified (Chang et al. 2011). This study explored the performance of Wald tests under varied conditions, assuming that the Wald test has equally good performance as observed in the CDM-based settings. However, the proposed method may not be the best effective and efficient in the CD-CAT context. Other more generalized CDM, such as the G-DINA model (de la Torre, 2011) and LCDM (Henson, Templin, & Willse, 2009), allow for various probabilities of success for different attribute profiles, which can be further investigated. Researchers for future studies also need to explore non-CDM-based or non-IRT-based DIF detection methods.

So far, many constrained assumptions are imposed, such as fixed item parameters and attributes, relatively long test length conditions, and so on. We have reasons to believe item parameters vary across items, and test conditions vary in the real

setting than in the conveniently simulated. Future improvements to DIF detection methods could explore the Bayesian approach of DIF detection and other item parameter values from a more realistic prior distribution. In addition, simulation conditions need to mimic more realistic settings, such as exploring the relations of a highly diagnostic Q-matrix and its related people skill distribution, for item selection algorithms, and for more simulation conditions.

Reference

Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement*, *40*(2), 129-145.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*, 619-632.

Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: The modified maximum global discrimination index method. *Educational and Psychological Measurement*, *70*(6), 902-913

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179-199.

de la Torre, J., & Douglas, J. (2004). Higher order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333-353. doi:10.1007/BF02295640

de la Torre, J., & Lee, Y. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, *50*(4), 355-373. https://doi.org/10.1111/jedm.12022

Feng, X. (2004). *Statistical detection and estimation of differential item functioning in computerized adaptive testing* [Doctoral dissertation]. University of Columbia, New York.

George, A. C., & Robitzsch, A. (2014). Multiple group cognitive diagnosis models, with an emphasis on differential item functioning. *Psychological Test and Assessment Modeling*, *56*(4), 405.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191-210.

Hou, L., de la Torre, J. D., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*(1), 98-125.

Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment*, *15*(3), 1-7.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258-272.

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement*, *39*(3), 167-188.

Lei, P., Chen, S., & Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement, 43*(3), 245-264. http://www.jstor.org/stable/20461826

Li, F. (2008). *A modified higher-order DINA model for detecting differential item functioning and differential attribute functioning* [Doctoral dissertation]. University of Georgia, Athens.

Li, X., & Wang, W. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement*, *52*(1), 28-54. https://doi.org/10.1111/jedm.12061

Lin, C., & Chang, H. (2018). Item selection criteria with practical constraints in cognitive diagnostic computerized adaptive testing. *Educational and Psychological Measurement*, *79*(2), 335-357. https://doi.org/10.1177/0013164418790634

Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, *93*(14), 1-26.

Ma, W., Terzi, R., & de la Torre, J. (2021). Detecting differential item functioning using multiple-group cognitive diagnosis models. *Applied Psychological Measurement*, *45*(1), 37-53. https://doi.org/10.1177/0146621620965745

Nandakumar, R., & Roussos, L. (2004). Evaluation of the CATSIB DIF procedure in a pretest setting. *Journal of Educational and Behavioral Statistics*, *29*(2), 177-199.

Paulsen, J., Svetina, D., Feng, Y., & Valdivia, M. (2020). Examining the impact of differential item functioning on classification accuracy in cognitive diagnostic models. *Applied Psychological Measurement*, *44*(4), 267-281. https://doi.org/10.1177/0146621619858675

Piromsombat, C. (2014). *Differential item functioning in computerized adaptive testing: can CAT self-adjust enough?*. University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/163281.

Qiu, X. L., Li, X., & Wang, W. C. (2019). Differential item functioning in diagnostic classification models. In *Handbook of diagnostic classification models* (pp. 379-393). Springer, Cham.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197-207.

Rogers, S. J., & Swaminathan, H. (1993). A comparison of logistic regression and MH procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105-116.

Svetina, D., Feng, Y., Paulsen, J., Valdivia, M., Valdivia, A., & Dai, S. (2018). Examining DIF in the context of CDMs when the Q-matrix is Misspecified. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00696

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological methods*, *11*(3), 287.

Wald, A. (1943). On stochastic limit and order relationships. *The Annals of Mathematical Statistics*, *14*(3), 217-226. https://doi.org/10.1214/aoms/1177731415

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement*, *73*(6), 1017-1035.

