# Developing a Technology-Enhanced Solution to Language Inequality in English-Based Mathematics Tests

by

Kevin Close

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved August 2021 by the
Graduate Supervisory Committee:

Yi Zheng, Co-Chair
Audrey Amrein-Beardsley, Co-Chair
Kate Anderson

ARIZONA STATE UNIVERSITY

December 2021

# ABSTRACT

In this mixed-methods study, I sought to design and develop a test delivery method to reduce linguistic bias in English-based mathematics tests. Guided by translanguaging, a recent linguistic theory recognizing the complexity of multilingualism, I designed a computer-based test delivery method allowing test-takers to toggle between English and their self-identified dominant language.

This three-part study asks and answers research questions from all phases of the novel test delivery design. In the first phase, I conducted cognitive interviews with 11 dominant Mandarin Chinese and 11 dominant Spanish speaking undergraduate students while taking a well-regarded calculus conceptual exam, the Precalculus Concept Assessment (PCA). In the second phase, I designed and developed the linguistically adaptive test (LAT) version of the PCA using the *Concerto* test delivery platform. In the third phase, I conducted a within-subjects random-assignment study of the efficacy the LAT. I also conducted in-depth interviews with a subset of the test-takers.

Nine items on the PCA revealed linguistic issues during the cognitive interviews demonstrating the need to improve the linguistic bias on the test items. Additionally, the newly developed LAT demonstrated evidence of reliability and validity. However, the large-scale efficacy study showed that the LAT did not appear to make a significant difference in scores for dominant speakers of Spanish or dominant speakers of Mandarin Chinese. This finding held true for overall test scores as well as at the item level indicating that the LAT test delivery system does not appear to reduce linguistic bias in testing.

Additionally, in-depth interviews revealed that many students felt that the linguistically adaptive test was either the same or essentially the same as the non-LAT version of the test. Some participants felt that the toggle button was not necessary if they could understand the mathematics item well enough. As one participant noted, “It's math, It's math. It doesn't matter if it's in English or in Spanish.” This dissertation concludes with a discussion about the implications for test developers and suggestions for future direction of study.

# DEDICATION

To my wife, who provided perspective, support, encouragement, and humor every step of the way.

To my parents, who provided a passion for education and who opened their home to me during a pandemic.

To my advisors and mentors at Arizona State and Utah State, for modeling the pairing of positive attitudes with strong scholarship.

To my cohort and classmates, for sharing their time, expertise, and friendship on this journey.

Lastly, to my former students, who inspired me to pursue my own academic dreams.

# ACKNOWLEDGEMENTS

First of all, I would like to thank my dissertation committee members: Yi Zheng, Audrey Amrein-Beardsley, and Kate Anderson. Without your guidance, mentorship, and role modeling I would not have made it so far in this journey. Kate, your humor and openness in work and in play inspires me. Audrey, your brilliance and can-do attitude pushed me to write more, think harder, and contemplate how to best create impact with my work inside and outside of the academy. Yi, there is more to say than I can write in this section, but I just want to thank you for your encouragement, guidance, and kindness throughout this journey. Thank you all.

Outside of my committee, I need to thank the other members of the Arizona State University community. In particular, I am grateful to Punya Mishra, who not only provided spaces for me to develop ideas, but who also served as a mentor to me outside of my committee. Thank you, as well, to Katie Bernstein, for both being an incredible co-author and for sharing job hunt advice. I also need to thank Clarin Collins for always being there for advice, help, or just to crack a joke at the expense of the Chicago Bears. Thank you, too, to Dr. Marilyn Carlson for sharing your Precalculus Concept Assessment with me for use in this study. To another Carlson, David Carlson, thank you for being there at a moment's notice to chat. To Pamela Kulinna, I so enjoyed teaching with you. To Billie Cruz, thank you so much for your hard work. Lastly, to my cohort members and classmates, thank you. A special thank you to my writing buddy, Megan Troxel Deeg, and my Talking About Design groupmates (too many to list!). There are so so many others I would like to thank at Arizona State, if I had the space.

Outside of Arizona State University, I want to thank Pamela Paek. You were such a wondeful mentor. Thank you to the whole ACT Inc. team, it was great to spend the summer working with you all. Thank you, as well, to my mentors and colleagues at Utah State University: Jody Clarke-Midura, Sarah Brasiel, and Taylor Martin. Thank you for showing me the ropes in my first few years.

Lastly, Yueying Li, I need to further acknowledge your support throughout this process. I cannot imagine a better person with whom to share this journey. Thank you.

# TABLE OF CONTENTS

APPENDIX Page

## LIST OF TABLES

Table Page

# LIST OF FIGURES

Figure Page

# CHAPTER 1

# INTRODUCTION

“Any test is to some degree a test of English language proficiency and may be [an] invalid [and] unreliable measure of English language learners' academic proficiencies," stated the American Educational Research Association (AERA) in 1985 (AERA, 1995, as cited in Solano-Flores et al., 2001, p. 50). In other words, a student’s English language proficiency presents a hurdle for accurately measuring other proficiencies. Yet, overcoming such a hurdle is paramount to creating valid and reliable tests for all subgroups.

Recently, the National Council on Measurement in Education (NCME) released a position statement on testing multilingual learners in which they stated that, “test developers can improve fairness by enabling response methods that enable students to demonstrate their abilities on the target construct or domain” (NCME, 2020, p. 1). In other words, proper test development adjusting for English language proficiency is an issue of fairness, where fairness is defined as a fundamental validity issue centered on a combination of bias, accessibility, and universal design (AERA et al., 2014). Fundamentally, fairness ensures that an assessment yields valid inferences in various contexts for various individuals meaning that tests should not be utilitarian (e.g., work the most well for the most people), but rather, tests should be exhaustive (e.g., work the most well for all groups of people).

However, even defining groups of people has pitfalls. Designating certain test-takers to certain bins (e.g., binning test-takers by language) can fail to incorporate complex linguistic backgrounds. Scholars who study anti-essentialism and intersectionality emphasize that many factors play significant roles, and interact in significant ways, in the formation of an identity or ways of experiencing the world (Crenshaw, 1990). For example, Feuchtwang (1990) critiqued multiculturalism by examining the simplistic underlying assumptions:

> The premise of sorting populations by ethnic origins according to presumed cultural essence is that culture is a community of deep-seated values. For values one may also read social roles and meanings, or customs and traditions. But what makes cultural origin a category of population is the additional assumption that a culture is a community of original identity, to which individuals belong by birth. By the common sense of being and belonging which sets the tone of this cultural recognition, all those born into a community absorb and ineradicably sediment within themselves its customary ways of thinking, feeling and being. Even if they do not so identify themselves, they are properly identified with that community, whatever subsequent layers of other cultures they may have absorbed to cover the original sediment. (p. 4)

In a linguistic example, prescribing a label to students as bilingual and multilingual does not fully capture the complex phenomenon of multilingualism wherein students may speak various forms of various languages in various contexts. Such students may have learned some content in one language, other content in a different language, and feel most comfortable speaking a third language informally with family or friends.

However, in the face of such an issue (e.g., accounting for the complexity of language among students in something meant to be standardized like a placement test), I pursue a potential and practical step forward. Finding solutions to testing problems are fundamentally practical concerns because tests are artifacts that have real world impact. Not only do test designers need to have the theoretical knowledge to recognize how to improve a test (e.g., reduce linguistic bias issues for multilingual test-takers), they need to figure out practical ways to do it (e.g., limit the word count on test items, change passive verb tenses into simple active verb tenses, provide glossaries). Poorly worded test items, technical glitches in an online test, unclear images, bad color schemes, and other concerns can all cause some students to perform differently than others.

For students, the quality and rigor of test design has real world impacts. Receiving a higher score can mean landing on one side or the other of a cutoff line. Receiving 149 instead of a 150 may mean the difference between being considered for admission, for acceptance to an honors college, or for having a scholarship application read (even if that falls within two times the standard error of measurement from the cutoff line 150). Even more so, authors of recent economic and human resources studies have shown that scores provide personal feedback to students which in turn affect their decision to either pursue more school (e.g., attend college) or enter the workforce. Papay et al. (2011, 2016) found that test labels such as “advanced” or “proficient” affected student college-going decisions significantly for urban, low-income students. Therefore, even small improvements to accuracy and fairness that can tip a student to one side or the other of hard cut-offs, such as how they are labeled by a test, has implications for important life decisions.

In this dissertation, I address how English language proficiency presents a hurdle for accurately measuring other proficiencies by building, with a three-phase mixed-methods empirical study, a potential solution called a linguistically adaptive test (LAT). The idea of LAT is a) informed by a recent linguistic theory, called *translanguaging*, and b) leverages adaptive testing to change the language of the test for different students in such a way that the students can use their full multilingual repertoire to interact with test items. In the second chapter, I summarize the steps that others have taken to create better tests for multilingual learners by touching not only on research about test accommodations, but also about translanguaging.

**The Linguistically Adaptive Test (LAT)**

In this dissertation, I built and tested a test delivery mechanism that leveraged the computerized-adaptive testing (CAT) technology to operationalize translanguaging theory, thereby allowing test-takers to change the language of test items freely during a test. I call this novel test delivery mechanism a linguistically adaptive test (LAT), a term coined for the purpose of this dissertation.

A LAT is a test design that attempts to reduce language bias in tests by taking advantage of the affordances of CATs. With CATs, test items no longer follow the typical standardized format in which every student receives the same items in the same order (Wainer et al., 2001). Instead, the test learns as the student takes the test, and provides items, from a standardized test bank, that are more appropriate to the knowledge level of the student. This results in tests that arrive at the appropriate score with fewer test items (Gershon, 2005). Having fewer items leads to less testing time, less fatigue, and, ultimately, less cost. The benefits of CAT testing are compelling many large testing companies to adopt the method around the world. However, the affordances of CATs may extend beyond simply speeding up an assessment. The key point is that many students are no longer receiving the same test. The test items they receive are often personalized to provide a better testing experience and a more accurate test (Casaletto et al., 2016).

LATs extend the boundary of CAT to allow tests to adapt the *language* the items are presented in. During the test, test-takers can freely change the language of the items by clicking a toggle button. For example, a test-taker who states that that they typically speak Spanish at home, but English at school, may receive items to a mathematics test in English with the option to toggle back and forth into Spanish. The purpose is to remove construct-irrelevant barriers from subgroups who may be disadvantaged by the format of delivery for the test item.

On the other hand, the LAT treats multilingualism as a resource by allowing students to transition from one language to another for each item (Hornberger & Link, 2012). In this way, LATs are meant to incorporate translanguaging theory, which will be unpacked in the next chapter, into a type of test delivery meant to increase fairness for test-takers with various linguistic backgrounds.

## Definitions of Terminology

To provide better clarity for sections that follow, the following list of definitions clarifies some of the terms in the dissertation.

### *Definition of Multilingual Learner*

I use the term “multilingual learner” to describe students often called English language learners (ELL), English as a Second Language (ESL) students, bilingual students, limited English proficient (LEP) students, or emergent bilingual learners (García, 2009). However, using the term “multilingual learner” emphasizes that the student has a surplus of language skills (e.g., the ability to speak two or more languages) instead of emphasizing a deficit of English language proficiency. Names matter. Using the term ‘multilingual learner’ is a signal to focus away from linguistic limits or deficiencies instead reframing linguistic differences as something outside of any narrative about deficiencies or limits. Also, by choosing to emphasize multilingualism instead of bilingualism, I am deliberately recognizing that many students speak multiple languages and multiple forms of those languages.

### *Definition of Construct*

In the education and psychology disciplines, a construct, also called a latent variable, is an unobserved characteristic of a person meant to be measured (Messick, 1989a). Researchers consider constructs to be an underlying trait that cannot be *directly* measured. To measure constructs, researchers use several indicators or observable characteristics. For example, researchers measure the construct of depression by examining the answers of a several-question depression scale because there is no direct way to measure depression.

### ***Definition of Validity***

The classic concept of validity started out with a simple idea: tests need to prove that they measure what they claim to measure. According to DeVellis (2016), "validity concerns whether the variable, [the construct being measured by the scale or the exam], is the underlying cause of item covariance" (p. 59). A valid test is a test that can make strong arguments showing that when scores vary between test-takers, that variance is caused by the target construct as opposed to a problem with the test. In other words, two test-takers who get different scores on a test should be getting those scores because their knowledge, skill, or ability level of the construct is different.

Messick (1989) and Kane (2013) both extended this definition to examine the validity of how the test is interpreted and used (Kane, 1992, 2006, 2013; Messick, 1989a). Messick and Kane would examine the myriad ways in which stakeholders interpret and use the tests to determine if all, some, or none of these ways have strong validity claims. Importantly, Messick and later Kane emphasized that validity is an argument of an interpretation or a claim that relies on an "integrated, or unified, evaluation of the interpretation (Kane, 2001, p. 329). In other words, validation efforts should include specifying inferences and claims, providing evidence for those claims, and explaining plausible alternatives.

For the purposes of this dissertation and in the field of educational measurement, the key reference document outlining how to create a valid test is the *Standards for Educational and Psychological Testing*, henceforth referred to as the *Standards* (AERA et al., 2014). The *Standards* mirror many of the ideas put forth by Kane regarding building an argument for validation. However, the authors of the *Standards* cite five specific sources of evidence to be used when arguing for the validity of a test (2014):

1. Evidence based on test content (e.g., test blueprint, expert panel review)
2. Evidence based on response process (e.g., cognitive interview)
3. Evidence based on internal structure (e.g., factor analysis)
4. Evidence based on relations to other variables (e.g., convergent/divergent validity, predictive/concurrent validity, change across time, difference between groups)
5. Evidence based on related consequences (e.g., intended consequences, unintended consequences)

In this dissertation, I rely on the validity standards as outlined in the *Standards*

### ***Definition of Reliability***

Reliability according to the *Standards* is defined as reliability/precision, or consistency over replications of testing (AERA et al., 2014). According to the authors of the *Standards,* “We use the term reliability/precision to denote the more general notion of consistency of the scores across instances of the testing procedure” (2014, p. 33).

Traditionally, reliability is used to measure the correlation between two equivalent forms of tests; however, more generally, reliability refers to general test consistency and precision. So, even if the content of the test aligns with the intended construct, the consistency of the scores or the reliability could vary widely from test instance to test instance, making the measurement unreliable and, therefore, invalid.

### *The Relationship Between Reliability and Validity*

Both reliability and validity are words used to describe the quality of a measure. However, validity relates specifically to the accuracy of the measure (Heale & Twycross, 2015). Reliability, on the other hand, relates to the consistency of the measure. Heale and Twycross (2015) state it simply: “A simple example of validity and reliability is an alarm clock that rings at 7:00 each morning, but is set for 6:30. It is very reliable (it consistently rings the same time each day), but is not valid (it is not ringing at the desired time)” (p. 1).

### *Definition of Construct-Irrelevant Variance*

Messick (1989a) defines construct-irrelevant variance (CIV) when “the test contains excess reliable variance, making items or tasks easier or harder for some respondents in a manner irrelevant to the interpreted construct” (p. 7). In other words, CIV is aptly-named variance that is caused by some feature of the test that does not concern the construct or the target measure.

### ***Definition of Accommodation***

An accommodation, in the context of testing, indicates a modification of a test or the test-taking procedure from a standard condition (Thurlow et al., 2006). For example, common test accommodations include adding additional time to a time limit, having a helper read items or write, or adding translations or language dictionaries to a test (Abedi & Ewers, 2013; Elliott & Marquart, 2004). The purpose of accommodations is not to create unfair test conditions, but rather to make tests more equitable and, therefore, more valid measurements of the intended construct (National Research Council, 2004; Sireci et al., 2005).

### ***Definition of Fairness***

According to the *Standards*, fairness is a fundamental validity issue centered on a combination of bias, accessibility, and universal design (AERA et al., 2014). In short, fairness is ensuring that a test measures the construct validly in various contexts for a variety of individuals. According to the *Standards*, "A test that is fair within the meaning of the Standards reflects the same construct(s) for all test-takers, and scores from it have the same meaning for all individuals in the intended population; a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct" (2014, p. 50). Fairness can be thought of as a measurement principle (i.e., test should not demonstrate bias) and as an accessibility principle (i.e., test-takers should be given an opportunity to demonstrate their knowledge, skill, or ability level with regards to the construct).

### ***Definition of Measurement Bias***

According to the *Standards*, measurement bias is the central threat to fairness in testing (AERA et al., 2014, p. 49). Measurement bias is the result of construct-irrelevant features of a test causing variance in scores for certain groups of test-takers. In other words, measurement bias is when some features of the test or testing process, features that are not related to the intended measurement goal, cause advantages or disadvantages for some subgroup of test-takers (Freedle, 2003).

### ***Definition of Cultural Bias***

A broad term relating to systematically favoring one cultural group over another. Cultural bias can be found in textbooks, teacher attitudes, curriculum, and tests. As related to testing, Kruse (2016) defines cultural bias in testing as, "the existence of unexpected differences in test results for subgroups of similar ability levels within a tested population" (p. 23). Within the realm of cultural bias, some scholars have defined language or linguistic bias as bias against those from a particular language group or linguistic background (Aguirre-Muñoz, 2000; Chen & Henning, 1985; te Nijenhuis et al., 2016).

### ***Definition of Language/Linguistic Bias***

Language bias, sometimes called linguistic bias, is a sub-type of CIV in measurement caused by language differences among test-takers (Haladyna & Downing, 2004). Language bias is the systematic variance caused by language regardless of the test-taker's true score on the underlying construct.

## Research Questions and Overview of Research Flow

In this three-phase dissertation study, I aim to test the efficacy of the LAT using mixed-methods including a mix of cognitive interviews, inferential statistics using a within-subjects experiment, and in-depth interviews. All phases build on the Precalculus Concept Assessment (PCA; See Appendix A), a test developed at a large university in the U.S. Southwest and proctored semi-annually in calculus courses (Carlson et al., 2010). Participants included dominant Mandarin Chinese speakers and dominant Spanish speakers because those populations are the largest non-dominant English speaking populations in the United States (U.S. Census Bureau, 2015). I also recruited dominant English speakers to serve as the control group when testing the efficacy of the LAT. I selected a conceptual test, as opposed to a test of procedural knowledge or a test that predicts future performance, because the language in a conceptual exam is theoretically construct-irrelevant (Haladyna & Downing, 2004). As an opposite example, an American college entrance exam is written to measure predicted college success. Likelihood of succeeding in college is the underlying target construct. In this case, the mathematics section of the college entrance exam should include factors that are related to college mathematics success such as reading in English and working quickly. These factors, reading in English and working quickly are construct relevant. However, test writers who write a conceptual exam mean to measure test-takers' underlying conceptual understanding, not their ability to work through mathematics problems in English or to work though mathematics problems quickly.

The research plan included two pre-phases, needed to ensure the feasibility of the main study, and three phases of the main study. The first pre-phase was a linguistic analysis of the PCA. The second pre-phase was a differential item functioning (DIF) analysis of the PCA. For the main study, the first phase was a qualitative analysis of the linguistic complexity of the questions and cognitive lab analysis. The second phase was a linguistically sensitive translation of the PCA and configuration of the CAT. The third and final phase was a study of the efficacy of the new computer-based language-adaptive PCA.