Wang, C., Chang, H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement*, *48*(3), 255-273.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied psychological measurement*, *6*(4), 473-492.

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups. *Educational and Psychological Measurement*, *73*(3), 532-547. https://doi.org/10.1177/0013164412464875

Xu, X., Chang, H., & Douglas, J. (2003, April*). Computerized adaptive testing strategies for cognitive diagnosis* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Montreal, Canada.

Xu, G., Wang, C., & Shang, Z. (2016). On initial item selection in cognitive diagnostic computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 291-315.

Zhang, W. (2006). *Detecting differential item functioning using the DINA Model* [Doctoral dissertation]. University of North Carolina, Greensboro.

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

Zwick, R., Thayer, D. T., & Wingersky, M. (1993). A simulation study of methods for assessing differential item functioning in computer-adaptive tests. *ETS Research Report Series*, *1993*(1), i-110.

Chapter 6. Conclusion

Diving into the psychometric and analytic details for cultural validity in contextualized assessment and adaptive test infrastructure, we have highlighted the critical aspects of culturally responsive and equitable testing (CRET) for addressing equity, fairness, and diversity issues in modern assessment. In conclusion, we hope to summarize how those psychometric considerations are necessary for future research by connecting each study with existing theoretical and practical implications and discussing the limitations and future lines of research.

Chapter 3 discussed ways in which cultural context represented in contextualized items can leverage the knowledge construction of testing. It is a starting point for understanding context features while setting the intentions explicitly for an equitable, fair, and culturally responsive agenda. It calls the larger educational community toward a theory and program of equitable measurement. Informed by cultural theories, our results show positive evidence of understanding cultural contexts to investigate measurement invariances in the quality of items. Though the regression coefficients on the item-level have limitations due to the small sample size of contexts, the provided approach opens the possibility of examining item contexts through mixed methods of analyses. The study brings out further thinking on whether psychometric analyses go beyond student test performance to respecting culturally diverse students' cultural knowledge and prior experiences. More discussion is needed to examine contextualized measurement for its cultural validity and fairness claims. Future research can explore the conceptual tool of cultural validity or conduct cognitive interview studies to delineate meaningful bridges between item performance and students' actual understanding.

Next, chapter 4 systematically highlights different aspects, codes, and psychometric evidence of item contexts linking to student performance. Our analytical context coding approach is integrated through multilevel item response modeling to account for clustered context effects and student cultural backgrounds. In probabilistic terms, we demonstrate isolating and experimenting with context effects on the subtestlet level for disentangling construct-irrelevant variances. The findings reveal the potential effects of the context's sociocognitive and sociocultural functions on students' success in responses. Methodologically, the Bayesian multilevel explanatory item response model describes an approach to adjust potential item bias in the meanwhile accounting for dependencies in the hierarchical structures of repeated measures in student, item, and contexts.

Chapter 4 shows that context coding by experts enables probing the influence of context or heterogeneity across contexts and students and provides an exemplary item context calibration for psychometrics. The results from both chapter 3 and chapter 4 should be taken to guide and inform item development and calibration.

Lastly, chapter 5 delves into the item bias detention method itself, exploring differential item functioning for complex testing such as CD-CAT. This type of testing demonstrates its advantage by creating item-level and learner-level attribute profiles. Specifically, our studies show that good item selection algorithms are sufficient to recover persons' patterns and attributes under various conditions of DIF presence. Knowledge gained about the DIF phenomena and detection performance in CD-CAT will help create better designs for adaptive testing. As pointed out, CD-CAT can incorporate contextualized items, and the study can support a fair development of future contextualized adaptive testing.

6.1 Future Direction

Taken together, the chapters demonstrate from theory to measurement models, from items to tests, the advantages for culturally responsive, equitable, fair, and individualized testing technology. However, more studies are needed to advocate and demonstrate the validity and equity claims for CRET. Specifically, this work points to three directions for future research.