The matching research questions for each phase include the following:

***Phase One***

1. Which test items, and how many test items, cause confusion or misconceptions for linguistic reasons for Chinese and Spanish dominant speakers when reading the PCA?
2. Is there a pattern of confusion or misconception that matches the types of test items as classified by Jamal Abedi and colleagues' linguistic complexity framework (2002; 2012; 2001)?

***Phase Two***

3. Is there preliminary evidence for the validity of the new language-adaptive concept assessment as evidenced by participant explanations of their selected answers?

***Phase Three***

4. Is the new language-adaptive concept assessment reliable and valid?
5. Do Chinese and Spanish speakers perform better on the new language-adaptive concept assessment compared to the original English version and the non-adaptive translated version?
6. What perceptions do test-takers have about the new language-adaptive concept assessment?

### Assumptions Underlying the Study

Carolyn Gipps in her book (2002), *Beyond testing: Towards a theory of educational assessment,* and Jones and Thissen (2006) argue that the underlying theory of standardized testing is essentially the same underlying theory of the science of psychometrics. In other words, there is a central assumption that some trait or construct can be measured in a quantifiable way for use in decision making. This theory, however, includes some strong assumptions including that: (a) there is a fixed measurable construct, (b) scores can be interpreted in relation to norms, (c) tests are objective, (d) the construct being measured on a test exists universally for all persons taking the test, and that (e) the construct being measured on a test is relatively unidimensional. These assumptions, however, may prove problematic.

Prominent educational philosophers, psychologists, and researchers contend that culture affects how children learn (Bronfenbrenner, 2009; Vygotsky, 1980; Wertsch, 1993). Culture impacts how people interpret the world around them and how they learn, think, and solve problems (Semken & Freeman, 2008; Solano-Flores & Nelson-Barber, 2001). Greenfield (1997) showed that culture influences epistemology (Greenfield, 1997, as cited in Solano-Flores & Nelson-Barber, 2001, p. 554). Standardized test writers often assume that students from different cultures have the same epistemological beliefs (Solano-Flores & Nelson-Barber, 2001; Ward et al., 2014). Test items that account for diverse epistemological beliefs face other forms of cultural bias such as favoring the background knowledge associated with students from a certain region or delivering test items in an unfamiliar dialect (Young, 2003). Other examples of bias appear in standardized testing, which has driven researchers to define and classify cultural bias in testing. The examples above show that the assumptions of testing can be violated when tests are given across cultures. In particular, the assumption of universality and the aura of objectivity can be undermined in cross-cultural testing.

## Limitations

Although this study involved several phases of in-depth analysis, there are limitations. This study may not always effectively parse cultural differences between students in a nuanced way. Culture, as a concept is incredibly complex and linked to language (Rogoff, 2003; Wertsch, 1993). In this dissertation, I attempt to uncover and provide a solution for cultural bias in testing by examining linguistic bias, which is a subset of cultural bias. I do this to keep the study manageable. Though some researchers in the past studied specific types of cultural bias in tests (Kruse, 2016; Sternberg, 2007), such as determining that some test-takers reacted to the figures of authority in word problems differently than other test-takers, I will limit my study to linguistic bias. I choose this limit because I propose to not only examine types of bias in items, but also present a new type of solution. Rewriting items to account for other types of cultural bias could prove overly complex for this study.

Additionally, studying linguistic bias is a lens into studying cultural bias. Studying only linguistic bias provides a frame into culture because culture and language are intertwined (Rogoff, 2003). However, another limitation to this study is that test-takers were placed into large bins for my quantitative analyses (i.e., all speakers who identify as dominant Mandarin Chinese speakers are analyzed as Chinese speakers regardless of their linguistic ability or nuance linguistic or cultural background).

Lastly, I designed this study under the assumption that results with Spanish and Chinese speaking participants could be generalized to a larger population, but this may not be the case. I chose Spanish and Chinese speakers because they represent two of the more commonly spoken non-English languages in the United States. In the remaining sections of this dissertation, I will present a literature review, followed by a description of the research design, findings from all three phases, and a conclusion and discussion.

# CHAPTER 2

# LITERATURE REVIEW OF THEORIES

## Introduction to the Literature Review of Theories

In the following section, I review literature that informs and frames the current study. For the sake of simplifying the organization of this section, I present four distinct parts: a) a review of cultural bias in education and educational measurement, b) a review of validity and the *Standards*, c) a review of literature framing the theory and assumptions underlying standardized tests, and d) a review of the methods used to mitigate cultural bias in testing. In other words, this section addresses: a) the problem, b) the lens used to investigate the problem, c) the underlying assumptions, and d) the history of potential solutions. Lastly, this section ends with my argument for how this dissertation study builds on previous literature to test a novel solution to a small piece of cultural bias in testing.

## A Review of Cultural Bias in Education and Educational Measurement

### *The Foundation of Cultural Bias in Anthropology and Linguistics*

Before researchers addressed cultural bias in curriculum design, teaching methods, learning behaviors, and educational assessments, much of the academic literature about cultural bias in education could be found in the related field of anthropology. Culture and how it relates to education finds in roots in canonical anthropology and linguistics literature such as Lucien Lévy-Bruhl's *How Natives Think* (1926), Sapir's *Language: An Introduction to the Study of Speech* (2004), and Whorf's

*Language, Thought, and Reality* (2012). These authors argued that culture and language influenced the most basic ways of thinking which introduced a key point of study that influenced future scholarship on human psychology. Of note, these works are considered canonical from a Western academic point of view, but also deeply because these early works operated from a white supremacist gaze that exocticized non-Western cultures.

Importantly, according to the above authors, language and culture are more than just a conduit for human thought (e.g., a way of expressing the mind). Instead, the authors argued that language and culture are fundamentally intertwined with thought. Whorf, for example, argued that the Hopi Native American tribe thought of time differently than other cultures. Instead of describing or referring to time as a noun (e.g., a flowing thing), the Hopi addressed happenings (i.e., time) as manifested and manifesting (with the important distinction that the manifested could include things beyond physical understanding and objective truth – a realm of greater possibilities) (Whorf, 1950). These writings played a seminal role in past and current research regarding cultural bias by breaking down the notion that people from all cultures think in the same ways.

Though the inaccuracies of the work of the above authors has been much discussed—e.g. Lévy-Bruhl (1926) arguing that there are primitive minds and modern minds-- there is little doubt that such works had a profound impact on academic literature on culture and thought. Implicit in these arguments is the idea that people from different cultures who speak differently also learn differently. These implications are at the core of any argument about cultural bias in education. Though some researchers contradict the above argument (Glick et al., 1969), in general, the argument that culture and language affect ways of thinking gained popularity in academic circles.

### *The Foundation of Cultural Bias in Education and Psychology*

In the early to mid-20th century, contemporary with Lévy-Bruhl, Sapir, and Whorf, scholars of education such as Jean Piaget and Lev Vygotsky were developing learning theories that would also have a huge influence on questions about culture and the learning. Lev Vygotsky, who studied developmental education in Russia in the early 1900s, published a series of papers that would only later gain popularity among a wider audience of educational researchers in the 1980s. Piaget, who produced his work later than Vygotsky, experienced widespread popularity of his work before Vygotsky. Piaget's work, however, did not account for the interconnectedness of culture and thought. According to Wallace Lambert, "The Piaget school argues for the independence of language and thought while the Vygotsky school argues for an interdependence." (Lambert, 1973, pp. 2–3). In other words, though members of the Piaget school would no doubt argue that language and culture adjust the way that thoughts are expressed, they would not argue that language and thought are fundamentally linked, as members of the Vygotsky school would. Members of the Vygotsky school argue that, though thought and language start separately for young children, by the age of two they are intertwined. The divide between Piaget and Vygotsky schools of thought regarding the independence of language and culture and thought would become more important later in the century when scholars began to address multicultural education.

### ***Cultural Bias and Multiculturalism***

In previous paragraphs, I addressed some of the more theoretical roots of how culture became an important part of the conversation in educational research. However, as the civil rights movement ramped up in the United States, so too did efforts to address multiculturalism in the classroom. With addressing multiculturalism in the classroom, the problem space began to include not only high-level theory on learning (i.e., questions like, “does culture affect learning?” as addressed by scholars like Piaget and Vygotsky), but also include other aspects of education including, but not limited to, curriculum design, teaching methods, and educational assessment. Researchers began to address questions such as whether the content of the formal education system created a bias against certain cultural groups and whether even the classic educational experiments conducted by themselves were biased toward white males, or those of European descent.

Educational research at the time did not address *cultural bias* per se (e.g., there was no special issue addressing bias). Rather the research implied that there was an existing cultural bias by suggesting that students from different backgrounds may need a different type of pedagogy, different content, a different stance toward education, or even a different definition of knowledge. By implication, scholars such as Pierre Bourdieu argued that current education research and practices favored the dominant social group (1984). In this space, notably, in the late 1970s, Bourdieu introduced the idea of cultural capital as a way to explain how cultural background (i.e., ways of speaking and acting that are socialized) led to increased social mobility (Dasen, 1984). In other words, Bourdieu posited a theory about how one dominant social group could provide more social currency for their offspring, in the form of sharing their cultural ways of speaking and acting, in turn, allowing for increased social mobility.

For the first time, educational research was focusing on cultural differences and their effect on education. Piaget's theories were under attack for being too rigid to account for students with a variety of cultural background (Dasen, 1984; Greenfield, 1997). At around the same time, as part of the general paradigm shift regarding the importance of studying non-white middle class cultures, as certainly influenced by the scholars mentioned previously, a group of sociolinguists began to publish work about social language and literacy. These authors, including Shirley Brice Heath, William Labov, Geneva Smitherman, and Luís Moll revived and reinvigorated the earlier work of Vygotsky and other sociolinguists arguing that language and culture are fundamentally intertwined.

Shirley Brice Heath's research in *Ways with Words: Language, Life, and Work in Communities and Classrooms* (1983) served as an exemplar for socio-cultural research of an issue because her research centers on the complex social context surrounding that issue. Specifically, Heath argued that language and behavior are constructed from a young age through localized community interactions creating miscommunication and conflict when two (or more) cultures meet. In her work, Heath, revealed in detail how the children of two Southern mill town communities (one primarily black, the other primarily white) learned to communicate through spoken words, written language, behavior, and play. Heath’s main argument was that language and culture are interdependent and, therefore, claiming one form of language to be more appropriate than another disadvantages certain communities. “The place of language in the cultural life of each social group is interdependent with the habits and values shared among members of that group” (p. 11). Importantly, Health’s work showed that class was arguably just as important as race in terms of interdependence with language which adds nuance to exactly how culture is considered. As with Bourdieu’s theory of social capital with regards to formal schooling, Heath argued that context and cultural background play a significant, and often unrecognized role in how teachers interpret the skills and knowledge of students.

As the field of multicultural education developed, the definition of multicultural education developed as well. Sleeter and Grant (1987) stated that multicultural education can also go by the names emancipatory education, transformative education, and critical teaching. Later, this type of education gained other names and continues to be somewhat inconsistently identified (see also, the following section about teaching methods). According to Castagno and Brayboy (2008):

> It is also important to keep in mind that this literature is somewhat inconsistent with respect to the name given to these culturally based educational practices; some of the most commonly used names are culturally responsive, culturally relevant, culture-based, and multicultural education. (p. 946)

In other words, defining such a fluid phenomenon as culture has challenged recent academics who seek to create a term that recognizes the complexity and nuance without becoming too complex to be disfunctional.

### *History of Cultural Bias in Teaching Methods and Curricular Plans*

Gloria Ladson-Billing's article, *Toward a Theory of Culturally-Relevant Pedagogy* (1995), kicked off a new era in dealing with cultural bias in teaching practices. Drawing from the work of scholars before her who conducted work with Native Hawaiian and Native American students (K. Au & Jordan, 1981; Mohatt & Erickson, 1981), she argued for a pedagogy that would include aspects of students' home environment with students' school environment as opposed to a pedagogy that emphasized one-size-fits-all best practices. The pedagogy would be based on, "a theoretical model that not only addresses student achievement but also helps students to accept and affirm their cultural identity while developing critical perspectives that challenge inequities that schools (and other institutions) perpetuate" (Ladson-Billings, 1995, p. 469). According to Ladson Billings, teachers have more success when they incorporate the cultural behaviors and beliefs of their students that are conducive to academic success into the classroom. This stands in opposition to teachers who try to limit some behaviors that are culturally acceptable to students at home but may not fit into the traditional school behaviors. For a specific example, Au and Kawakami (1985) show how young Hawaiian children struggled learning to read because they actually struggled with the classroom rules of interaction. The teacher spent valuable time trying to get students to answer one by one rather than engaging in the material. When the researchers evaluated a different teaching style that allowed students to answer in groups, students appeared to show gains.

Culturally-relevant pedagogy as a concept was subsequently adapted by many in the academic community. As with multicultural education, culturally-relevant pedagogy goes by many names in the academic community, typically with different foci, including culturally-congruent, -appropriate, -compatible, -responsive, and -responsible (Harmon, 2012). However, the overall message is the same for each type of pedagogy. These pedagogies are a move towards personalizing classroom practices to students with an emphasis on students' life outside of the classroom. Culturally-relevant pedagogy, for example, would include positive perspectives toward student families, student-centered instruction, kinship with students, holding high expectations for students, drawing on the culture and history of the students. According to this pedagogical framework, successful teachers would need to spend a significant amount of effort to learn about their students' communities and lives outside of school to best support student learning.

### *History of Cultural Bias in Educational Assessment*

Though media and researchers often present evidence that educational assessments favor students of the culturally dominant race (Freedle, 2003; Taylor, 2016), few studies address feasible psychometric solutions. Many researchers, however, present solutions for wider issues about race and pedagogy by drawing on Gloria Ladson-

Billing's (1995) call for culturally relevant pedagogy (Carter, 2008; Milner IV, 2008; Vaught & Castagno, 2008).

Assessment designed to compare students is common in formal education. The most well-known comparative tests include college entrance exams such as the SAT and the ACT which, in essence, attempt to measure which students are most likely to succeed in college. These tests have a large impact on the trajectories of students, the rankings of universities, and various other economic factors. However, large-scale comparative tests are controversial and, potentially, biased against certain groups of students.

According to Wayne Au (2010), the standardization of tests may be the root of large scale biases against certain groups because the tests are standardized by and for those in power or part of the dominant group. In other words, these tests are standardized, but according to whose standards? They are standardized according to the accepted standards formed by those with the political capital and social standing to make those types of decisions. Intuitively, all students receiving the same items seems to be the fairest practice. However, test items are not perfect; they often measure unintended constructs (Noble et al., 2012). For example, mathematics items written in paragraph form test mathematics ability, but also reading ability. Thinking back about fairness, the fairest practice is not about presenting the same items, but about assessing the same underlying constructs (Crocker & Algina, 1986). The items themselves are simply delivery mechanisms. If changing the item results in a more accurate score (a score that more closely measures the underlying construct) then the items can and should be changed.

The potential error in test items goes well-beyond mathematics items written in paragraph form. Providing the same standardized test items to students from different cultures or backgrounds may result in validity issues because not every student will read or interpret items in the same way (Kruse, 2016; Lun et al., 2010; Sternberg, 2007). More precisely, the chance of a test item failing to accurately measure the intended construct may be directly connected to the interaction between how the test item is delivered (e.g., delivered in English, confusing or ambiguous wording) and students' cultural background and prior knowledge (e.g., the student is familiar with certain types of test items, the student holds a certain epistemology, the student struggles with specialized English vocabulary) (Luykx et al., 2007; Solano-Flores, 2006; Ward et al., 2014; Winter et al., 2006).

Modifying items for each student is not as radical as it sounds. New developments in CAT methods have changed the meaning of "standardized" tests. Students no longer receive the same test items in the same order. An algorithm, adjusting for each correct or incorrect answer, drives the order, difficulty, and even whether some items even appear at all. In short, test-takers receive personalized tests designed to trim away redundant items. Test designers do this to reduce testing time, fatigue, and cost. However, test designers could use the same principles to reduce other issues as well by considering other ways to personalize tests (Casaletto et al., 2016).

### *History of Research on Solutions to Cultural Bias in Educational Assessment*

Though research on feasible solutions to cultural bias are limited, there are some works worth highlighting. In 1998, the *Journal of Negro Education* published a special issue called *Assessment in the Context of Culture and Pedagogy* (Hood, 1998; Lee, 1998; Qualls, 1998). This special issue, released a few years after Gloria Ladson-Billing's landmark article on culturally relevant pedagogy, represents the first block of articles addressing assessments and culturally relevant pedagogy. Within this issue, three works present theoretical solutions and frameworks for building more culturally relevant assessments. Authors of two of these works argue that performance-based assessments can deliver a feasible solution (Hood, 1998; Lee, 1998). Performance-based assessments test students while they are enacting a skill and provide a context for students to use their knowledge. More importantly, authors of these two studies argued for performance-based assessments that use cultural funds of knowledge to determine which skills to test. Though the term *performance-based assessment* does not appear often in the subsequent literature, the concepts presented in these works do appear. Many of the works indicate how to use cultural funds of knowledge to create assessments (see Berardi et al., 2003; Rameka, 2011; Slee, 2010).

Several other researchers present practical solutions to solving such problems. These studies can be divided into two categories: those in which authors address how programs or policies should be designed, and those in which authors address what teachers should do. The papers in each category share a few similarities and exhibit some key differences. An in-depth look at the differences in Hue and Kennedy's (2015) findings and Berardi et al.'s (2003) approach show some of the contradictions between program-focused works and teacher–focused works. Hue and Kennedy (2015) claim the the following principles of culturally responsive assessment: a) integrating the 'part' of the assessment with the 'whole', b) managing diverse learning needs of students, c) removing language barriers from the assessment process, and d) examining the influence of public examinations on teachers' classroom assessment practices. Whereas Berardi et al. (2003) emphasize three key principles: a) integrating tribal and western knowledge, b) adopting a non-abandonment policy towards students, and c) devoting resources to developmental skills such as writing and computational skills. Unpacking each principle is beyond the scope of this synthesis; instead, I will highlight overlapping and contradicting concepts within the two papers.

Authors of both of these papers about programs and teacher practice emphasize using alternative forms of assessment such as group work, verbal work, and written work to assess students. Berardi et al. (2003) couch their alternative approach within the argument that tribal ways of knowing include multiple ways expressing that knowledge. Hue and Kennedy (2015), on the other hand, argue that alternative forms of assessment, such as observing group work, are best implemented through curriculum change. Authors of both works recommend the same practical change but vary in their reasoning. Essentially, Berardi et al. calls for a system-wide change whereas Hue and Kennedy highlight how teachers make small changes within the existing examination driven system.

The call for system-wide change versus small classroom-level changes divides the literature in this field. Tippins and Fichtman Dana (1992) present a list of practical tips for the teacher to incorporate *within* the current system. Most of these suggestions are for formative small-scale assessments, not summative evaluation. These suggestions fit within the arguments presented by Hue and Kennedy for making small changes within the classroom. On the opposite pole, Slee (2010) wrote a paper defending a systemic change to assessment policies. She addresses the need for revolutionary types of summative assessment. For example, Slee argued that students should not be punished for work they do not turn in. Additionally, she argued that students should not be graded on attendance. She argued for large-scale change *throughout* the system. Whereas Hue and Kennedy argued for smaller scale changes *within* the system.

Lastly, authors of the two studies employed critical action to employ change (Coles-Ritchie & Charles, 2011; Nagai, 2001). Both authors focused on creating curriculum and assessment based on cultural funds of knowledge, emphasizing the connection between community and education. Nagai (2001) argued that assessment takes place in and out of school. She also argued that assessments can be conducted by teachers, parent, or even peers. Coles-Ritchie and Charles (2011), on the other hand, centered around making changes in the classroom by adapting formative and summative assessments. The principal contribution of these works is to show a model of how to implement culturally relevant assessments in real life situations.

### *How This Dissertation Contributes to Research on Cultural Bias*

Cultural bias touches every aspect of education from teaching and learning to theory and practice. Educational theories that serve as cornerstones for learning science classes, such as Piaget's stages of cognitive development and Vygotsky's zone of proximal development, developed from research on white, European, middle-class children. In practice, scholars of culturally responsive pedagogy emphasize that teaching in the classroom must reflect and respect the cultural beliefs of students in the classroom.

The major underlying argument, that assessments need to change to account for students' various cultural backgrounds, led to various approaches to solving the problem. Some literature contributed to the knowledge base by defining the problem more clearly or more complexly. Other works contributed to the base by presenting theoretical or practical advice and solutions. Lastly, some work contributed by enacting change as researchers conducted studies used action research. Accordingly, the existing scholarly literature offers myriad approaches to solving the problem; yet holes in the literature base remain.

First, further research utilizing empirical methods would advance the research base. Multiple frameworks exist to assess students using cultural funds of knowledge and alternative assessments, but the field needs more empirical studies to measure success. Researchers should conduct more empirical research demonstrating when and how certain strategies succeed. The conditions for when to use strategies are severely under researched, yet these studies would be most helpful to practitioners. In addition, empirical studies on culturally relevant assessments may reach a wider academic audience. In this dissertation, I apply relatively recent theories of multilingualism and translanguaging, into practice by incorporating them into a new type of test delivery and testing the efficacy among multilingual students. Such research contributes to the gap in empirical research.

Second, the literature reveals a division between studies in which researchers address indigenous communities and studies in which researchers address non-indigenous students from historically marginalized groups. Though many findings and theoretical frameworks may extend to both groups, the context is not the same. Differences in Hue and Kennedy’s (2015) findings and Slee’s (2010) framework expose this reality. Studies with diverse participant groups would bolster the research in the field. In this dissertation, I recruited participants with diverse linguistic backgrounds, all of whom identified as dominant speakers of languages other than English. Hence in this dissertation, I studied a diverse community of linguistically diverse students.

Lastly, authors of the current literature do not adequately address how to fit culturally relevant assessments into the existing assessment frameworks used by universities for admissions and policymakers to determine success. Surely, research can make inroads without revolutionizing the whole system of standardized assessments. The field needs research that provides suggestions for intermediary steps. In fact, though scholars focused on educational measurement and scholars focused on cultural bias in education can sometimes be at odds, both groups have common ground. In this dissertation, I design and test the efficacy one potential practical solution which applies recent research on linguistics with existing tests. Such a solution has the potential to be applied immediately to existing systems of testing at schools around the country. Via the next section in the literature review I summarize and synthesize literature about validity and other standards for educational measurement which relate to reducing cultural bias in testing.

## Language and Testing

For this dissertation, I focus on only one small aspect of cultural background: test-taker language. However, to address test-taker language, I examine a recent theory of multilingualism called translanguaging. Then, I show how that theory has been applied to testing practices. Not many researchers have addressed how to bridge the gap between the theory of translanguaging and the applied practice of creating better tests for multilingual learners, which lends value to this dissertation.

### *Introduction to Translanguaging and How it Applies to Testing*

Solano-Flores and Trumbull (2003) argue that some students performed better in their dominant language than in English for some items, but better in English than in their dominant language for other items. According to Valdés and Figueroa (1994), "When a bilingual individual confronts a monolingual test, . . . both the test-taker and the test are asked to do something that they cannot. The bilingual test-taker cannot perform like a monolingual. The monolingual test cannot ‘measure’ in the other language” (p. 87). Valdés and Figueroa challenge the notion of bilingualism by implying that bilingual students use both languages, together, to take a test. Even the term bilingualism implies a divide between two languages that does not accurately reflect the way diverse languages interact. I use the term bilingualism only in reference to the work of previous scholars.

The notion of transitioning from one language to another fluidly within a test relates to the theory of translanguaging (García & Wei, 2014). Translanguaging is defined as “the deployment of a speaker’s full linguistic repertoire without regard for watchful adherence to the socially and politically (and usually national and state) defined boundaries of named languages” (Otheguy, García, & Reid, 2015, p. 283). In other words, translanguaging emphasizes fluidly using the knowledge of multiple languages (and variations of languages) to think and act without thinking about switching between official languages. Lewis et al. (2012) argue that the new concept of translanguaging replaces an older conceptualization of language such as code-switching. Primarily, translanguaging transcends the boundaries of language. Adherents to translanguaging argue that translanguaging stands in contrast to the view that students think or act in one language or another. Instead, translanguaging scholars recognize that having knowledge of multiple languages affects meaning-making. These languages, including the way these languages interact and entangle, serve as a resource and a fund of knowledge.

There is limited research on how translanguaging applies to testing (Dendrinos, 2013; Lopez et al., 2017). However, some scholars have discussed how translanguaging applies to testing. Otheguy and colleagues (2015) state that school testing techniques often inhibit translanguaging by forcing bilingual or multilingual students to act in one language during testing. In a more recent paper regarding translanguaging and testing, Ascenzi-Moreno (2018) argues that most testing accommodation designs disadvantage bilingual students "because they use only a portion of their language abilities" (p. 359). For example, even when test writers incorporate multilingual dictionaries or translations in a test, multilingual students still need to primarily read and answer in one single language. In theory, test writers accounting for translanguaging would need to design an accommodation that allows multilingual students to act and answer in multiple languages. The impact of research on such an accommodation has the potential to impact the linguistics and assessment academic communities by furthering research on the application of translanguaging to measurement.

### *Importance of Developing Linguistically Adaptive Tests*

Standardized testing is commonplace in formal schooling worldwide, often attached to high-stakes consequences, and contributes to achievement gaps among certain cultural groups (Tienken & Zhao, 2013). As a consequence, the design of standardized testing including every small alteration in item writing or score interpretation has the potential to have major consequences (Abedi, 2002). Standardized testing can often measure the characteristics of examinees incorrectly because quantitative measurement of psychological features is notoriously difficult (Gould, 1996), especially in a single test, sitting among a diverse set of test-takers (Dumas & McNeish, 2017). However, standardized tests are also an important and undeniable part of the current education system (Tienken & Zhao, 2013).

LATs specifically target language bias issues by providing a test that allows the test-taker to read items in both languages (i.e., by allowing translanguaging). LATs are a practical approach to a greater problem. LATs do not solve some of the more complex types of cultural bias such as testing constructs that have different meanings in one culture or another (Kruse, 2016). However, LATs take an important step toward fairness by, in theory, treating multilingualism as a resource (Abedi, 2002).

To treat multilingualism as a resource, test writers using LATs provide a platform for multilingual students to act and answer in many languages without any specified primary language. For LATs, there is no primary test language. Students can toggle back and forth between languages. The toggle feature means students can use multiple languages in unison to think about and solve test questions. The act of using multiple languages simultaneously in order to think and act reflects the principles of translanguaging.

Furthermore, creating a toggle system may be better than presenting languages side-by-side. By toggling back a forth, test-takers do not get overloaded with content by seeing two side-by-side test questions, one in one language and one in the other language. Such cognitive overload, when a test-taker’s visual attention is split, can affect performance and learning by effectively overloading working memory (Paas et al., 2004; Sweller, 1988). Some researchers have examined how cognitive load issues apply to testing environments with Elliot and colleagues (2009), for example, recommending that text should be used economically and avoid splitting test-takers' attention. This argument supports a toggle system over a side-by-side translation system; however, the toggle system still presents more cognitive load than a simple tests without translation. Therefore, LATs, like all accomodating tests, have limitations. The tests may provide improvements in some ways and increased difficulties in other ways. This complicated interaction supports the need for examing the results using multiple methods to parse when and how the toggle system works or otherwise.

In terms of increasing validity, LATs target CIV issues where the delivery of the test (i.e., the language of the test or the way a question is worded) causes variance in test scores among subgroups that has nothing to do with the test-taker's relationship to the underlying construct that is meant to be measured by the test (Haladyna & Downing, 2004).

## Validity

To answer how a LAT could be effective, I will give an overview of validity and bias in testing, then explain the theoretical basis for standardized testing, followed by an explanation about the conflicts behind the theoretical basis of testing.

### *Summary of Four Views of Validity*

Researchers in psychology, education, and measurement have been discussing validity for decades, but no every research views validity in the same way. As an overview, the following section highlights four views of validity from a more classic concept of validity as a concrete element of a test, championed by DeVellis (2016), to a more holistic concept of validity, championed by Messick (1989), then later advanced by Kane (2013), which includes more thought about how the test is used by others. Lastly, the fourth view, championed by Borsboom (2004), returns to a definition much closer to that championed by DeVellis. The following paragraphs explains these views with more detail and nuance.

The classic concept of validity started out with a simple idea: tests need to prove that they measure what they claim to measure. According to DeVellis (2016), “validity concerns whether the variable is the underlying cause of item covariance” (p. 59). In other words, a valid test is a test that can make strong arguments showing that the covariance in item scores is caused by the target construct. This question alone is complex enough to inspire lists of types of validity. For example, construct validity addresses whether the test is addressing the construct in question (e.g., does this depression index measure depression or does it measure anxiety?). Face validity basically asks, “does this look like it measures what it is supposed to measure?” Other types of validity include predictive validity which addresses whether the measure accurately predicts what it should predict. Conceptually, this notion of validity is oriented toward proving that one test is or is not validly measuring what it purports to measure.

Samuel Messick (1989a) extended the concept of validity to include, not only the test itself, but how the test is used. Perhaps because of his background working at a large testing company which dealt with practical issues related to testing, his concept of validity shifted the focus of validity from the test to the real-life use of the test. Specifically, Messick (1989) emphasized that arguments for validity need to support, “the adequacy and appropriateness of inferences and actions” (p.13). Messick’s argues there is no such thing as a valid test, only a valid use of a test. Validity is really an argument for how the test is used.

Michael Kane (Kane, 1992, 2006, 2013) followed Messick by creating a framework for how to argue for validity. He leaned on Toulmin’s model of inference to create a method for establishing validity (Toulmin, 2003). Kane deemphasized checking off types of validity in favor of creating a strong series of warrants to defend validity claims (Kane, 1992, 2001, 2006). He emphasized focusing on key inferences necessary to making a strong validity argument to stakeholders.

In particular, Kane pointed to four key inferences: scoring, generalization, extrapolation, and implications. The first key inference, scoring, refers to the inferential jump linking the observation (or test item response) to a score. The second key inference, generalization, refers to the inferential jump linking the test score with actual test performance. The third key inference, extrapolation, refers to the inferential jump linking the test score with real-world performance. The fourth and final key inference, refers to the inferential link between test scores and decision-making. Most importantly, for the sake of validation, each inference and each claim made, needs to be named and argued for by those validating the test (Kane, 2006, 2013).

Though Kane built on Messick's theory, not all recent validity scholars support Messick's validity concept. For example, Borsboom et al.'s (2004) paper represents a return to the idea that a test can or cannot be valid, with authors arguing that validity is something simpler than that proposed by Messick and "thus, validity theory has gradually come to treat every important test-related issue as relevant to the validity concept and aims to integrate all these issues under a single header" (Borsboom et al., 2004, p. 1061). Borsboom et al. argue that validity should be simply defined as whether the target construct causes variation in the test score.

Consider the validity of the SAT college entrance exam through the lens of each of these four schools of thought regarding validity (i.e., DeVellis, Messick, Kane, and Borsboom et al.). Starting simply, both DeVellis and Borsboom et al. would focus on different items than Messick and Kane. DeVellis and Borsboom et al. would employ tactics to prove the validity of the test itself, whereas Messick and Kane would examine the validity of how the test is interpreted and used. In other words, DeVellis and Borsboom et al. would read through the SAT test manual, note the arguments for why the SAT measures college success, then determine if the test is valid. Messick and Kane would instead look to see at the myriad way in which stakeholders interpret and use the SAT to determine if all, some, or none of these ways have strong validity claims.

At this point Messick and Kane may diverge in how they go about creating arguments for validity. Kane would be primarily interested in first establishing the claims made for a proposed use then analyzing whether each of those claims are valid through logic and empirical evidence (Kane, 1992). Kane would also be interesting in plausible alternatives for the empirical evidence. In other Kane, would expect a well-structured and coherent argument linking the test items, the test scores, and the test output, with the test use (Kane, 2001).

Messick does not have a framework as established as Kane, but, through his lens, he is more focused on whether the test serves its intended purpose. A valid test use, in Messick's view, is a use that successfully reaches its intent.

DeVellis and Borsboom et al. may also diverge in how they evaluate validity. Borsboom et al. (2004) essentially believe "a test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure" (p. 1061). Though Borsboom et al.'s concept of validity shares much with validity as described by DeVellis (e.g., construct and content validity are key; validity is a property of the test), Borsboom et al. decry the heavy-handed role of criterion validity in many validity studies. According to Borsboom et al., "not just criterion validity but any correlational conception of validity is hopeless." (p. 1067). In other words, Borsboom et al. believe that developing a strong causal explanation is much more important than finding that the scores correlate with the intended outcome.

Lastly, the four sources differ somewhat in how they conceive of validity epistemologically speaking. Both DeVellis' version of validity and Borsboom et al.'s version of validity reveals a scientific realist epistemology (Hood, 2009). In other words, they reveal a view of truth as somewhat hard. A test can be valid or not. Messick and Kane, on the other hand, reveal a more constructivist epistemology. To them, validity is determined by stakeholders according to the use and context of a test. As such, validity changes over time. A test use that is valid one year, may no longer be valid the next as the views of stakeholders change. Therefore, according to Messick and Kane, determining the extent of validity is a never-ending process.

To develop the LAT, I will lean more heavily on Messick and Kane's more holistic (and pragmatic) view of validity. In other words, developing a valid test will mean the need to also show how the test can be used with particular emphasis on the context of the test. Specifically, it is not just about making an objectively better test, but rather creating a test that is more valid than other similar tests in education used for purposes like placement or evaluation.

### *Bias in Testing*

Following the explanation of validity theory and before launching into an explanation for how a type of CAT could feasibly reduce bias, it is important to define bias in the context of testing. Bias can be defined in two ways, the popular way and the statistical way. In common parlance, if someone is biased it simply means they are acting unfair, whereas, in testing, bias refers to a systematic underestimation or overestimation of a value across subgroups (Maller, 2003; Reynolds & Suzuki, 2003). The bias that we discuss in this paper is limited to the systematic validity issues across subgroups. Fairness will also be discussed, as it is an important element affected by testing bias, and something explicitly appearing in the *Standards*, which, plays the role of the core reference text regarding testing spanning from academic to legal uses (AERA et al., 2014).

As for bias, though the broad definition is relatively simple, bias can be divided into different parts and the causes behind systematic validity issues across subgroups can be complex. Van de Vijver and Tanzer (2004) demonstrated a three-part classification of bias among psychological tests that can be, generally, extended to education tests: construct bias, method bias, and item bias.

Construct bias means having some issue with the underlying construct that causes a systematic error in measurement. This can be manifest in three ways such as a) having a construct that exists in one subgroup, but not another, b) eliciting test behaviors that relate to the construct for one subgroup, but not another, and c) selecting such behaviors poorly in such a way that one subgroup performs better on the selected behaviors. Method bias means having issues with the method of collecting data. This can be caused by sampling bias of test-takers indicating incorrectly differential scores among subgroups, instruments that are more familiar to one subgroup than another, or issues with administration of the test. Lastly, item bias means having bad translations, content that does not cross cultures, nuisance factors, connotations for certain subgroups, or unfamiliar content.

Additionally, the following seven causes are one categorization of the causes drawing on the taxonomy developed by Reynolds et al. (1999):

1. The tests can have inappropriate content that favors certain subgroups. This means requiring somewhat exclusive background knowledge for correct answers that may favor one subgroup over another.
2. The test can be incorrectly normed on a sample that has too few minorities (even if the proportion is correct). This can be particularly difficult if some of the affected subgroups are particularly uncommon in the test-taking population.
3. The language of the examiner may intimidate test-takers who may not be used to the academic or formal style of language. As an example, Word (1977) showed that African-American interviewees felt defensive about how White interviewers reacted to their informal language.
4. The social consequences of receiving a certain score may be different for different subgroups. This reason for test bias is related to how the test is used, more than test construction itself.
5. Measuring different constructs. In this case, the test may actually be measuring the incorrect construct for some subgroups. For example, a conceptual mathematics test that uses items that are written in long paragraphs with difficult vocabulary may be measuring different constructs for test-takers depending on their level of English proficiency.
6. The predictive validity of the test is predicated on the majority subgroup. In other words, the test is well designed to predict future outcomes for certain subgroups, but not others.

7. Non-universal types of characteristics between subgroups. In this more radical argument, different subgroups have fundamentally different types of characteristics and should not even be taking the same test.

These seven causes of bias can be combined further into two themes, according to Reynolds et al. (1999): “inequitable social impact and qualitatively distinct aptitude and personality” (p. 86). With these views of validity and bias in mind, in the following section I address the theoretical basis for standardized testing as based on classical psychometric theory and go on to discuss which theoretical pillars are facing challenges within the educational testing community.

## Theoretical Basis for Standardized Testing

Shifting from discussing the practicality of LATs, in this section, I address the underlying theory of standardized testing and the challenges to that theory. Standardized testing, as we know it, grew from the growing field of psychometrics which sought to develop the quantitative and rational side of psychology and then cemented itself after widespread use during wartime (Jones & Thissen, 2006). Haney (1981) argues that educational testing in general would have remained strictly for professional use except for the role that standardized testing played during World War I. Cronbach (1975) stated that psychologists, who considered army testing successful, decided to adapt such testing style for the educational community. However, the transition from conducting army aptitude tests to conducting large-scale educational testing meant controversy and an examination of the underlying assumptions and theoretical basis for standardized testing.

Carolyn Gipps in her book (2002), *Beyond testing: Towards a theory of educational assessment,* and Jones and Thissen (2006) argue that the underlying theory of standardized testing is essentially the same underlying theory of the science of psychometrics, meaning, that there is some trait or construct that can be measured in a quantifiable way for use in decision making. Gipps (2002) put forward six underlying assumptions of this underlying theory, some of which are considered somewhat traditional and are already being challenged within the assessment community:

1. There is a fixed measurable construct. Originally, psychometricians were trying to capture the fixed notion of intelligence. However, more recently this definition construct has been expanded to include anything from reading ability to depression level.
2. Scores can be interpreted in relation to norms, meaning test scores are placed in relation to other peer test scores. In education this approach is controversial. Most educational tests, instead, base scores in relation to criteria or standards. This approach is called criterion-referenced testing (Glaser, 1963).
3. Primacy of technical issues. A focus on developing statistical methods to determine tests that are reliable and tests that reduce variance.
4. Aura of objectivity. Related to the above point, and certainly problematic from a larger social sense. This assumption of standardized testing is that the numbers derived from such test are somewhat objective and therefore, it is easier to make decisions with the scores.