First, the ideas from chapter 2 should motivate and encourage more detailed conceptualizations and designs transform culturally responsive practices. Studies need to identify and implement innovative ways for testing to make reference to culture, as well as allow students to draw on their social and cultural literacies. Second, while laying out the psychometric and analytic details for individualized and adaptive testing, we find the field of adaptive testing is still missing efforts for including context designs and attributes. For future studies, we hope more research and practice in adaptive testing can adaptively select item contexts instead of items so that item contexts can be responsive to culturally diverse learners' frames of reference, needs, and testing styles. Specifically, a responsive item context can have context stimuli relating to students' cultural identities and performance styles to make testing encounters more relevant and engaging. In addition, methods for examining fairness at the item context level in adaptive testing are still needed. Lastly, measurement research can utilize machine learning to automate bias detection and adaptive algorithms. While testing should aim to approximate and optimize the measurement of a person's latent ability, we need culturally responsive principles to be operationalized through adaptive testing via item configuration, selection, administration, estimation, and evaluation. A rich layer of data including process data should be used for making validity arguments and score inferences.

APPENDIX 1.A

Equitable Item Context -- Review Protocol

(Simplified Version)

Part 1: *Cultural Equity (CE) rubrics*

1. Item context has elements that show explicit prejudice, stereotypes or biases against certain Group Identity (GI) or their communities of practice.
2. Item context has elements that (un)intentionally privilege certain GI.
3. Item context has elements of language (text and visual) that assign agency or proactivity tocertain GI over others represented.
4. Item context has elements that perpetuate or reinforce the status quo of social inequality.
5. Implicitly, Item context has elements that shows culturally insensitive or identity-based information that could potentially trigger negative mental, psychological, or emotional effectsto certain GIs (as readers).

Part 2: *Equitable Context Representation (ECR) rubrics*

6. Item context is relatable and familiar for all students from diverse cultural backgrounds.
7. Item context can be applied to students' and their communities' everyday life, hobbies, interests,and benefits as real-world, everyday life scenarios.
8. Item context is constructive in scaffolding scientific thinking and skills.
9. Item context captures a coherent story or message by appropriate levels of modality, dimensions,and details of information for multicultural learners.
10. Item context has language and design, in principle, is accessible to all students, includingemergent bilingual students.

Part 3: *Knowledge Construction (KC) rubrics*

11. Item context has elements that transmit knowledge or cultural assumptions from the groups in power with agenda.
12. Item context has elements that represent its knowledge structure as: linear, abstract, schematic,theory-based, oversimplified, monotonous, or field-independent.
13. Item context has elements that represent its knowledge structure as: dynamic, Complex,multidimensional, context-based, or field-

dependent.

14. Item context has elements that represent its epistemology as: transformative, critical, culturallysustaining, inclusive, egalitarian, strong-objectivity, or for the greater social good.
15. Item context has elements that represent its epistemology as: suppressive, partisan, exclusive,colonial-oriented, hegemonic, materialistic, or deeply oppressive.

APPENDIX 1.B

Coder's Procedure

Step 1	Identify resources: text, tables, charts, figures, and images
Step 2	Identifying the five essential aspects of context information by asking the keywords of "when, where, who, what, and how" of a context.
Step 3	Sort out the above elements into a categorical identification of Group Identity (GI), such as race/ethnicity, gender, and class.
Step 4	Identify structures and locations of contextualized items, including community space, cultural heritage, sociohistorical sites;
Step 5	Identify inherent values, culture, knowledge, and ideological domains manifested in item context.
Methods	Discourse analysis: Apply modality analysis, critical analysis, and cultural sensitivity analysis to analyze the relationships between identified elements in the item context and their sociocultural meanings; Reflective analysis: reflect on the complex interplay of factors and derive meanings embedded in item context, for its inherent agenda in relation to sociocultural values and assumptions; Power structure analysis: analyze power relations in the text; aware of power relations in social relations, education, and knowledge production in the society.
Step 6	Consolidate initial codes by consulting the codebook, closely read the rating tips, coding rules and examples.