5. Assumption of universality. This assumption states that the construct being measured on a test exists for all persons taking the test.
6. Unidimensionality. This assumption which is that the construct being measured on a test are relatively unidimensional.

In the following section, I reveal some of the issues with the previously stated theoretical basis for standardized tests.

### ***The Theory of Standardized Testing and Conflicts with Recent Theories of Learning and Knowledge***

Prominent educational philosophers, psychologists, and researchers contend that culture affects how children learn (Bronfenbrenner, 2009; Vygotsky, 1980; Wertsch, 1993). Culture impacts how people interpret the world around them and how they learn, think, and solve problems (Semken & Freeman, 2008; Solano-Flores & Nelson-Barber, 2001). Greenfield (1997) showed how culture influences epistemology by conducting experiments to determine if children demonstrated the Piagetian construct of conversion with children from the Wolof tribe in Senegal (Greenfield, 1997, as cited in Solano-Flores & Nelson-Barber, 2001, p. 554) The children did not answer questions such as, "Why do *you think* …?" or "Why do *you say* …?", remaining silent and failing the assessment. When the question was changed to "Why is the water the same (or more, or less)?" the students suddenly offered coherent explanations. According to Greenfield (1997), the Wolof children did not believe that knowledge and reality were separate, so asking about their thought or opinion made no sense. The children exhibited an epistemology called mental realism which rendered the assessment questions, the same questions that worked for Piaget and his colleagues, useless. However, only a slight change to the assessment questions to account for the epistemology of the Wolof children allowed for accurate measurement. According to Greenfield (1997):

> Had an exact translation of the Cambridge conservation procedure been used, it would have been erroneously concluded that the unschooled Wolof children were not able to explain the reasoning behind their quantity judgments. Their theory of mind would have been confounded with their reasoning about the world. The research publication would have incorrectly concluded that unschooled Wolof children had a major cognitive lack in reasoning skill. Instead, the conclusion from pilot testing was that unschooled Wolof children had a different epistemology and therefore required a different interview procedure. When tested with an epistemologically appropriate procedure, the cognitive deficit in reasoning about the world disappeared. (p. 313)

Greenfield's analysis shows that assessing multiple cultures with the same test items may result in false results. Standardized test writers often assume that students from multiple cultures have the same epistemological beliefs (Solano-Flores & Nelson-Barber, 2001; Ward et al., 2014). Test items that account for diverse epistemological beliefs face other forms of cultural bias such as favoring the background knowledge of a certain region of students or delivering test items in an unfamiliar dialect. Young (2003) gives an example from the 1998 SAT:

The dance company rejects_________, preferring to present_____ dances in a manner that underscores their traditional appeal.

A. invention emergent

B. fidelity long-maligned

C. ceremony ritualistic

D. innovation time-honored

E. custom ancient

The correct answer is D. The racial background of those answering the question correctly was 38% Black and 62% White (Young, 2003). Test items referring to specific cultural spheres, such as items about traditional dance companies, may be unfamiliar to students from other cultures who rely on their own worldview and experiences to make sense of test items. In a different article, Rosner (2000) gives an example of when a question favors African American students instead of white students:

Choose the best antonym for RENEGE

A. dispute

B. acquire

C. fulfill

D. terminate

E. relent

African American students performed *better* on average than white students on this trial question which test-makers ultimately excluded from the SAT. Researchers speculated that African American students may use the term renege more often with regards to African American history in the United States (Rosner, 2000). Other examples of bias appear in standardized testing driving researchers to define and classify cultural bias in testing. The examples above show that the assumptions of testing, as presented by Gipps (2002) can be violated when tests are given across cultures. In particular, the assumption of universality and the aura of objectivity can be undermined in cross-cultural testing.

The complexity of test development, especially regarding how tests are used to make decisions, led three major academic bodies, AERA, the American Psychological Association (APA), and NCME, to come up with a set of standards for educational and psychological measurement. Though educational measurement may have many flaws, this set of standards represents the best consensus from the field of education and psychology about how to create valid tests. In this dissertation, I use these standards as the key document to guide test development. In the following section I also highlight which standards guide the development of LATs.

## Using the Standards to Create a Valid Test

The key reference document outlining how to create a valid test is the *Standards* (AERA et al., 2014). According to Camara (2014), the *Standards* provides "definitive technical, professional and operational standards for all forms of assessments that are professionally developed and used in a variety of settings." Within the *Standards*, there are several specific references that relate to LATs. In the following section, I will walk through the standards which apply and discuss the implications of those standards and how LATs would meet those standards.

The first relevant standard, because it is heavily related to the focus of LATs, is Standard 3.0: "All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the intended uses for all examinees in the intended population." To argue for the validity of a LAT, one would have to make the argument explicitly that the goal of a LAT is to reduce the CIV that may be caused by the language of the test items.

Related, and perhaps just as important are other standards related to Standard 3. For example, Standard 3.1 says, "Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population." In other words, test designers need to attempt as best as possible to create tests that increase the validity among various subgroups. Additionally, Standard 3.2, which is similar to Standard 3.0, more specifically calls out the possibility that a test can be affected by cultural bias and, specifically, language related bias. Standard 3.2 notes, "Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics." Related, and more specific to education, Standard 12.3 notes, "Those responsible for the development and use of educational assessments should design all relevant steps of the testing process to promote access to the construct for all individuals and subgroups for whom the assessment is intended." So, this standard related to providing increased access which LATs should, in theory, do well.

Simply attempting to reduce construct-irrelevant issues, however, is not enough. The following standards address more practically how to do so and how to show that attempts have been made. For example, Standard 3.3 notes, "Those responsible for test development should include relevant subgroups in validity, reliability/precision, and other preliminary studies used when constructing the test." The key point here is that these subgroups need to be involved early in the test development process. This is an argument that Solano et al. (2002a) emphasize in their work on dual language assessments. Additionally, documentation is important, according to Standard 3.5, "Test developers should specify and document provisions that have been made to test administration and scoring procedures to remove construct-irrelevant barriers for all relevant subgroups in the test-taker population.” In other words, when designing a LAT, care must be taken to document what steps were taken and the justification for those steps in the lens of removing CIV. I say in the lens of removing CIV because, I emphasize that, to meet standards, the argument for why to create a test that is linguistically adaptive should be based on improving the validity for more subgroups. To fit with the requirements of the standards, the argument should be less focused on the theoretical reasoning for LATs, but rather on the pragmatic goal of simply reducing unwanted variance and, hence, improving validity.

Lastly, when discussing standards related to Standard 3.0, language translation can be tricky, hence Standard 3.12 notes, “When a test is translated and adapted from one language to another, test developers and/or test users are responsible for describing the methods used in establishing the adequacy of the adaptation and documenting empirical and logical evidence for the validity of test score interpretations for intended use.” Meaning, the process of translating the test must be carefully and validly done, including back translating test items to ensure accuracy and giving pilot items to various subgroups to ensure proper interpretation. Also related to language, Standard 3.13 indicates, “a test should be administered in the language that is most relevant and appropriate to the test purpose.” Here, again, this Standard is an argument for a LAT, and again, related to reducing CIV. Finally, as a test developer, focus must also be on the use of the test not just on receiving a more valid score. Standard 3.15 suggests, “The test developers and publishers who claim that a test can be used with examinees from specific subgroups are responsible for providing the necessary information to support appropriate test score interpretations for their intended uses for individuals from these subgroups." In other words, one should not just design a test to be accurate for various subgroups, but also ensure that those who receive the scores know how to use the test scores received.

The previous paragraph related to Standard 3.0, which most closely tied to the specifications of a LAT; however, other standards, though potentially less specific to LATs, are probably more important to creating a valid test. For example, the first standard, Standard 1.0, indicates that a "clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided." This standard, which appears to reflect Messick and Kane's beliefs about validity means that a LAT, and is related to Standard 3.15, would need to be used to help indicate how the test is meant to be used with regards to specific populations. The next standard, Standard 1.1, suggests that "the test developer should set forth clearly how test scores are intended to be interpreted and consequently used. The population(s) for which a test is intended should be delimited clearly, and the construct or constructs that the test is intended to assess should be described clearly." Developers of a LAT would need to clearly state the construct being measured and the target population. Following on that standard, the next standard, Standard 1.2 notes that "a rationale should be presented for each intended interpretation of test scores for a given use, together with a summary of the evidence and theory bearing on the intended interpretation." So, for a LAT, proof would also need to be presented justifying why the items get at the construct.

Other more obvious standards that would need to be followed to create a valid LAT include, Standard 1.8, which notes that the sample used to obtain validity evidence would need to be described in detail. In addition, Standard 2.0 indicates "Appropriate evidence of reliability/ precision should be provided for the interpretation for each intended score use." So, the test would need to be proven to be reliable, including for various subgroups (which is, incidentally, the content of Standard 2.3).

Perhaps of more interest is Standard 1.25, which indicates that "when unintended consequences result from test use, an attempt should be made to investigate whether such consequences arise from the test's sensitivity to characteristics other than those it is intended to assess or from the test's failure to fully represent the intended construct." In this case, one should not just develop the test, but also monitor the unintended consequences of the test to see if it is caused by construct-irrelevant issues. This is also evident in Standard 12.1 which is more specific to educational use:

> When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described by those who mandate the tests. It is also the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences as feasible. Consequences resulting

> from the uses of the test, both intended and unintended, should also be examined by the test developer and/or user. (p. 195)

Finally, several of the standards relate to documentation. Standard 2.1 notes, “The range of replications over which reliability/precision is being evaluated should be clearly stated, along with a rationale for the choice of this definition, given the testing situation.” Standard 4.0 suggests, “Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, validity for intended uses for individuals in the intended examinee population. Also, Standard 4.3 notes:

> Test developers should document the rationale and supporting evidence for the administration, scoring, and reporting rules used in computer-adaptive, multistage-adaptive, or other tests delivered using computer algorithms to select items. This documentation should include procedures used in selecting items or sets of items for administration, in determining the starting point and termination conditions for the test, in scoring the test, and in controlling item exposure. (p. 86)

Additionally, Standards 4.4, 4.7, and 4.10 all relate to documenting the validity of different test versions, the procedures used to select items from the item pool, and the psychometric properties used to screen items, respectively.

### How A Linguistically Adaptive Test Could Effectively Target Language Bias

A LAT is not a silver bullet meant to solve all the above discussed issues with validity and bias. In fact, creating a LAT is not radical in the scheme of other attempts to solve cultural bias. This is not a method that challenges the fundamental theory of psychometrics. Instead, a LAT specifically targets van de Vijver and Tanzer's (2004) item bias and Reynolds et al. (1999) third potential cause of testing bias, language related issues, by providing a test that allows the test-taker to read items in both languages. Note that this is only the tip of the bias iceberg. LATs are a somewhat practical approach to a greater problem. They do not attempt to solve some of the more complex types of cultural bias such as testing constructs that have different meanings in one culture or another. However, even small amounts of progress on high-stakes tests can have profound effects on certan subgroups who may have experienced bias on licensure exams, college entrance exams, and placement exams which have a direct impact on future earning and social status.

In terms of validity, LATs can be used to target CIV issues where the delivery of the test (i.e., the language of the test or the way a question is worded) may cause variance in test scores among subgroups that has nothing to do with the test-takers relationship to the underlying construct that is meant to be measured by the test (Haladyna & Downing, 2004).

A LAT is a test that attempts to reduce language bias in tests by taking advantage of the affordances of CATs. With CATs, test items no longer follow the typical standardized format in which every student receives the same items in the same order. Instead, CATs adapt the items that students receive based on an algorithm. A LAT would follow this format by changing the language of the items that students receive based on whether they click a toggle button. For example, a test-taker who states that that they typically speak Spanish at home, but English at school, may receive items to a mathematics test in English with the option to toggle back and forth into Spanish. The purpose is to remove construct-irrelevant barriers from subgroups who may be disadvantaged by the format of delivery for the test item.

Typically, CATs reduce the overall time that students spend taking a test but provide the same or result in a more accurate score. CATs achieve this feat by reducing the number of items that a student answers and avoiding repetitive items after a student has already demonstrated a certain level of mastery. For example, students who answer several "easy" items correctly in the first part of the exam will prove mastery and see fewer and fewer easy items. CATs, which were imagined in the 1970s and developed and launched over the next two decades are increasing in popularity and already have a foothold in several high profile tests including the Graduate Record Examination (GRE), the Graduate Management Admission Test (GMAT), the U.S. Armed Service Vocational Aptitude Battery (ASVAB). Even the recently passed Every Student Succeeds Act (ESSA) supports increased use of CATs for student assessments (Patz, 2015).

A linguistically adaptive CAT would follow the same logic as current CATs except the goal would not be to reduce the time that a test takes, but rather give students more appropriate questions regarding their linguistic background. The approach is also like tests that make accommodations for different learning abilities or students with various linguistic backgrounds. In a LAT, students would be allowed to freely choose between English and their dominant language, essentially allowing students to read items in multiple languages helping to eliminate bias caused by the language of item delivery.

## Conclusion

A LAT cannot be used to address every deep-rooted theoretical conflict between psychometrics and sociocultural theory, but it can present a pragmatic approach to incremental improvements targeted at improving the validity of scores for those who may be hamstrung by the presentation of items in only one language. Ultimately, a LAT aims to meet the *Standards* to improve validity by reducing CIV and, therefore, improving fairness for a wider range of subgroups.

# CHAPTER 3

# LITERATURE REVIEW OF METHODS

The options for testing the efficacy of a LAT are very similar to testing the efficacy of any new type of exam (i.e., testing to see how the new type of test compares to the old type of test for a target population). However, a LAT could cause unexpected effects for the test-takers leading to alternative CIV issues. The following section will review relevant experimental or quasi-experimental design options for testing the efficacy of the tests in question.

## Experimental Designs that Test for Mean Differences and Equivalence

Rivera and Stansfield (2001) conducted a study on the efficacy of a linguistically simplified standardized test on LEP students. To do so, they randomly assigned eight test forms, four of which were linguistically simplified, to 11,415 students, 109 of which were LEP students. To test for differences, they conducted a t-test on mean raw scores. According to Cook et al. (2002), this experimental design is the basic design for random assignment because there are two conditions (simplified and not simplified exams) and a posttest. The random assignment of the two conditions theoretically accounts for potential selection bias making this design simpler to conduct than most quasi-experimental designs.

In a more complicated experimental design on linguistically simplified standardized test items, Abedi et al. (1998) conducted a study in which they randomly assigned three forms of the National Assessment of Educational Progress (NAEP) mathematics tests and reading tests (original English, linguistically modified English, original Spanish) to 1394 Grade 8 students. To analyze the results, they conducted a two-factor design by booklet type and LEP status to test mean differences between the groups and interactions. They also solicited responses via in-depth questionnaires about the language and cultural backgrounds of the participants, and then used those answers to conduct a multiple regression analysis with mathematics and reading scores as the dependent variable to see if any of the factors from the questionnaire were good predictors for student scores. Another study published in 2000 (Kiplinger et al., 2000), also conducted to test various language accommodations on testing, also implemented a factorial analytical design and analyzed results with a set of 2x2 ANOVAs to test the mean differences of various conditions. This design is also found within articles in which the authors test for the effects of computer-based tests versus paper and pencil tests (Booth-Kewley et al., 2007; Carlbring et al., 2007). Conducting a factorial design would be a strong choice for a LAT because there are several potential levels of factors including participants' dominant language and whether the participant received a linguistically adaptive test or not. Plus this design allows for interactions to be probed and allows for smaller sample sizes that may otherwise be needed (Cook et al., 2002).

Other researchers who examined mean differences in test types took slightly different approaches to their research designs. McCoy et al. (2004) in a somewhat problematic design tested a computer-based questionnaire with a class of students and then, two weeks later, tested a paper and pencil-based version of the questionnaire. They then compared the mean scores of the two times. Such a design does not have a strong control for practice effects (i.e., improving on answering the questionnaire) or effects from time passing (e.g., the class has two more weeks of exposure to the material). Williams and McCord (2006), on the other hand, tested using random assignment and a counter-balanced repeated measures design with four conditions: a group that took the computer-based version of the test then later the computer-based version again, a group that took the standard version of the test then the standard version again, a group that took the computer-based version first followed by the standard version, and, finally, a group that took the standard version of the test followed by the computer-based version. The counter-balanced design avoids a common issue with repeated measures effects by controlling for any testing or time effects (Cook et al., 2002).

A few other authors used experimental designs that did not use random assignment. Wolfe and Manalo (2004), for example, allowed participants to select their own condition. To control for selection bias, they conducted a general linear model controlling for demographic characteristics and English language proficiency. Puhan et al. (2007), in an interesting design, also assigned conditions in a non-random way: however, they used propensity score matching to control for selection bias. Additionally, they did some analysis at a smaller grain size, at the item-level. In this case, they conducted DIF analysis on the mode of test (i.e., computer-based, or otherwise).

Gallagher et al. (2000) of the Educational Testing Service (ETS), also did a counter-balanced repeated measures approach like Williams and McCord (2006), but their goal was to explicitly measure CIV between a computer-based test and a paper version. So, the design was very similar, with just the hypothesis differed.

## Accommodations and Experimental Designs

The intent of accommodations is not to give unfair assistance. Instead, accommodations should, "make an assessment more accessible for English language learners and students with disabilities and to produce results that are valid for these students" (Abedi & Ewers, 2013, p. 4). Accommodations help to control for CIV: "As such, they may also be considered for all students as accessibility features because they control for sources of construct-irrelevant variance" (Abedi & Ewers, 2013, p. 5).

In the previous sections I focused on techniques to measure if one test form was equivalent to another among various populations. However, when testing the efficacy of a LAT, the most important point is to prove that the test has reduced CIV for the targeted subgroups. Some of the literature most closely related to testing this type of efficacy (i.e., not just testing if two forms are equivalent) is the literature regarding test accommodations. According to Sireci et al. (2005) research on test accommodations raises the following questions, "Do the test scores that come from nonstandard test administrations have the same meaning as test scores resulting from standard administrations? And do current test accommodations lead to more valid test score interpretations for certain groups of students?" (p. 457). To test accommodations, many authors set up an experimental design to test a certain hypothesis called the *interaction hypothesis* or the *maximum potential thesis* (National Research Council, 2004; Zuriff, 2000). The basic concept is that a test accommodation should help the group for which it is targeted, but not help a group for which it is not targeted. This would show that the accommodation is not an unfair crutch, but rather a way of reducing CIV that only affects a certain subgroup.

Many of the studies in literature testing the interaction hypothesis involve experimental designs that share features with some of the experimental designs testing for equivalence in the previous section. For example, Elliott and Marquart (2004) conducted a repeated measures counterbalanced design in which three groups of students completed two different and equivalent test forms in different testing conditions, one with an accommodation for time and another without an accommodation for time. This design is similar to two studies from the previous section (Gallagher et al., 2000; Williams & McCord, 2006). Interestingly, Elliot and Marquart (2004) also sent a follow-up survey to gauge participant reactions to the accommodations. Though, theoretically a mixed-methods design, the follow-up survey was almost an afterthought and not complementary to the main analysis of their hypothesis.