APPENDIX 2.A

Table 1

Descriptive statistics of DECISA Item of Booklet 1

Booklet 1 Row	Missings	Mean	SD	Skew	Item Difficulty	Item Discrimination	α if deleted
bus.1.1	0.48%	0.33	0.47	0.7	0.33	0.25	0.51
bus.1.2	0.36%	0.11	0.32	2.47	0.11	0.41	0.49
bus.3.1	0.36%	0.45	0.5	0.18	0.45	0.16	0.53
bus.3.2	0.48%	0.13	0.34	2.18	0.13	0.01	0.55
sled.2.1	1.21%	0.2	0.4	1.48	0.2	0.04	0.55
sled.2.2	1.21%	0.27	0.44	1.05	0.27	0.14	0.53
sled.4.1	1.21%	0.37	0.48	0.54	0.37	0.2	0.52
sled.4.2	1.33%	0.35	0.48	0.64	0.35	0.27	0.5
box.1.1	0.12%	0.26	0.44	1.11	0.26	0.25	0.51
box.1.2	0.36%	0.13	0.34	2.17	0.13	0.4	0.49
box.2.1	0.61%	0.24	0.43	1.24	0.24	0.14	0.53
box.2.2	0.36%	0.47	0.5	0.13	0.47	0.07	0.55
box.3.1	0.24%	0.66	0.47	-0.68	0.66	0.14	0.53
box.3.2	0.36%	0.46	0.5	0.15	0.46	0.23	0.51
box.4.1	0.36%	0.42	0.49	0.33	0.42	0.2	0.52
box.4.2	0.12%	0.3	0.46	0.85	0.3	0.21	0.52
hockey	0.36%	0.32	0.47	0.79	0.32	0.23	0.54
carpush	0.61%	0.16	0.37	1.87	0.16	0.11	0.55
cabinet	0.73%	0.11	0.31	2.56	0.11	0.04	0.56
Mean inter-item-correlation=0.063 · Cronbach's α=0.556							

Table 2

Descriptive statistics of DECISA Item of Booklet 2

Row	Missings	Mean	SD	Skew	Item Difficulty	Item Discrimination	α if deleted
			Booklet 2				
cart.1.1	0.00%	0.25	0.44	1.14	0.25	0.3	0.54
cart.1.2	0.25%	0.17	0.38	1.76	0.17	0.38	0.53
cart.2.1	0.12%	0.2	0.4	1.54	0.2	0.08	0.57
cart.2.2	0.00%	0.42	0.49	0.31	0.42	-0.01	0.59
cart.3.1	0.12%	0.71	0.45	-0.93	0.71	0.17	0.56
cart.3.2	0.12%	0.5	0.5	-0.01	0.5	0.27	0.54
cart.4.1	0.25%	0.47	0.5	0.11	0.47	0.21	0.55
cart.4.2	0.62%	0.33	0.47	0.73	0.33	0.28	0.54
sled.1.1	0.50%	0.28	0.45	0.99	0.28	0.25	0.55
sled.1.2	0.87%	0.14	0.35	2.05	0.14	0.34	0.54
sled.3.1	0.50%	0.49	0.5	0.04	0.49	0.21	0.55
sled.3.2	0.99%	0.3	0.46	0.9	0.3	0.25	0.55
bus.2.1	1.61%	0.38	0.49	0.5	0.38	0.05	0.58
bus.2.2	1.61%	0.37	0.48	0.53	0.37	0.13	0.57
bus.4.1	1.74%	0.43	0.5	0.27	0.43	0.23	0.55
bus.4.2	1.74%	0.42	0.49	0.32	0.42	0.27	0.54
hockey	1.12%	0.35	0.48	0.64	0.35	0.31	0.58
carpush	1.24%	0.18	0.38	1.67	0.18	0.23	0.6
cabinet	1.49%	0.12	0.33	2.29	0.12	0.06	0.61
Mean inter-item-correlation=0.078 · Cronbach's α=0.610							

APPENDIX 2.B

Two-step regression approach of modeling item difficulty as an outcome variable

Step 1: Item statistics: see Appendix 2.A

Step 2: Using Hierarchical Linear Modeling (HLM): Process and Results

Context variables:

Table 1

A description of key features

Variable details					
Block	Categories/Dimensions	Question/ Variable	Question Asked	Type of Response	Options
General	Characterizing Context	Q16	Item has context?	One option	No
Item	Item Characteristics	Q90	Illustration in item?	One option	Yes
Context	Context Characteristics	Q20	Context Length	One option	One sentence
Context	Context Characteristics	Q21	Reading Characteristics	One option	Context has everyday language, **common usage** of **language**
Context	Context Characteristics	Q22	Reading Load	One option	Low
Context	Context Characteristics	Q23	Ease of Reading	One option	Yes, only one reading is necessary
Context	Context Characteristics	Q26	Abstraction	One option	Main ideas in the context are **concrete**
Context	Context Characteristics	Q27	Focus	One option	**Focused. It is unlikely that the content activates irrelevant information**
Context	Context Characteristics	Q35	Sequence of stages/events	One option	**No**, the context does not include stages/phases.
Context	Bias	Q38	Cultural Bias	One option	The context **is little or not biased.**
Context	Bias	Q39	Sociolinguist-Unfamiliar	One option	**No, most likely the language is familiar to all or most of the students**
Context	Single Object	Q48	Plane of Motion	One option	Horizontal
Context	Single Object	Q52	Type of Motion for Context	One option	Yes, the motion is **specified**; it can be clearly identified
Context	Single Object	Q54	Type of force applied	One option	**No force is represented**

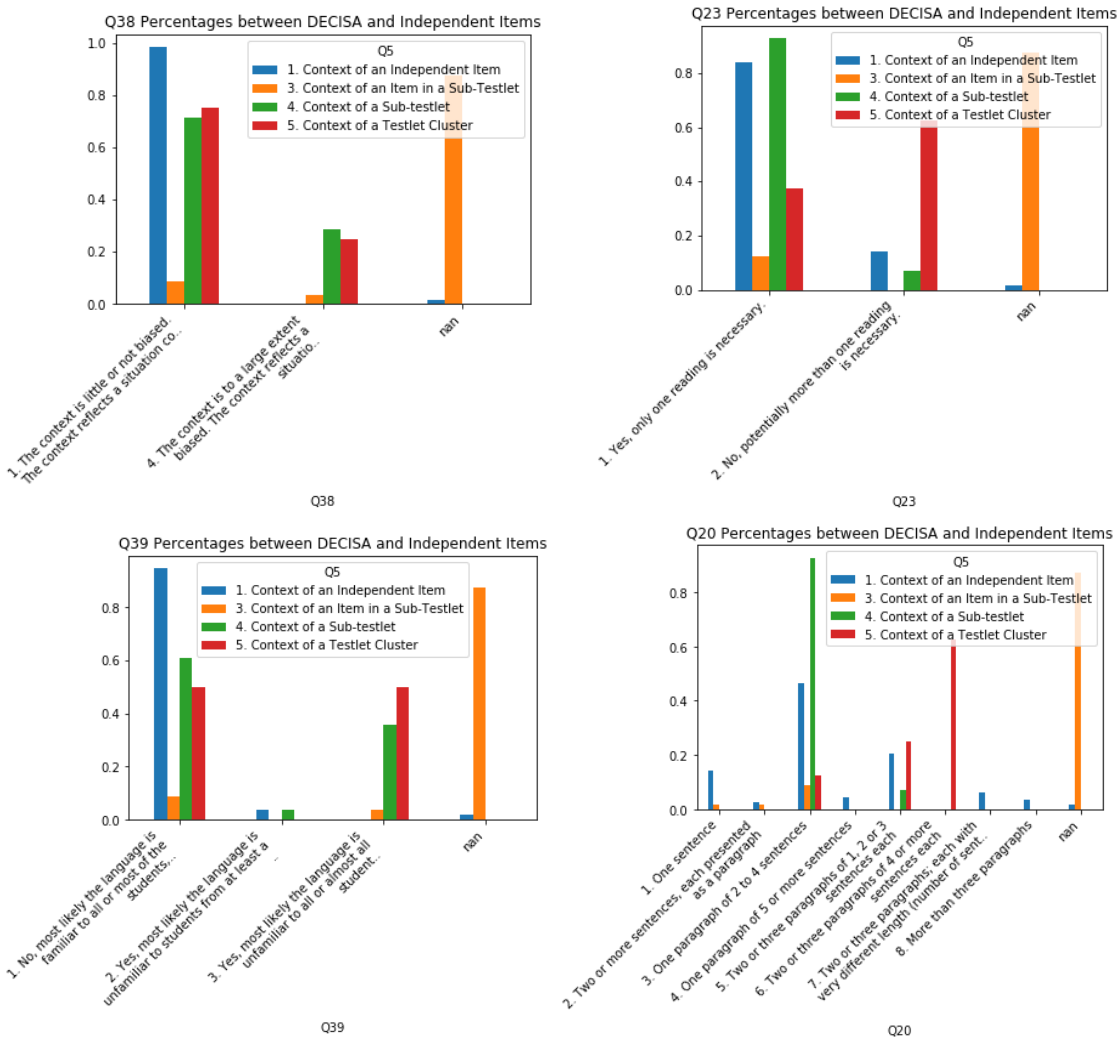


Figure 1. Frequency statistics about context variables across context levels.

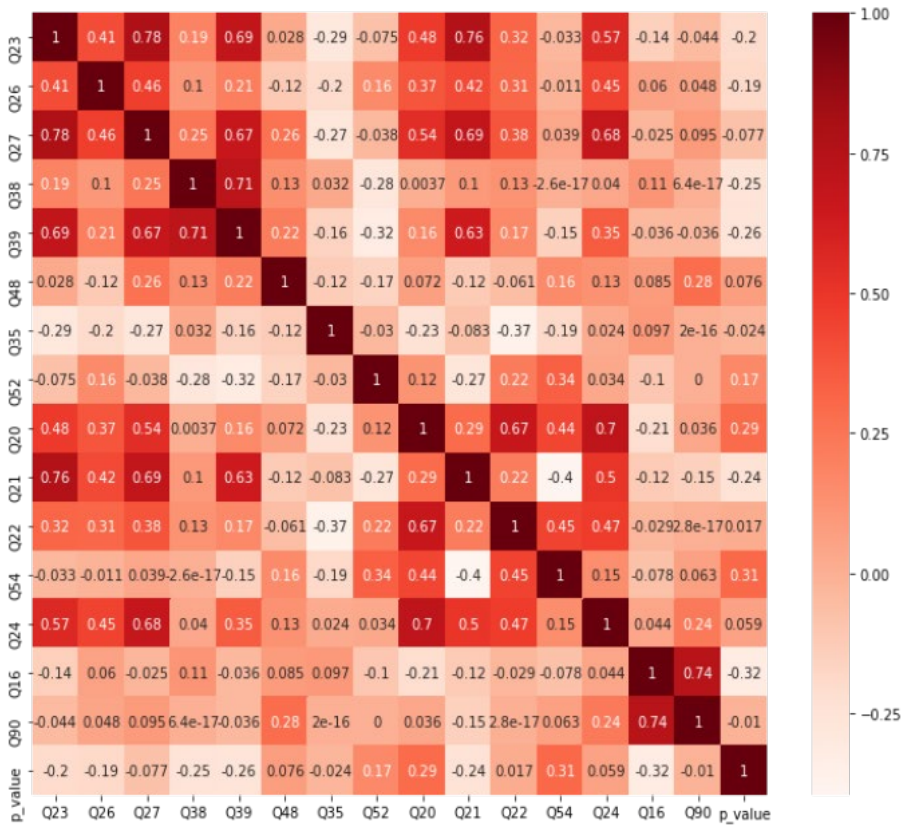


Figure 2. Correlation between ordered variables.