Weston (2003) conducted a similar experimental design to test an accommodation, but this particular experimental design also tested the effects on students with various reading proficiency levels. The strong experimental design has many pieces that could be potentially adopted by a test of a LAT. For the first part of the design every student took two forms of equivalent mathematics tests, one with the accommodation and one without; however, the researchers randomly assigned the participants the order in which they took the two tests to create a counter-balanced design. Then the researchers had students take reading tests, and they also collected a survey from teachers rating their students' mathematics and reading abilities. Lastly, the researchers conducted interviews with teachers and students to discuss their views of the accommodation. The key to the relevance of this research study to a LAT is that the authors did not simply test to see if the accommodation affected only the target group, the researchers also designed the experiment to evaluate if reading ability played a role by comparing the scores on the reading test with differences in the two testing conditions on the mathematics tests. Lastly, the mathematics tests themselves were cleverly designed by the researchers to include some items that were word problems and some that were calculation problems. Examining the participant scores on those items would give even more insight into the role that reading ability may play and show that, "learning disabled students benefited more than regular classroom students from the accommodation on word problems, but the groups benefited equally from the accommodation on calculation items" (p. 10).

## Evaluating the Dual Language Translation of Test Accommodations

There are too many types of accommodations to write a comprehensive literature review, however I can focus on accommodations for ELLs. Pennock-Roman and Rivera (2011) state that there are 104 types of test accommodations, although many of those accommodations relate to students with disabilities. Fewer accommodations are designed for ELLs. Some of these types of accommodations include having extra time, access to an

English dictionary, and access to a computer-based pop-up glossary. Pennock-Roman and Rivera (2011) classify 11 types of accommodations related to ELLs:

> Plain English (called “simplified English” by Kieffer et al.), bilingual glossary (paper and pencil versions), Spanish version, and extra time. A fifth type, English dictionary/glossary, reflected paper and pencil versions of this accommodation with the computer-administered version, pop-up English glossary, comprising a sixth category. The seventh category, dual language (DL), … Three additional types of accommodations in the present synthesis—picture dictionary, pop-up bilingual glossary, and read aloud (in English) that were administered individually or in pairs … The eleventh type was the small groups accommodation where the unchanged test was administered in small groups. (p. 13-14)

Within these 11 types of accommodations, dual language translation of test (DLTT) is an accommodation mostly closely related to LATs. For DLTT, tests are translated and placed side-by-side. Essentially, through my study I extend this accommodation to include a computer-based toggle screen to allow for better translanguaging, as noted earlier. LATs combine some of the features of bilingual glossaries and dominant language version of tests into a package that is more cognitively simple.  Pennock-Roman and Rivera (2011) also state that such types of accommodation research is needed: “Although the research shows that the most effective methods were English dictionary/glossary and bilingual accommodations with generous time limits (for most ELLs) or dominant language versions (for ELLs with low ELP), studies of these accommodations are scarce" (p. 21).

According to Abedi and Ewers (2013), there are authors of four studies who conducted DLTTs (Pennock-Roman & Rivera, 2011; Abedi, Courtney, Leon, Kao, & Azzam, 2006; Duncan et al., 2005; Sireci et al., 2003). Pennock-Roman and Rivera (2011), conducted a meta-analysis to find that DLTT was only slightly effective given extra time. Within their meta-analysis, they found three studies in which the authors studied dual language forms (Aguirre-Muñoz, 2000; Duncan et al., 2005; Liu et al., 1999). Authors of each of these studies used random assignment to compare two groups of students in the testing condition. Abedi and others (2006) also studied DLTT, finding that DLTT was not effective and required generous amounts of time, though they cited some limitations to their study. They used hierarchical linear modeling (HLM) to study the effects of classroom differences, two types of randomly assigned accommodations (including DLTT), and ELL status on their dependent variable, test scores. Results indicated that the two accommodations made no significant difference in test performance. However, they stated that one major limitation to their study was that they offered the accommodation (e.g., the DLTT with Spanish and English) to students who were not fluent in the language of the accommodation (i.e., not fluent Spanish speakers). In other words, the accommodation may have been superfluous for many of their participants.

Duncan and colleagues (2005), on the other hand, found that DLTT was an effective intervention on Grade 8 mathematics assessments. Duncan et al. tested the efficacy of the intervention by randomly assigning two conditions, the typical test booklet and the new side-by-side translation test booklet, to a group of over 400 Grade 8 students. The researchers then used DIF analysis and multivariate regression to test for problematic items and significant mean differences, splitting the participants into groups based on whether they had three years or more of English-language instruction. Duncan found that English proficiency levels played a large role, but, in general the accommodation proved helpful for students with less English proficiency.

I tested a similar type of intervention, in a similar style; however, the accommodation in question took advantage of a computer-based toggle system to avoid some of the pitfalls of DLTT such that, according to Pennock-Roman and Rivera (2011), “the booklet can be very long, requiring additional time for students to page through the materials. Alternatively, the presentation of items in two languages may have been confusing to students unfamiliar with this format” (p. 20). However, instead of one randomly assigned observation, I tested two observations at different time points in a counterbalanced design allowing for smaller sample sizes. Importantly, authors of these studies emphasized the need to control for language background and the proficiency levels of the target construct. As such, I ensured such demographic details were captured in my research design.

## Cognitive Interviews

Many of the previous designs were meant to test the result of an intervention or to test if one test form was equivalent to another test form. However, there are other ways to detect CIV in tests. Quantitative methods may be good at detecting bias, but qualitative methods may be better at explaining this bias (Benítez et al., 2018). One such qualitative method is cognitive interviewing, which encapsulates several slightly different interview methods.

Cognitive interviewing is typically used in one of two ways, either as a way of ensuring that test items make sense during test development or as a technique to determine and explain some bias that has already been detected. Miller et al. (2014) break that distinction down even further stating that cognitive interviews can be used to study construct validity and also comparability across socio-cultural groups. As far as using cognitive interviews to ensure the quality of test items, Collins (2003) states that cognitive interviews are primarily used for: "Pre-testing questions, particularly pre-testing questions in their questionnaire context, enables us to establish whether: – respondents can understand the question concept or task, – they do so in a consistent way, and, – in a way the researcher intended" (p. 231). Collins continues to explain that there are two main cognitive techniques: think aloud interviewing and probing. In the think-aloud method, the participants are prompted to think aloud as they go through the items, essentially giving the researchers rich verbal data about reasoning and showing which parts of items garner attention or confusion. Probing, on the other hand, gives the researcher more flexibility and control. Probing, which means the researcher asks specific questions, can take place as the participant is taking the test or afterward. Also the researcher can choose to ask questions about comprehension (e.g., What does that term mean to you?), retrieval (e.g., how did you remember that?), confidence judgement (e.g., how sure are you of your answer?), and response (e.g., how did you feel about answering that question?) (Collins, 2003, p. 235). Both think-aloud and probing can be combined.

As for using cognitive interviewing to explain some detected bias, authors of many studies show that the technique can be used to explore what the causes of the bias may be. Benítez et al. (2018) conducted cognitive interviews to detect bias in a test retrospectively, with participants first taking the scale then answering probing questions. In another study, Benítez and Padilla (Benítez & Padilla, 2014) first conducted DIF analysis to test for biased items between linguistic groups, then followed up on the DIF analysis by conducting cognitive interviews with participants to show different interpretation pattern between groups. Winter et al. (2006) took it a step further and conducted a very interesting design for their cognitive interview in order to investigate interactions between item types and individual factors. They explored how, "a diverse group of students (including current and former English language learners, poor readers, and special education students) attempted to solve mathematics problems and identify features of items that inhibited or promoted their problem-solving" (p. 269). They used focused interviews to specifically explore how certain participant factors affected three stages of mathematics problem solving: apprehension of task demands, formulation of the solution, and articulation of the solution. Smits (2005) also explored bias by conducting semi-structured interviews using van de Vijver and Leung's (1997) bias framework to frame their analysis of the interview transcripts.

## Summary

In the previous section, I covered a review of cultural bias, validity, the assumptions of standardizes tests, and methods used in previous similar studies. Through literature, I identified how linguistic bias is a small piece of wider cultural bias found in many facets of education. I also identified how other researcher have attempted to study and counteract these types of biases found in testing. In the following section, the methods section, I will detail how I approached the design of a novel way to deliver test items meant to counteract potential linguistic bias.

# CHAPTER 4

# METHODOLOGY

Via this research study, I sought to address language bias by building and testing the efficacy of a novel type of test delivery that leverages adaptive testing to change the language of test items for different students. The following overarching question guided the study: *Can a linguistically adaptive test provide a more valid measure of the target construct than the usual test for multilingual learners?* However, to answer the question, I needed to a) select a test to adapt, b) develop the test in a culturally sensitive way, and c) test the efficacy of the test. Therefore, for this study I incorporated multiple phases of implementation and inquiry: a) a pre-phase to ensure that the selected test demonstrates linguistic issues, b) a first phase of cognitive interviews to learn the magnitude and the ways in which items show linguistic bias, c) a second phase consisting of test development, and d) a third and final phase to test the efficacy of the new test. Using mixed-methods as a guiding methodology and pragmatism as a guiding philosophy, I split the overarching question into a series of sub-questions for each phase of the study:

Phase One:

1. Which test items, and how many test items, cause confusion or misconceptions for linguistic reasons for Chinese and Spanish dominant speakers when reading the PCA?

2. Is there a pattern of confusion or misconception that matches the types of test items as classified by Jamal Abedi and colleagues' linguistic complexity framework (2002; 2012; 2001)?

Phase Two:

3. Is there preliminary evidence for the validity of the new language-adaptive concept assessment as evidenced by participant explanations of their selected answers?

Phase Three:

4. Is the new language-adaptive concept assessment reliable and valid?
5. Do Chinese and Spanish speakers perform better on the new language-adaptive concept assessment compared to the original English version and the non-adaptive translated version?
6. What perceptions do test-takers have about the new language-adaptive concept assessment?

The variety of research study phases and research questions necessitated myriad methods. Hence, I relied on methods ranging from cognitive interviews with students to larger scale within-subjects random-assignment trials. The following section lays out the outline of methods used.

**Outline of Methods**

For the first phase, I conducted cognitive interviews. For the second phase, I conducted cognitive interviews as well. For the third phase, I conducted a within-subjects random-assignment trial comparing the effect of the test type (i.e., non-LAT and the LAT). In addition, I conducted in-depth interviews with a subset of the participants to understand their perceptions about the test (see Table 1).

**Table 1**

*How Methods Map onto Phases and Research Questions*

| Phase | Research Question | Method |
|---|---|---|
| One | Which test items, and how many test items, cause confusion or misconceptions for linguistic reasons for Chinese and Spanish dominant speakers when reading the PCA? | Cognitive Interviews |
| | Is there a pattern of confusion or misconception that matches the types of test items as classified by Jamal Abedi et al.'s linguistic complexity framework | Cognitive Interviews |
| Two | Is there preliminary evidence for the validity of the new language-adaptive concept assessment as evidenced by participant explanations of their selected answers? | Cognitive Interviews |
| Three | Is the new language-adaptive concept assessment reliable and valid? | Calculate Cronbach's Alpha, Calculate Test-Retest Reliability, Demonstrate the Five Sources of Validity Evidence |
| | Do Chinese and Spanish speakers perform better on the new language-adaptive concept assessment compared to the original English version and the non-adaptive translated version? | Within-Subjects Random-Assignment Trial |
| | What perceptions do test-takers have about the new language-adaptive concept assessment? | In-depth Interviews |

## Explanation of Methods Chosen: A Mixed-Methods Approach

The strength of using mixed methods in this study, such as combining findings from cognitive interviews and test score comparisons, is the ability to explore the tension between the forced standardization of standardized tests and the unique and multifaceted nature of test-takers. Along these lines, Saakvitne et al. (2005), discussing the design of culturally-sensitive measures, recommend that researchers combine both personal narratives and population research to create measures, and this combination should happen as measures are developed, not afterward. So, for a LAT, the key is not about how to create a new test, but rather how to adapt and translate a test to become a valid test across cultures. The goal is a pragmatic improvement to the current testing procedure with an emphasis on improving the validity for non-dominant English speakers. As worded by Ungar and Liebenberg (2011), "Cross-national studies face a related problem: How do we balance assumptions of homogeneity across Minority and Majority World contexts with the need for sensitivity to within group and between group heterogeneity?" (p. 129).

When looking at a successful adaptation there are several aspects to consider: a) the test format may be more familiar to one cultural group than another, b) the test administration may have issues that affect bias, c) the construct may not have equivalence across cultures, and d) a speeded test may be more familiar to one culture than another (Hambleton et al., 2004; Van Leest & Bleichrodt, 1990). Most of these issues are related to familiarity with the process of taking the test which means that finding solutions to these issues requires asking people who are part of the target group to give their opinions on their comfort level with these factors. The anecdote of this issue is that during the test development process a group of people who is familiar with the target culture ensures that the question format makes sense. Via this three-phase dissertation, accordingly, I emphasize a pragmatic approach to solving a problem by increasing understanding of a problem. The approach uses a mixed-methods approach to enrich data analysis.

***Pragmatism and Mixed Methods***

Johnson and Onwuegbuzie (2004) argue that the time for mixed methods has arrived. They argue that the previously common belief in the incompatibility thesis, an idea introduced by Howe (1988), was slowly dying in education research and the social sciences in general, to be replaced by more pragmatic approaches. According to Howe (1998), the incompatibility thesis rests on the following beliefs: "The positivist and interpretivist paradigms are incompatible; the positivist paradigm supports quantitative methods, and the interpretivist paradigm supports qualitative methods. Therefore, quantitative and qualitative methods are, despite the appearance that research practice might give, incompatible" (p. 13). Howe stated that the number of those believing in the incompatibility thesis was on the wane in 1988 for a few major reasons. One, the objectiveness of the positivist tradition was beginning to be questioned in the academy. Two, pragmatism following from the tradition of Dewey and others became to be more widespread which allowed for more emphasis on what works rather than what is theoretically logical.

Johnson and Onwuegbuzie (2004) state that mixed methods is "the class of research where the researcher mixes or combines quantitative and qualitative research techniques, methods, approaches, concepts or language into a single study" (p. 17). However, good mixed methods should go beyond simply mixing or combing quantitative and qualitative inquiry into a single study. A strong mixed method design uses quantitative and qualitative inquiry to enrich the findings of each other.

Creswell and Plano Clark (2017) state that there are three ways to mix methods: a) converge the datasets by bringing them together, b) connecting the two datasets so one set of data builds on the other, and c) embedding one dataset in the other in a supportive role. This can be thought of, for example, as a) collecting two types of data for the same analysis, b) collecting some exploratory qualitative data followed by quantitative inquiry, and c) collecting quantitative results then confirming them using qualitative data. According to pragmatism, there is no best way. In fact, some data may not always agree.

Luyt (2012) rejects the notion that the strength is in triangulation, which is a common paradigm of mixed methods that has been part of the conversation since the 1970s (Denzin, 2012). Triangulation, to see the same thing from different viewpoints, in Luyt's opinion, implies that there is some objective truth to be discovered if looked at from multiple perspectives. Hesse-Biber (2010) agrees stating that triangulation puts the qualitative data in the position of simply providing support for the positivistic findings of the quantitative data. For Luyt, the key distinction is that a focus on the complementarity of methods means that researchers must accept and learn from times when the data are inconsistent, and that inconsistency is acceptable. The reason this inconsistency is acceptable is that the methodology underlying the method allows for multiple truths. As such, not every one of the findings to the research questions may match. Mismatched findings would be significant and reflect the difficulty and nuance needed in test design. The purpose of approaching a dissertation with mixed methods is to allow a space for multiple truths when evaluating a new type of test accommodation.

## The Validity and Reliability of the PCA

The creators of the PCA followed a three-phase approach to developing the PCA items (Carlson et al., 2010). In the first phase, researchers identified a theoretical framework that included 11 abilities critical for calculus success (Carlson, 1998). Working from this framework, the item development team appropriated 34 items from other instruments that they believed addressed abilities within the framework. Additionally, they wrote another 13 items that they believed addressed previously unaddressed abilities within the framework. A series of four experts in mathematics education reviewed items to ensure that they tested the abilities in question. Additionally, the research team tested the items in clinical interviews with potential test-takers to ensure that the question meanings were clear and interpretable. The test items were then converted to multiple-choice items.

In the second phase, the test items were used to conduct studies about how potential test-takers reason about and understand the core concepts being tested. A series of seven studies were conducted from 1997 to 2008 (Carlson, 1997, 1998; Carlson &

Bloom, 2005; Carlson et al., 2010; Engelke, 2007; Engelke, Oehrtman, & Carlson, 2005; Oehrtman, Carlson, & Thompson, 2008). The data were used to write theoretically strong distractors that match with common misconceptions or faulty reasoning. These distractors would later be piloted in the third phase.

In the third phase, to establish internal content-related validity evidence, the creators of the PCA cycled through eight iterations of testing and rewriting. During these iterations, if any distractors attracted less than five percent of respondents, the researchers replaced the distractors. Additionally, researchers required students to write justifications for their answer choices with each iteration of test writing. The research team compared the justifications with the answer choice selected to ensure that the process of thinking matched the choice selected. Lastly, the research team conducted clinical interviews with each iteration of test writing to gather more data on whether each test-taker's thought process matched the meaning of the answers that they selected.

Overall, the research team followed test creation best practices by first defining the construct carefully through literature, multiple studies, and checking with experts. Second, they wrote initial items, drawing from literature then checked those items with experts. Third, they conducted studies and drew from literature to understand which distractors would be most effective. Fourth, they wrote and rewrote their items through an iterative process that incorporated pilot groups and clinical interviews.

## Pre-Phase: Linguistic Analysis and Differential Item Functioning

I conducted a pre-phase to ensure that the PCA demonstrates the potential to include linguistic issues. I needed to select a test, like the PCA, with a strong history of demonstrated validity in multiple contexts, but also with the potential to be improved significantly by being delivered in a linguistically adaptive style. Hence, I approached the pre-phase by examining both the linguistic content of the test items, but also to see if there were discrepancies among test scores among various subgroups. In other words, even in the pre-phase I used a mixed-methods approach to evaluate the feasibility of using the test as my base test.

### *Pre-Phase Study One: Linguistic Analysis Methodology Study of the PCA*

Jamal Abedi and colleagues (2002; 2012; 2001) developed and used a method for analyzing and adjusting complex language in mathematics items. Abedi created a list of ways that mathematics question language can be difficult to interpret including unfamiliar vocabulary, complex grammatical structures, nominalization, multiple embedded clauses, and passive voice constructions (see Table 2). In Table 2, I present each type of issue with a good and bad example showing what the linguistic issues look like as pieces of text within a test item.

**Table 2**

*Framework of Linguistic Complexity Issues with Examples*

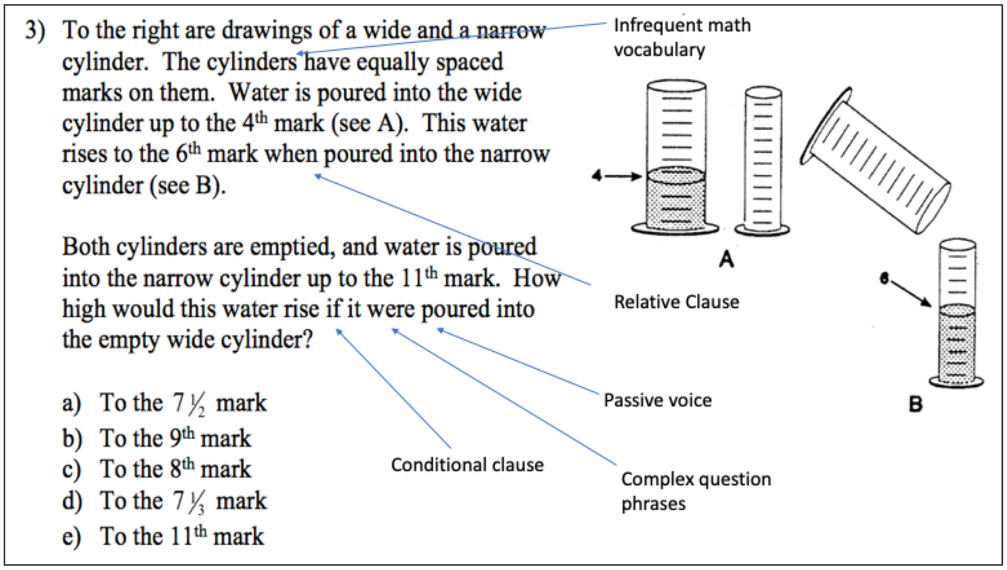
| Issue | Bad | Good |
|---|---|---|
| Infrequent non-mathematics vocabulary | "a certain reference file" | "Mack's company" |
| Passive Voice of verb phrase | "if a marble is taken from the bag" | "if you take a marble from the bag" |
| Long nominals | "the pattern of the puppy's weight gain" | "the pattern above" |
| Conditional clauses | "if two batteries in the sample were found to be dead" | "he found three broken skateboards in the sample" |
| Relative clauses | "the total number of newspapers that Lee delivers in 5 days" | "how many newspapers does he deliver in 5 days" |
| Question phrases | "which is the best approximation of the number" | "approximately how many" |
| Abstract or impersonal presentations | "… 2,675 radios sold" | "… 2,675 radios that Mrs. Jones sold" |

Based on the preceding issues and using the examples to guide interpretation, I walked through each of the 25 items on the PCA simply indicating whether the question exhibited one of the seven issues in Abedi and colleague's framework. I used the sum to calculate a rudimentary level of complication. I also counted the total number of non-mathematics words or expressions that were required for each item as a separate indicator of complexity (Abedi, 2002).

As an example of coding, Item three contains several complex language and grammatical structures. I coded the word "cylinder" as infrequent mathematics vocabulary. I coded the sentence, "This water rises to the 6th mark when poured into the narrow cylinder," as a relative clause (see Figure 1).

**Figure 1**

*An Exemplar of Complex Language in Mathematics Items*



3) To the right are drawings of a wide and a narrow cylinder. The cylinders have equally spaced marks on them. Water is poured into the wide cylinder up to the 4th mark (see A). This water rises to the 6th mark when poured into the narrow cylinder (see B).

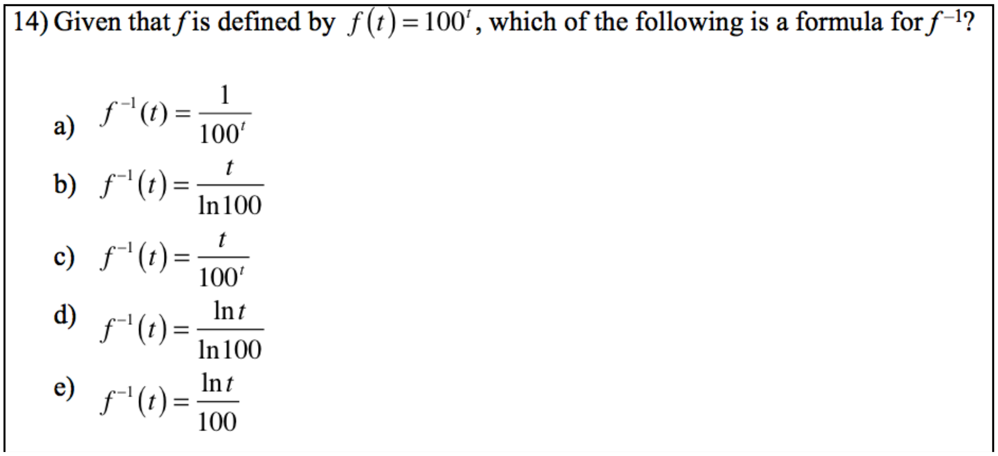
Both cylinders are emptied, and water is poured into the narrow cylinder up to the 11th mark. How high would this water rise if it were poured into the empty wide cylinder?

a) To the $7\frac{1}{2}$ mark
b) To the 9th mark
c) To the 8th mark
d) To the $7\frac{1}{3}$ mark
e) To the 11th mark

Additionally, I found passive voice, a conditional clause, and a complex question phrase in the second paragraph of the item. Hence, I coded item 3 as the most complex item in the PCA. In contrast to item three, other items demonstrated straightforward language. For example, Item 14, contains a short question without complicated language (see Figure 2).

**Figure 2**

*An Exemplar of Straightforward Language in Mathematics Items*



14) Given that $f$ is defined by $f(t) = 100^t$, which of the following is a formula for $f^{-1}$?

a) $f^{-1}(t) = \frac{1}{100^t}$

b) $f^{-1}(t) = \frac{t}{\ln 100}$

c) $f^{-1}(t) = \frac{t}{100^t}$

d) $f^{-1}(t) = \frac{\ln t}{\ln 100}$

e) $f^{-1}(t) = \frac{\ln t}{100}$

Item 14 has a simple straightforward question statement using common mathematics vocabulary, "which of the following is a formula …". Additionally, test-takers may be able to solve item 14 without reading the words because the mathematics statements present a clear problem to solve. Applying these findings led to a better understanding of the results of the cognitive interviews in phase one discussed later.

### *Pre-Phase Study Two: Differential Item Functioning of the PCA*

I conducted the second study of the pre-phase to determine the likelihood that the PCA would include items that exhibit some form of linguistic bias among students by looking for evidence of linguistic bias in a similar, older version of the PCA, called the PPCA. I chose to evaluate linguistic bias in the PPCA because I had access to data from 346 test-takers.

**Data Description.** This data had been previously collected at the beginning of the Spring 2017 from nine classrooms of students who took the PPCA at a large university in the U.S. Southwest. I calculated basic descriptive statistics on the sample of 346 test-takers to ensure a wide range of scores. The mean score was 19.45 with a standard deviation of 4.15. Importantly, the scores did not seem to indicate ceiling or floor effects because not one single test-taker achieved either full points or no points on the exam. The maximum score was 29 out of 30 (96.7%) and the minimum score was 8 out of 30 (26.7%) points (See Table 3).

**Table 3**

*Sample Size and Descriptive Statistics for the PPCA*

| Descriptive Statistic | PPCA |
|---|---|
| Sample Size | 346 |
| Median Score | 20/30 |
| Mean Score | 19.45/30 |
| Standard Deviation | 4.15 |
| Min Score | 8/30 |
| Max Score | 29/30 |
| Skew | 0.01 |
| Kurtosis | -0.41 |
| Standard Error | 0.22 |

However, as this was secondary data, I did not have control over the test proctoring or the instrument. Hence, the only data I received included test scores and participant names.

**Analysis.** Using only participant names, based on my assumptions, I coded the participants into two groups: dominant English speakers and non-dominant English speakers. My coding method, based on names only, was clearly imprecise and biased, but, as this was a pilot test, a rough analysis was acceptable. According to Vartanian (2010), "One of the problems with secondary data is lack of control over the framing and wording of survey items. This may mean that questions important to your study are not included in the data" (p. 15). In this case, my data lacked important demographic data, so I estimated as best as possible.

To look for evidence of linguistic bias, I conducted a statistical test to determine whether test items functioned differently for non-dominant English speakers compared to dominant English speakers. So, I conducted Mantel-Haenszel DIF using linguistically simple items as an anchor (items 13,17,21, and 22) and non-dominant English speakers as the focal group (Clauser & Mazor, 1998; Woods, 2009). The Mantel-Haenszel DIF indicates differences in item function between two groups using a chi-square statistic. If the chi-square statistic is significant with a *p*-value less than .05 than the item statistically significantly functions differently.

Additionally, the effect size codes of Mantel-Haenszel DIF include a letter system based on a scale developed by the ETS where A is negligible or nonsignificant, B is slight to moderate, and C is moderate to large (Zwick, 2012). Lastly, the Mantel-Haenszel DIF analysis includes a measure called the deltaMH which is important for interpretation. A negative number indicates that the item functions differently for the focal group (in this case, non-dominant English speakers) with non-dominant English speakers scoring worse on these items than they should considering their overall ability level. The findings from the analysis served to set the stage for phases one through three, by ensuring that the PCA, or earlier versions of the PCA, exhibited linguistic complexity among the items.

## Phase One: Qualitative Analysis of Linguistic Complexity

This phase addressed research question one (Which test items, and how many test items, cause confusion or misconceptions for linguistic reasons for Chinese and Spanish dominant speakers when reading the PCA?) and research question two (Is there a pattern of confusion or misconception that matches the types of test items as classified by Abedi and colleagues' linguistic complexity framework (2002; 2012; 2001)?).

### *Participants and Recruitment*

Before beginning to recruit participants, I obtained authorization for the Institutional Review Board (IRB) at my university (see Appendix B). First, I coded items from the 25-item PCA to identify whether the items showed linguistic complexity according to Abedi's framework. Next, I recruited Spanish and Mandarin Chinese speakers from a large university in the U.S. Southwest for cognitive interviews with the 25-item English language version of the test. I selected speakers of those languages because those subgroups are the largest non-dominant English speaking U.S. populations.

I recruited participants by asking for volunteers through an official announcement at the university and then through referral (see Appendix C). The referral sampling, called snowball sampling or chain sampling, carries an increased risk of sampling error above and beyond the sampling error introduced by asking for volunteer participants (Ritchie et al., 2013). However, snowball sampling was an effective way to recruit participants with specific criteria such as speaking Mandarin Chinese or Spanish as first languages. I excluded participants that did not a) identify as dominant Spanish or Mandarin Chinese speakers; b) hold a first- or second-year student status at a university or college; c) take a mathematics class in the last year. I paid participants $40 for their participation. I selected this amount because taking a standardized test in front of a researcher is not an intrinsically motivating task. According to Takahashi et al. (2016), when the task is not intrinsically motivating, incentives equivalent to standard minimum wage encouraged the most effort. This task took a little less than two hours, so I decided on the equivalent of $40 per participant or about $20 per hour.

Participants included 11 dominant Mandarin Chinese speakers (five males, six females) and 11 dominant Spanish speakers (two males, nine females) aged 18-26. I sampled until reaching theoretical saturation, meaning that I began to analyze results when collected and stopped sampling when the results did not add new information for answering the research question (Sandelowski, 2008). Practically speaking, if three participants in a row no longer added significantly to findings, meaning that the same questions present problems in the same way with only slight variations, then I condsidered saturation to be reached. I sought participants with different genders and socioeconomic statuses because research shows that such factors can affect test taking (Sadker & Zittleman, 2009). I limited my interviews to less than fifteen participants per language to limit the amount of work because cognitive labs are time-consuming (Benítez & Padilla, 2014).

### *Materials*

As I described previously, I used the PCA. In this case, I used a paper and pencil version of the test so participants could highlight the test (see Appendix A). I also gave participants a demographics instrument asking about their math background, self-reported math ability, and questions about their linguistic background (see Appendix D). Before beginning, the study, each participant signed a consent form approved by my institution's IRB (see Appendix E). A cell phone was used to audio record the cognitive interview.

### *Procedure*

For the cognitive interviews, I began by reading the script for each participant (see Appendix F). I prompted participants to read each question out loud, think aloud, and write down their work. During the two-hour interview, I recorded the conversation and asked probing questions according to Collins' (2003) probing technique to clarify comprehension, retrieval, confidence, and response (p. 235). Miller et al. (2014) state that cognitive interviews can be used to study construct validity and also comparability across socio-cultural groups.

### *Analysis*

I examined test items for each student to determine if the items met any of the following conditions for at least three of the 22 (~14%) students:

1. Participant showed mathematics ability, but either did not select the correct answer or had to guess;

2. Participant did not understand the item enough to have an opportunity to demonstrate their mathematics ability;
3. Participant selected the correct answer but showed that they did not understand the underlying math.

If so, I coded the item as problematic. If more than three students demonstrated a disconnect, I coded the items as problematic. If less than three, but at least one student demonstrated the above behaviors, I coded it as rarely problematic. My justification for this was that I wanted find a way to separate the magnitude of problematic items in a way that could match the way I separated items by magnitude when analyzing the items using Jamal Abedi et al.'s (2002, 2013, 2001) framework. Notably, some of the issues may not have been related to linguistic complexity (e.g., participants selecting the correct answer, without demonstrating the correct ability). Lastly, I compared problematic items in the cognitive interviews with items coded for Abedi's framework issues. I compared these results to the linguistic analysis from pre-phase one to address the first research question. Knowing which test items cause the most confusion and, in what way they cause confusion, leads to better test translations since tests need to be adapted as well as literally translated (Hambleton et al., 2004; Van Leest & Bleichrodt, 1990).

## Phase Two: Translation of the PCA and Configuring the LAT

This phase, including developing the LAT test into a usable format, was designed to check that the newly developed test operated as expected to a preliminary degree. This stage meant to discover user experience issues with the software, typos, issues with flow, and information about whether the computer-based items functioned in much the same way as the paper-based items. Essentially, this phase served as a pilot study to ensure that the following phase would be measuring the efficacy of the intervention, rather than inadvertently measuring the effect of test design issues. Hence, to test for preliminary evidence for validity of the newly designed LAT test, I conducted a series of cognitive interviews with 20 total participants (*n* = 20), including 10 self-identified dominant Spanish speakers and 10 self-identified dominant Mandarin Chinese speakers. Before this phase and the following phase, I received updated approval from my institution’s IRB (See Appendix G). In the following sections, I detail my methods for this phase.

### *Data Collection and Participants*

I recruited participants for this phase in the same manner as phase one. However, during this time the spread of COVID-19 limited in-person contact, so the cognitive interviews were held via Zoom (Zoom Video Communications Inc., 2020), an online video chat platform, in the late spring and early summer of 2020. My institution’s IRB reviewed and approved the changes to my collection procedure and my study necessitated by the spread of COVID-19 (see Appendix G). First, I posted an online advertisement at a large university in the U.S. Southwest (see Appendix H). Then, I asked participants to refer their acquaintances who fit my exclusion criteria (e.g., recruitment through snowball sampling). For this study, I required that participants a) identify as dominant Spanish or Mandarin Chinese speakers; b) be a first- or second-year student at a university or college; c) have taken a mathematics class in the last year. Each participant received a $40 incentive for their participation. I distributed the computer-based LAT through a link to the *Concerto* platform (Kosinski & Rust, 2011). I collected consent forms via the first page of the computer-based test (see Appendix I).

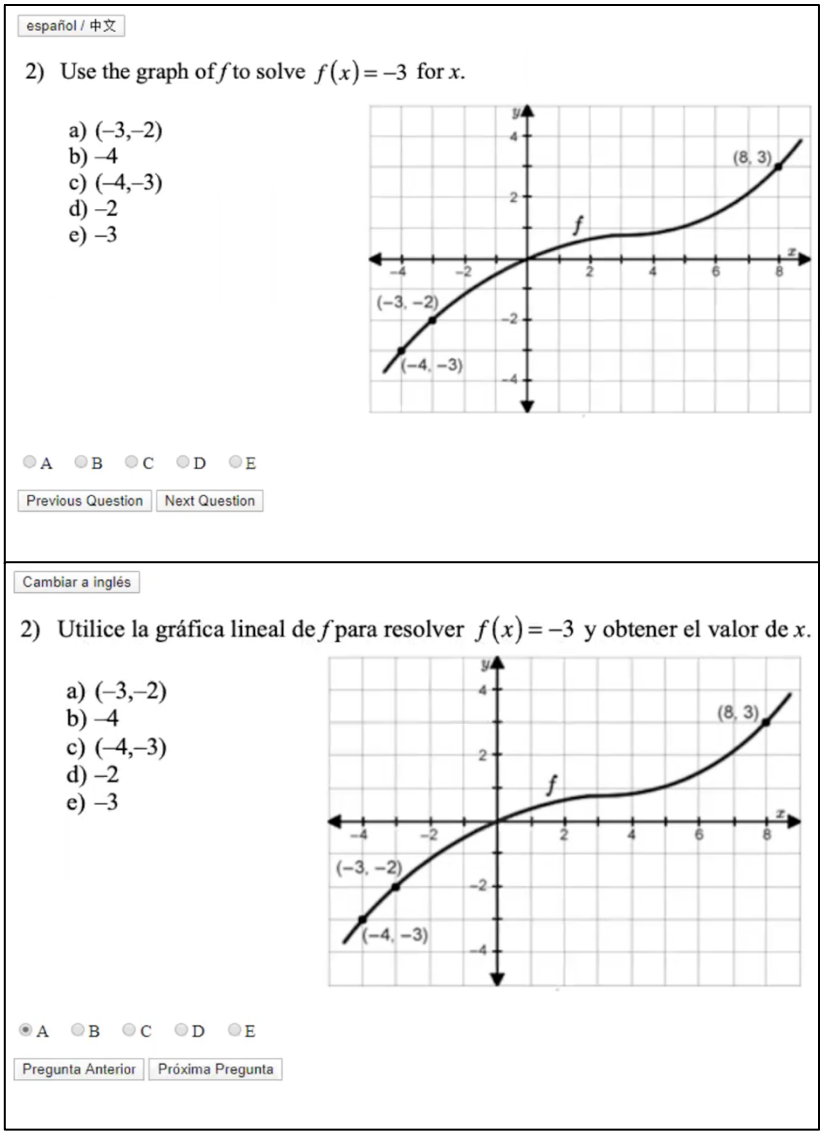
I asked participants not only the same type of cognitive interview questions that I asked in phase one (e.g., when I prompted participants to read each question out loud, think aloud, and write down their work), but I also gave participants the information that this test was newly designed and that I needed their help to discover potential issues. For example, one paragraph from the script read to each participant prior said, “I asked you to participate to help me make sure that the user experience of the test makes sense, to make sure the translation makes sense, and to get any other type of feedback that will help with my study. So, feel free to critique anything at any time. Or to pose questions. Or challenge things” (see Appendix J).

### *Materials*

I developed a language-adaptive version of the PCA, using *Concerto* as the open-source computer platform for test development and delivery (Kosinski & Rust, 2011). Specifically, I developed graphic user interface (GUI) features and source codes to allow test-takers to freely switch back-and-forth between English and their dominant language for each question (see Figure 3).