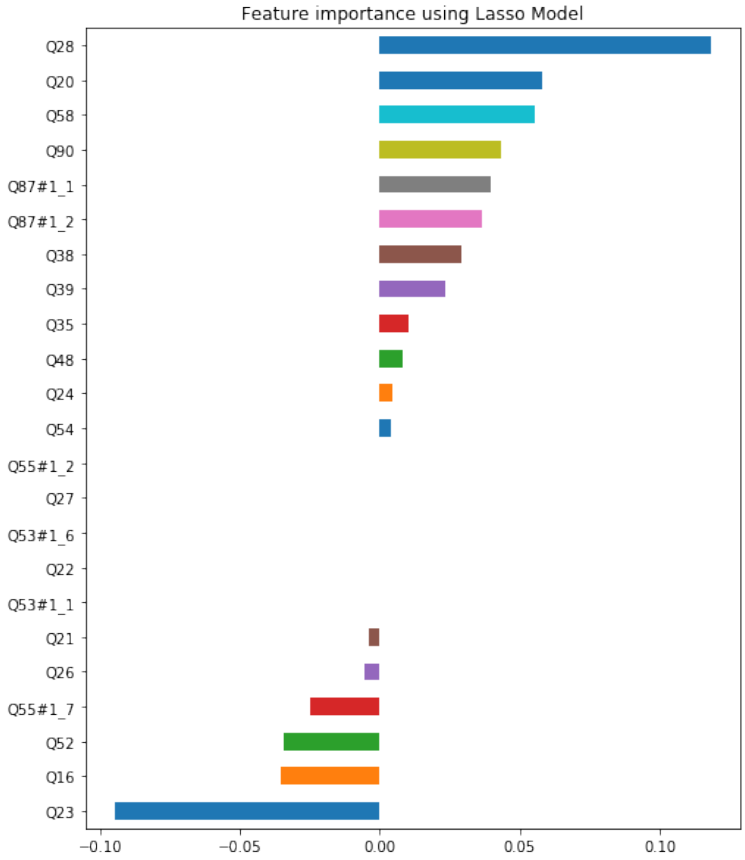


Figure 3. Feature selection based on LASSO regression.

Summary of effects of item and context characteristics on item parameters

The HLM approach essential treats context feature variables from DECISA codes as Level 2 predictors, with a sample size of n=28. Level 1 has item feature variables (data is sparse, so only 4 variables are tested) as predictors, with sample size n= 56. If adopting a three-level structure, level 3 only has an intercept and uses no context feature variables for predictors, despite we coded them for testlet context data. The reason is that sample size only equals to 7-- too small degrees of freedom for any inferential test in HLM. So the part of information on testlet level is ignored. For Rasch scores, we recommend a two-level HLM with the same model specification as above. We choose two-level model because the level-three variance is no longer statistically significant.

For estimating the direction of effects on outcome variables, both p-value and Rasch models give the same result, meaning they always predict same positive or negative effect of variable on the outcome.

For parameter estimates such as their significance level, the CTT item difficult and IRT Rasch parameter do give slightly different results. For example, the variable Q28 Genetic setting is no longer statistically significant for Rasch model. We recommend using IRT Rasch model in this case mainly because CTT item difficulty tend to be more sample-dependent and in this case not normally distributed, whereas IRT parameter reflects more general relationship.

The table below displays the two-level random intercepts model for Rasch on the right, but one that adds a General Context effect at level 3 on the left to further model the level 1 intercept for p-value scores. Approximate Pseudo-R2 for 3-level p-value model is 0.3801. Overall the predictors in the model explained approximately 38% of the variance in p-value. Approximate Pseudo-R2 for 2-level Rasch model is

0.4576, which means the predictors in the model explained approximately 46% of the variance in Rasch.

Table 2

Multilevel model main effect estimates for random intercept model with different within- and between-group regressions

	p-value		Rasch	
Fixed Effect	Coefficient	S.E.	Coefficient	S.E.
Intercept	-0.207	0.095	**3.183**	0.296
Ease of Reading (Q23)	-0.095	0.060	**0.429**	0.105
Sociolinguist-Unfamiliar (Q39)	**0.052**	0.023	**-0.228**	0.065
Context Length (Q20)	**0.190**	0.026	**-0.977**	0.069
Genetic setting (Q28_G)	**0.104**	0.046	-0.148	0.128
Applied force (Q55_force)	**-0.094**	0.029	**0.529**	0.116
At rest (Q53_motion)	**-0.108**	0.036	**0.447**	0.133
Constant speed (Q53_motion)	**-0.128**	0.032	**0.710**	0.103
Illustration in item (Q90)	**0.061**	0.027	**-0.371**	0.208
Force, Predict (Q87)	0.013	0.019	-0.051	0.076
Random Part	Parameters	S.E.	Parameters	S.E.
Level-three variance	0.003	0.055		
Level-two variance	<0.01	0.001	<0.01	0.005
Level-one variance	0.078	0.006	0.197	0.444
Deviance (-2LL)	-115.798		68.022	
No. Params Est	13		12	

Note. $N = 56$ items within 28 Subtestlet Context groups and 7 General Context groups; Maximum Likelihood and Robust Standard Error estimates shown. Q16, Q90, and Q87 are level 1 predictors, while the rest are level 2 predictors. HLM7 program used to estimate models; between within method for *df* used for coefficient t -tests. Coefficients with $p < .05$ are bolded.