**Figure 3**

*Two Screenshots of the LAT Showing the Toggle Function*



2) Use the graph of *f* to solve $f(x) = -3$ for *x*.

a) (–3,–2)
b) –4
c) (–4,–3)
d) –2
e) –3

2) Utilice la gráfica lineal de *f* para resolver $f(x) = -3$ y obtener el valor de *x*.

a) (–3,–2)
b) –4
c) (–4,–3)
d) –2
e) –3

*Note.* The English version (above) and the Spanish version (below) of the same item.

Test-takers can toggle back-and-forth between the two versions of the item.

In Figure 3, the top screenshot shows how the LAT first presents the item to the test-taker. The bottom screenshot shows how the LAT presents the item after the test-taker clicks on the toggle button on the top left of the screen. For each item, the test-taker could toggle back and forth between languages indefinitely.

Regarding the LAT item translation, a professional dominant Spanish speaking translator, through a translation company, translated the Spanish items and then a different dominant Spanish speaker back translated the items to ensure no meaning was lost. Additionally, for the Mandarin Chinese items, one dominant Mandarin Chinese speaker translated the items then another dominant Mandarin Chinese speaker back translated the items. The fully translated versions of the test can be found in the appendices (see Appendix K, see Appendix L).

### *Procedure*

Each participant, one at a time, via Zoom, received a scripted introduction to the task from me before clicking on a link to start the newly created version of the LAT exam. As mentioned previously, the first page of the exam asked participants to read and accept consent for the study. The second page presented a practice item, so participants could see a sample item showing how the toggle button could be used. This practice item was meant to call attention to the toggle button on the test showing participants how to use the button to translate an item. The next set of pages displayed the LAT version of the test. After taking the test, participants received a link leading to the same demographics instrument presented in phase one. Throughout the test-taking process, I conducted cognitive interviews with each participant. As mentioned earlier, I used the same interviewing technique described in phase one (See Appendix J). At the end of the interview, I collected information to send the $40 incentive to the participant.

***Analysis***

This phase of the dissertation was more of a production phase than a research phase, however, to prove that the LAT test was ready for larger groups of participants, I conducted a preliminary analysis, using the cognitive interview data (i.e., process data) to establish an argument for validity. My research question for this phase was: Is there preliminary evidence for the validity of the new language-adaptive concept assessment? I used the *Standards'* five sources of validity evidence as my guiding framework to argue for the initial validity of the test (2014):

1. Evidence based on test content (e.g., test blueprint, expert panel review)
2. Evidence based on response process (e.g., cognitive interview)
3. Evidence based on internal structure (e.g., factor analysis)
4. Evidence based on relations to other variables (e.g., convergent/divergent validity, predictive/concurrent validity, change across time, difference between groups)
5. Evidence based on related consequences (e.g., intended consequences, unintended consequences)

Since the LAT is based on content developed previously by Carlson et al. (2010), I was not concerned about presenting new evidence for validity based on test content.

However, as the LAT presents test items in a new way to test-takers, finding preliminary evidence based on response process (i.e., how the test-takers interacted with the test items) is important. As for evidence based on related consequences, I decided to pursue that question in more depth with a greater sample size in the later phase three.

Hence, I gathered validity evidence based on response processes. I coded each cognitive interview for any unique or new type of issue not seen in phase one. I was looking for any differences in the response process. For example, one participant on item 2 stated:

> Participant: See, right here. I was looking for the "next" button and I immediately went over here to the bottom-right and I couldn't find it. Just for. A quick second.

This example revealed a unique type of computer-based issue not found elsewhere. Further differences of functioning could be probed in the following phase with a larger sample size. I also asked probing questions about the translated items as to see if they appeared to operate in much the same way.

Usually during pilot studies, care is taken to a) address the reliability of the measure, b) ensure there a limited number of construct-irrelevant features driving test scores, and c) ensure that items are not functioning differently for various subgroups in the target population. For this phase, I made an initial argument of the validity of the LAT with a particular focus on ensuring that the translated version of the items and the way of presenting the items via an online platform did not make the items function significantly different then the non-LAT version of the PCA. The stronger argument for the validity of the LAT is made in the following phase, phase three, because in phase three I collected the responses from 78 participants who took both the PCA as a LAT and a non-LAT version.

## Phase Three: Study the Efficacy of the LAT

Through this phase I sought to test the efficacy of the newly developed assessment. As such, through this phase I addressed research question four (Is the new language-adaptive concept assessment reliable and valid?), research question five (Do Chinese and Spanish speakers perform better on the new language-adaptive concept assessment compared to the original English version and the non-adaptive translated version?), and research question six (What perceptions do test-takers have about the new language-adaptive concept assessment?).

### *Data Collection and Participants*

For this phase, I recruited 93 students in total in a similar manner as previous phases. I recruited participants primarily from a large university in the U.S. Southwest. through an online announcement then asked to recommend other participants who fit the qualifying criteria (See Appendix M). I required that participants a) identify as dominant English, Spanish, or Mandarin Chinese speakers; b) be a first- or second-year student at a university or college; c) have taken a mathematics class in the last year. I gave participants a short quiz to ensure that they qualified before being invited to the take part in the study.

Of the 93 students who took part, 36 students identified as dominant English speakers (38.7%), 28 identified as dominant Spanish speakers (30.1%), and 29 identified as dominant Mandarin Chinese speakers (31.2%). The numbers are uneven because, though 50 or more students per language were recruited, attrition rates, especially during the transition to online learning during a pandemic meant only about 30 students per language participated in both test taking sessions.

The participants made up a diverse group identifying as Asian/Pacific Islander (*n* = 33; 35.5%), Black or African American (*n* = 3, 3.2%), Hispanic or Latino/a (*n* = 32; 34.4%), and White (*n* = 25, 26.9%). Ages ranged from 18-56 with a mean of 22.3 and a median of 20. For gender, 60.2% of the participants identified as female (*n* = 56), 38.7% identified as male (*n* = 36), and 1.1% identified as non-binary (*n* = 1). Childhood household incomes levels varied widely with 21.5% (*n* = 20) from households earning $0 - $24,999, 34.4% (*n* = 32) from households earning $25,000 - $49,999, 20.4% (*n* = 19) from households earning $50,000 - $74,999, 5.4% (*n* = 5) from households earning $75,000 - $99,999, 7.5% (*n* = 7) from households earning $100,00 - $149,999, and 10.8% (*n* = 10) from households earning $150,000 and up. Participants were not students who previously participated in phases one or two.

### *Materials*

I combined the language-adaptive version of the PCA created in the second phase with attached demographics questions to ask about participants' linguistic background (See Appendix N).  I administered the instrument online on the *Concerto* platform

(Kosinski & Rust, 2011), so all participants had access to the computer-based test. The first online page instructed participants to read and consent to the study (See Appendix O). The *Concerto* platform stored the data in an SQLite database. I also deidentified the data with unique user identification used in lieu of names.

***Procedures***

Participants took two forms of the test, one at each time (i.e., at the beginning and end of the semester). The first form was standard (i.e., not language adaptive test), the other was a language-adaptive version of the test (with a toggle ability to switch between languages). I counterbalanced the study randomly with some students taking the linguistically adaptive version first and others taking the linguistically adaptive version second. The dominant English speakers took two tests of the same form because their language-adaptive version is the same as the control version. Participants were given a $20 incentive to participate in the first test session and $30 to participate in the second test session which took place typically 4-8 weeks after the first session. I used the counter-balanced design, and the 4-8 week delay between tests, to counteract possible testing effects where students could learn from the test or remember the test content (Cook et al., 2002).

I also carried out in-depth follow-up interviews with 19 participants, 8 Spanish speakers and 11 Mandarin Chinese speakers, recruited via follow-up emails (See Appendix P) (Sandelowski, 2008). As in phases one and two, I selected participants who represented different genders and social economic statuses, to learn how they felt about their testing experience, such as the confusion and anxiety they may have experienced, and attitudinal variances towards language-adaptive tests to address research question 6 (See Appendix Q).

***Measures***

**Age.** I measured participants' age via the demographic instrument asking for precise age. Of the 78 participants, ages ranged from 18 to 56 with a median age of 20, a mean age of 21.9 and a standard deviation of 6.1.

**Gender.** I recorded participants' gender via the demographic instrument asking for female, male, or other. Of the 78 participants, 49 identified as female, 28 identified as male, and one identified as other.

**Math Ability Self Report.** One of my measures for math ability asked participants to rate their own math ability via a six point scale from very weak to very strong. Of the 78 participants, four identified their math ability as "very weak", 13 identified their math ability as "weak", 21 participants identified their math ability as "slightly weak", 18 participants identified their math ability as "slightly strong", 18 identified their math ability as "strong", and four identified their math ability as "very strong." On a scale from 0 to 5, the median score was 3 and the mean score was 2.58 with a standard deviation of 1.30.

**Math Ability As Measured By Linguistically Simple Items.** As another method for measuring math ability, I summed the score from each item that I identified qualitatively as linguistically simple using Abedi et al.'s framework in phase one. This sum of scores is meant to indicate each participants' math ability with less noise from linguistically complex items. For this measure, I scored 16 items. Participants scores ranged from one to 16 with a median of 8 and an average score of 8.38 with a standard deviation of 4.05.

**PCA Test Scores.** I scored each item on each version of the test (the LAT and the non-LAT version) as correct or incorrect with no partial credit given. For each test, I summed the total score for all items to calculate a total score. The scores on the non-LAT version of the test ranged from two to 25 points out of a total of 25 points possible. Of the 78 scores on the non-LAT version of the test, the median score was 12 and the mean score was 12.4 with a standard deviation of 6.28 points. The scores on the LAT version ranged from two to 25 points as well. Of the 78 scores on the non-LAT version of the test, the median was also 12 points, and the mean was 12.2 with a standard deviation of 6.06 points.

**Socioeconomic Status (Household Income).** My measurement for socioeconomic status asked participants to select their household income growing up from six different ranges. Eighteen participants selected "$0 to $24,999", 27 participants selected "$25,000 to $49,000", 15 participants selected "$50,000 to $74,999", four participants selected "$75,000 to $99,999", six participants selected "$100,000 to $149,999", and eight participants selected "$150,000 and up". For the sake of analysis, I recoded each bin of incomes on a scale from 0 (i.e., "$0 to $24,999") to 5 (i.e., "$150,000 and up"). Of the 78 participants, the median score was 1, the mean score was 1.70 and the standard deviation was 1.58.

**Other Measures.** Other measures collected on the demographic instrument proved to be less useful for analysis because of missing data or flaws in the instrument. For example, I attempted to collect other outside measures of math and language ability by asking participants to list their scores on language proficiency exams, typically required for international students to enter U.S. universities, like the IELTS or TOEFL test**.** These exams, in particular, measure reading ability at the college level. Additionally, I asked the participants to list their SAT or ACT math scores as an outside measure of math ability. However, due to the unique nature of admissions at the particular university where I conducted my study, many students had neither taken the SAT, ACT, TOEFL, or IELTS exam. Of those who did take the exam, many could not remember and could not access their scores. Hence, the items had rates of missingness above 50% rendering the measures less useful for my analysis.

### *Analysis*

***Analysis for Research Question 4: Is the new language-adaptive concept assessment reliable and valid?*** Firstly, I examined the reliability and validity of the translated version and the language-adaptive version of PCA. As in phase two, I used the five sources of validity evidence from the *Standards* as my guiding framework to evaluate the validity of the test (2014):

1. Evidence based on test content (e.g., test blueprint, expert panel review)
2. Evidence based on response process (e.g., cognitive interview)

3. Evidence based on internal structure (e.g., factor analysis)
4. Evidence based on relations to other variables (e.g., convergent/divergent validity, predictive/concurrent validity, change across time, difference between groups)
5. Evidence based on related consequences (e.g., intended consequences, unintended consequences)

Validity evidence based on test content was previously demonstrated to some extent by Carlson et al. (2010) in their detailed description of test design. The LAT format, however, shares the same underlying content as the PCA, hence shares the same evidence of test content. Therefore, to analyze this aspect of validity evidence, I detailed my own process of test construction and translation, a procedure that Lissitz and Samuelsen (2007) argue is the most important way to determine the validity of a new instrument (a process started in the previous phase).

To evaluate the reliability of the test and validity evidence based on the internal structure, I calculated the alpha coefficient (Cronbach, 1951). Though Carlson et al. (2010) cautioned against using *coefficient alpha* as a measure of reliability of the PCA, Carlson et al. still reported the coefficient alpha to demonstrate a lower bound of reliability (Cronbach, 1975). I did the same for this research study. I also showed *test-retest reliability* among the items in which the participants did not use the features of the LAT test (e.g., did not click on the toggle button). Test-retest reliability is a test of score stability over time. The statistical test for test-retest reliability is an intra class correlation test between the two PCA tests for each participant who took both PCA tests. In this case, using the testRetest function from the R package *psych* (Revelle, 2017).

To evaluate validity evidence in response processes, I created an argument based on validity evidence from the cognitive interviews used in phase one and phase two. During phase two cognitive interviews, in particular, I targeted validity evidence of response processes by asking particular questions about translation or user experience issues while conducting the LAT test.

To evaluate validity evidence based on relations to other variables, I examined the correlation between scores on the PCA and math ability levels. I expected a strong positive relationship between those two variables as evidence for validity. Lastly, for validity evidence based on related consequences, in conducted the in-depth interviews to follow-up with participants about their testing experience. The particular methods of analysis will be described in more depth in the explanation of the analysis for research question six: *What perceptions do test-takers have about the new language-adaptive concept assessment?*

***Analysis for Research Question 5: Do Chinese and Spanish speakers perform better on the new language-adaptive concept assessment compared to the original English version and the non-adaptive translated version?***

To begin exploring the data, I first calculated some descriptive statistics and tested the assumptions required to run inferential statistical tests like an ANOVA. Importantly, I based my research design on the interaction hypothesis, as previously highlighted in the literature review. As a reminder, the interaction hypothesis, typically meant to evaluate new test accommodations, compares how the test accommodation interacts with both the target group (i.e., those who the accommodation was designed for) and the non-target group (National Research Council, 2004; Zuriff, 2000). In theory, the accommodation should help the target group, by reducing CIV, but have no effect on the non-target group. My research design borrowed from Elliott and Marquart (2004) who conducted a repeated measures counterbalanced design in which three groups of students completed two different and equivalent test forms in different testing conditions. I did the same with three groups of students (English, Spanish, and Mandarin Chinese) taking two versions of a test at two different time point. I also counterbalanced the tests through random-assignment. Of note, the English version of the LAT test was the same as the non-LAT test, because both versions of the test included the participants' dominant language. Although the test taken by the English dominant speakers appeared the same between the two different periods of time, one was recorded as a LAT and the other a non-LAT following the random-assignment counterbalanced design, as with the Chinese and Spanish dominant speakers. To explore the interaction hypothesis, I conducted two mixed ANOVAs.

**Mixed ANOVAs.** I conducted two different two-way mixed ANOVAs. The first ANOVA compared dominant Spanish speakers with dominant English speakers. The second compared dominant Mandarin Chinese speakers with dominant English speakers. I set both ANOVAs up with test type as the within-subjects factor (i.e., non-LAT version or LAT version) and dominant language as the between-subjects factor (i.e., English and Spanish or Mandarin Chinese). Recall that the central tenet of the interaction hypothesis is that there should be a significant interaction between the novel test accommodation and the characteristic of interest (i.e., dominant language). Hence, I tested for an interaction between the dominant language and test type. A significant interaction would signal, according to Elliot and Marquart (2004), “evidence of a meaningful change in students’ test scores” (p. 357). If no significant interaction, a main effect signifying a significant difference between test type means could indicate that the novel test type assisted all students regardless of dominant language which would also be a finding of interest. As noted, there may be a significant main effect when comparing mean scores on the between-subject factor (i.e., dominant language). However, finding a significant difference in test scores among different language groups, lies outside of the scope of my research questions which focus on how the test type impacts scores.

To begin conducting the two different two-way mixed ANOVAs, I checked for the robustness of assumption required for such a statistical test. Running such an ANOVA requires testing for testing for outliers in each group (i.e., for each test type and each language), testing for a normal distribution in each group, checking for a homogeneity of variance for between-subjects factor (e.g., language), testing for a homogeneity of covariances of the between-subjects factor, and testing of adjusting for sphericity.

First, I checked for outliers using the *rstatix* package in the R programming language to check for datapoints that lie above the 3rd quartile plus 1.5 times the interquartile range or below the 1st quartile plus 1.5 times the interquartile range (Kassambara, 2013). Secondly, I tested for a normal distribution among test scores, which can often prove problematic when tests are either too easy for the test-takers, creating a ceiling effect, or too hard for test-takers, creating a floor effect. To test for normality, I conducted a Shapiro-Wilks test for each combination of factor levels. If data is normally distributed the *p*-value should be above .05.

Then I checked for the homogeneity of variance for each type of test using the Levene's test with the null hypothesis being that the variances are homogeneous. According to the results, I rejected the null hypothesis indicating that there is a statistically significant difference in the homogeneity of the variances. So, instead of running typical mixed ANOVAs, I ran a more robust two-way between-within subjects ANOVAs on the trimmed means using the R package *WRS2* (Mair & Wilcox, 2020).

I also checked for the homogeneity of covariances of the between-subjects factor (e.g., test type) using Box's M-test (Box, 1949). If the test was not statistically significant with a *p*-value > .001 then I could accept the null hypothesis that the covariances are homogeneous. Lastly, *WRS2* automatically checks sphericity using Mauchly's test. If the program identifies a violation of sphericity, the program applies the Greenhouse-Geisser correction. I probed ANOVA effects further by performing pairwise *t*-tests.

**Linear Mixed-Effect Model Results.** To statistically control for the confounding effects of potential covariates, I conducted a series of linear mixed-effects regression models using test score as the outcome variable, test type as the primary predictor of interest, and math ability, gender, and socioeconomic status as controlled covariates. The models were run in *lme4* R package (Bates et al., 2014). As part of the computation, I used restricted maximum likelihood estimation to estimate covariance parameters. I ran two separate models to evaluate the Spanish speaker data and then the Mandarin Chinese speaker data. I ran these models using only total score for linguistically complex items (as identified in phase one) as the outcome variable.

**Item-Level Analysis.** My next task was to conduct tests on the item-level. Particularly because I flagged some items in phase one as more linguistically complex than other items, I was interested to see whether LAT made a difference in participants' test performance on those linguistically complex items, and how the effects of LAT compare across items with varied linguistical complexity. First, I compared mean scores for each language group on each item on each test type. Then, to examine individual items, I performed multiple paired *t*-tests, using the R base package, comparing the mean differences between non-LAT test scores and LAT test scores. There was no need to adjust for Type I error because even without adjustment no *p*-values indicated statistical significance. The data used for this analysis included the LAT and non-LAT test data from the Mandarin Chinese speakers and Spanish speakers only, excluding the dominant

English speakers, because this analysis focused on the main effect of test type, but the dominant English speakers received two identical tests.

Next, I fit a two-parameter IRT model. To do so, I fit a 2-parameter IRT model, using the *mirt* package in R (Chalmers, 2012), to both the non-LAT and the LAT item responses for Mandarin Chinese and Spanish speakers. To stabilize the small sample size for the model, I used Bayesian priors applied to the model setting the distribution of the *a*-parameter as a log normal distribution with a mean of 0.2 and a standard deviation of 0.2 and the *b*-parameter as a normal prior of 0 with a standard deviation of 1. The parameters were standardized on theta to make the numbers comparable between the models so I could see how the LAT items compared to non-LAT items.

For the model using only Mandarin Chinese speaker data and the non-LAT test results, the model fit strongly. The Root Mean Square Error of Approximation (RMSEA) = 0, with a suggested cutoff value of less than or equal to .06 as a suggested guideline for assessing fit (Hu & Bentler, 1999). Additionally, the Comparative Fit Index (CFI) = 1 which is above the suggested threshold of .9. For the same data, but the LAT test results, the model also had a strong fit with the RMSEA = 0 and the CFI = 1.

For the model using only the Spanish speaker data and the non-LAT test results, the model also fit strongly. The RMSEA = 0, with a suggested cutoff value of less than or equal to .06 as a suggested guideline for assessing fit. Additionally, the CFI = 1 which is above the suggested threshold of .95. For the same data, but with the LAT test results, the

RMSEA = 0.03 which is still below the threshold and CFI = 0.92. Hence, the models fit well enough to examine the resulting coefficients and interpret their meaning.

Lastly, I conducted logistic regressions for each individual item using math ability, gender, and socio-economic status as predictors, to predict whether participants selected the correct answer on each item.  For these models, I conducted a series of repeated-measures generalized mixed-effects logistic regressions for each item using the *lme4* R package in order to see if the test type was a significant predictor in any of the models (Bates et al., 2014). As part of the computation, I bound optimization by quadratic approximation (i.e., BOBYQA) method of optimization with 10 points per axis for evaluating the adaptive Gauss-Hermite approximation to the log-likelihood. I estimated *p*-values via Wald-tests using the Satterthwaite approximations to degrees of freedom.

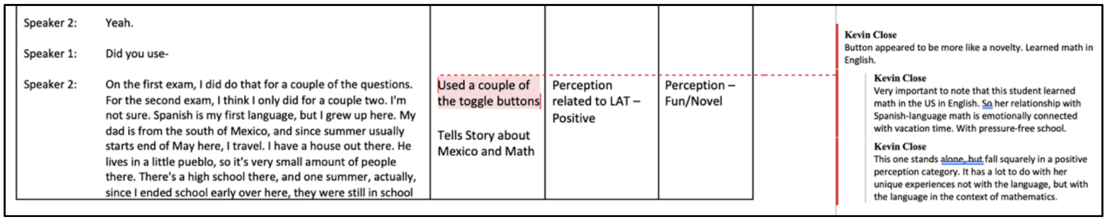
### ***Analysis for Research Question 6: What perceptions do test-takers have about the new language-adaptive concept assessment?***

After analyzing the quantitative test scores, I analyzed the data derived via my in-depth interviews to address research questions four and six. I conducted semi-structured interviews with a focus on the participants' testing experience. I sent the interviews to a company to transcribe verbatim. I coded the material in three rounds starting with a mixture of descriptive and in vivo coding, combined with writing analytical memos, to identify areas of overlap and difference between participants’ experiences with the adapted test in order to generate themes to complexify findings derived from statistically generated findings (Boeije, 2002; Saldaña, 2021; Strauss & Corbin, 1997).

In terms of data analysis, my qualitative coding process can be attributed to information provided in books by Johnny Saldaña (2021) and Miles et al. (2018). In particular, I think of the coding process through the lens of Miles et al. which discuss three flows of activity: data condensation, data display, and drawing and verifying conclusions (Miles et al., 2018); although, I followed more practical coding techniques from Saldaña (Saldaña, 2021). For example, once I collected my data and sent the data to a professional transcription company (rev.com) yielding 19 files with 138 pages of double-spaced text and 42,318 words. On average each participant yielded an average of 7.26 page of text and 2227.26 words. I worked through three cycles of coding making analytical memos throughout the process. I worked in a four column Microsoft Word document with one wide column dedicated to the interview transcription and the other three columns dedicated to the first, second, and third cycles of coding (see Figure 4).

**Figure 4**

*Example of Coding Worksheet*



| | | | | |
|---|---|---|---|---|
| Speaker 2: | Yeah. | | | |
| Speaker 1: | Did you use- | | | |
| Speaker 2: | On the first exam, I did do that for a couple of the questions. For the second exam, I think I only did for a couple two. I'm not sure. Spanish is my first language, but I grew up here. My dad is from the south of Mexico, and since summer usually starts end of May here, I travel. I have a house out there. He lives in a little pueblo, so it's very small amount of people there. There's a high school there, and one summer, actually, since I ended school early over here, they were still in school | Used a couple of the toggle buttons<br>Tells Story about Mexico and Math | Perception related to LAT – Positive | Perception – Fun/Novel |

**Kevin Close**
Button appeared to be more like a novelty. Learned math in English.

**Kevin Close**
Very important to note that this student learned math in the US in English. So her relationship with Spanish-language math is emotionally connected with vacation time. With pressure-free school.

**Kevin Close**
This one stands alone, but fall squarely in a positive perception category. It has a lot to do with her unique experiences not with the language, but with the language in the context of mathematics.

*Note.* From left to right: Transcribed interview, coding round one, coding round two, coding round three, and analytical memos.

I kept memos as comments within the word document. I also replied to my own comments to add additionally layers of memos within the document. I kept all interviews as part of the same Microsoft word table, so they could be more quickly compared or referenced. I kept a copy of research question written on a piece of paper nearby following the advice of Auerbach and Silverstein in order to keep me focused and allay my anxieties when making coding decisions (as cited by Saldaña, 2021, p. 21). Practically speaking, the piece of paper served as a visual reminder of my purpose and guide for my decision making while coding and memoing.

For the first round of coding, I did not approach the data with a particular technique of coding. I coded descriptively most of the time, but sometimes I coded in vivo to capture particularly outstanding phrases. I also took care to note what struck me which Saldaña notes can be a code that is "surprising, unusual, or conceptually interesting" (2021, p. 18). Importantly, the first round was not a rote coding exercise, but rather a form of data analysis and data processing in itself (Evers, 2015). After the first round, I copied and pasted my unique codes into a single document to compare and collapse similar themes/ideas. Much of this work was parsing information salient to my research question: "What perceptions do test-takers have about the new language-adaptive concept assessment?" In the end, I uncovered five major groupings for the first round of coding to help parse the information into research question focused content.

After conducting the second round of coding which primarily parsed data into comments or thoughts related to the LAT, I reread my analytical memos and decided to conduct a third round of coding with a focus on the phrases coded “Perception related to LAT - Positive,” “Perception related to LAT – Negative,” and “The tests are the same.” This third round included more analysis with the focus on findings themes for these perceptions of positivity, negativity, or sameness. The guiding sub-question I used to keep focus was “What links these perceptions?” I came up with four major themes which will be presented in the following results  section.

## Summary

In this chapter, I introduced the overarching questions driving this dissertation, explained how the overarching questions could be broken down and mapped to phases of my research study, argued for the importance of mixed-methods in my approach, then walked through participants, procedure, instruments, and analyses used during each phase of the study. In short, the qualitative analysis procedures complemented quantitative measures by illuminating participants’ own perspectives and experiences rather than testing outcomes. In the next chapter, I will lay out the findings from each phase.

# CHAPTER 5

# FINDINGS

In this chapter, I examine the results of my data analysis processes through phases one, two, and three to answer my main research questions:

Phase One:

1. Which test items, and how many test items, cause confusion or misconceptions for linguistic reasons for Chinese and Spanish dominant speakers when reading the PCA?
2. Is there a pattern of confusion or misconception that matches the types of test items as classified by Jamal Abedi and colleagues' linguistic complexity framework (2002; 2012; 2001)

Phase Two:

3. Is there preliminary evidence for the validity of the new language-adaptive concept assessment as evidenced by participant explanations of their selected answers?

Phase Three:

4. Is the new language-adaptive concept assessment reliable and valid?
5. Do Chinese and Spanish speakers perform better on the new language-adaptive concept assessment compared to the original English version and the non-adaptive translated version?
6. What perceptions do test-takers have about the new language-adaptive concept

assessment?

I have organized this section, accordingly, by phase and research question.

## Pre-Phase Results: Linguistic Analysis and Differential Item Functioning

As stated in the methods section, I conducted a pre-phase to ensure that the PCA demonstrates the potential to include linguistic issues. Hence, I examined both the linguistic content of the test items, but also the quantitative discrepancies among test scores among various subgroups. In the following section, I will also describe the results which defend using the PCA as an instrument in the rest of this research study.

### *Pre-Phase Study One: Linguistic Analysis Methodology Study of the PCA*

Using Jamal Abedi and colleagues' framework of linguistic complexity (2002; 2012; 2001), I analyzed each of the 25 items on the PCA simply indicating whether the question exhibited one of the seven issues in Abedi and colleague's framework. I used the sum to calculate a rudimentary level of complication. Based on Abedi et al.'s framework, which provides a general but inexact measurement of difficulty, items 3, 7, 10, 11, 17 were most likely to be difficult mathematics items for English language learners (See Table 4). I presented a full version of the PCA in the Appendix for reference (see Appendix A). For those items, I found levels of complication of 6, 5, 4, 4, and 5, respectively. Whereas items 2, 5, 6, 9, 12, 13, 14, 16, and 21 were likely to be less difficult, all with levels of complication of 0, indicating that I found none of the seven issues.

**Table 4**

*Framework of Linguistic Complexity Applied to All 25 PCA Test Items*

| Q | Infrequent non-mathematics vocabulary? | Passive Voice of verb phrase | Long nominals | Conditional clauses | Relative clauses | Question phrases | Abstract or impersonal presentations | Level of Complication |
|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | X | | 1 |
| 2 | | | | | | | | 0 |
| 3 | X | X | | X | X | X | X | 6 |
| 4 | | | | | | X | | 1 |
| 5 | | | | | | | | 0 |
| 6 | | | | | | | | 0 |
| 7 | X | X | | | X | X | X | 5 |
| 8 | | | | | | | X | 1 |
| 9 | | | | | | | | 0 |
| 10 | X | X | | | | X | X | 4 |
| 11 | X | X | | | X | | X | 4 |
| 12 | | | | | | | | 0 |
| 13 | | | | | | | | 0 |
| 14 | | | | | | | | 0 |
| 15 | X | X | | | | | X | 3 |
| 16 | | | | | | | | 0 |
| 17 | X | X | | | X | X | X | 5 |
| 18 | X | | | | | X | | 2 |
| 19 | | | | | | X | | 1 |
| 20 | X | | | X | X | | | 3 |
| 21 | | | | | | | | 0 |
| 22 | X | | | | X | X | | 3 |
| 23 | | | | | | X | | 1 |
| 24 | | | | | | X | | 1 |
| 25 | | | | | | X | | 1 |

### ***Pre-Phase Study Two: Differential Item Functioning of the PCA***

I conducted the second pre-phase to determine the likelihood that the PCA would include items that exhibit some form of linguistic bias among students by looking for evidence of linguistic bias in a similar, older version of the PCA, Using the DIF analysis previously described in the method chapter, I detected three items that performed differently for non-dominant English speaker as indicated by a *p*-value of less than .05: items 1, 3, and 11 (see Table 5). As a general trend, 21 of the 26 tested items seem biased against non-dominant English speakers shown by the negative deltaMH numbers. The effect size, in most cases, is negligible to moderate, but the pattern of deltaMH being negative for so many of the items indicates a general bias.

**Table 5**

*Mantel-Haenszel Chi-Square Statistics and Effect Sizes*

| | Stat. | *p*-value | alphaMH | deltaMH | Effect Size |
|---|---|---|---|---|---|
| Q1 | 12.03 | .001 *** | 3.02 | -2.60 | C |
| Q2 | 2.23 | .136 | 1.84 | -1.43 | B |
| Q3 | 11.27 | .001 *** | 3.88 | -3.19 | C |
| Q4 | 0.04 | .836 | 1.11 | -0.25 | A |
| Q5 | 0.03 | .861 | 1.23 | -0.48 | A |
| Q6 | 0.05 | .830 | 1.23 | -0.49 | A |
| Q7 | 1.92 | .166 | 2.04 | -1.68 | C |
| Q8 | 0.59 | .442 | 1.29 | -0.60 | A |
| Q9 | 0.003 | .954 | 0.94 | 0.14 | A |
| Q10 | 0.29 | .593 | 1.25 | -0.53 | A |
| Q11 | 4.03 | .045 * | 1.98 | -1.60 | C |
| Q12 | 0.62 | .432 | 0.77 | 0.61 | A |
| Q14 | 1.72 | .190 | 1.61 | -1.11 | B |
| Q15 | 0.68 | .408 | 1.31 | -0.64 | A |
| Q16 | 0.19 | .660 | 0.80 | 0.53 | A |
| Q18 | 0.12 | .731 | 1.17 | -0.36 | A |
| Q19 | 0.01 | .927 | 0.96 | 0.10 | A |
| Q20 | 0.55 | .459 | 1.29 | -0.60 | A |
| Q23 | 0.02 | .898 | 1.02 | -0.04 | A |
| Q24 | 0.003 | .952 | 1.08 | -0.18 | A |
| Q25 | 2.84 | .092 | 1.72 | -1.27 | B |
| Q26 | 0.89 | .346 | 1.42 | -0.83 | A |
| Q27 | 1.98 | .159 | 1.73 | -1.28 | B |
| Q28 | 1.16 | .282 | 1.54 | -1.01 | B |
| Q29 | 0.44 | .506 | 1.24 | -0.50 | A |
| Q30 | 3.67 | .055 | 1.70 | -1.24 | B |

*Note.* Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1; Effect size codes: 0 'A' 1.0 'B' 1.5 'C' (for absolute values of 'deltaMH')

On the item level, language bias is evident in the PPCA in items 1, 3, 7, and 11 because they graded an effect size of C. Item 7, though not statistically significant, hence shown without an asterisk in Table 5 shows an effect size worthy of concern. The general trend toward a negative deltaMH and the four evidently biased items indicate that this test was an acceptable candidate for this dissertation. Though there is no exact overlap between the PPCA and the PCA items, the style of the test items is similar. Unfortunately, there are no items that are direct analogs in term of both item content (i.e., testing the knowledge of limits) and item style (i.e., using graphs and text in combination).

## Phase One Results

### ***Which test items, and how many test items, cause confusion or misconceptions for linguistic reasons for Chinese and Spanish dominant speakers when reading the PCA?***

According to the cognitive interview results, I identified five test items as problematic and another four test items as rarely problematic. The other 16 test items did not show evidence of being problematic items. I coded those as non-problematic. Operationalizing problematic, rarely problematic, and non-problematic meant, as stated in the methods, analyzing whether there was a demonstrated disconnect between mathematics conceptual ability (i.e., the construct being measured) and performance on the item (i.e., did the participant select the correct answer). To evaluate this, I examined each interaction between a test item and a participant. This meant examining 25 items for each of the 22 participants for a total of 550 total interactions. For each interaction, I looked for evidence of the type of disconnect written about above. I recorded a disconnect if the following situations occurred:

1. Participant showed mathematics ability, but either did not select the correct answer or had to guess;
2. Participant did not understand the item enough to have an opportunity to demonstrate their mathematics ability;
3. Participant selected the correct answer but showed that they did not understand the underlying math.

For example, a student failing to understand the word “ripple” may fail to select the correct answer for a problem about the area of an expanding ripple in the water after a stone is dropped into a lake. As another example, some participants skipped an item after failing to comprehend the question, which meant they did not have the opportunity to demonstrate their mathematics ability.

Participants only demonstrated a disconnect on nine of the 25 items. Among those, the items ranged in terms of causing issues. Item 3, for example, showed a disconnect for six of the 22 participants whereas items 20 and 24 only showed a single instance of disconnect (see Table 6).

**Table 6**

*The Number of Participants Who Showed an Instance of Disconnect on Problematic and Rarely Problematic Items*

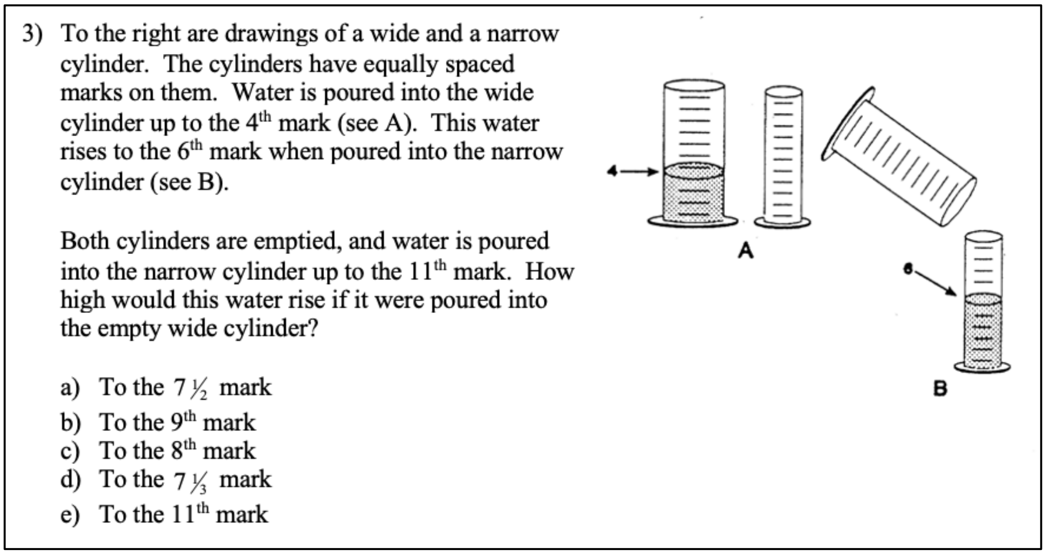
| | Item 3 | Item 4 | Item 8 | Item 17 | Item 18 | Item 7 | Item 10 | Item 20 | Item 24 |
|---|---|---|---|---|---|---|---|---|---|
| Number of Participants | 6 | 3 | 3 | 3 | 4 | 2 | 2 | 1 | 1 |

Items 20 and 24 were clearly rarely problematic with only one instance of each revealing a disconnect. Items 3, 18, 8, and 4 were clearly more problematic compared to the others with several instances of disconnect per item (6, 4, 3, and 3 instances of disconnect, respectively). However, items 7 and 10, with two instances of disconnect, could probably be classified as either problematic or rarely problematic. I chose to classify them as rarely problematic based on the need to separate items into grouping for easier analyses. Based on the categories as explained previously, the problematic test items included items 3, 4, 8, 17, and 18. The rarely problematic test items included items 7, 10, 20, and 24. The problematic and rarely problematic items are included in figures in the appendices (See Appendix A). I was the only one to code the items. So, one limitation for further study would be to include multiple coders and a check for coding agreement.

To add some detail to how I defined instances of disconnect, and to illustrate the ways in which I located test items that demonstrated construct-irrelevant issues, I also include some example instances from my cognitive interviews. Item 3, showing two paragraphs of text and an illustration of the situation, proved to be one of the most problematic items for the participants (See Figure 5).

**Figure 5**

*Item 3 classified as problematic*



3) To the right are drawings of a wide and a narrow cylinder. The cylinders have equally spaced marks on them. Water is poured into the wide cylinder up to the 4th mark (see A). This water rises to the 6th mark when poured into the narrow cylinder (see B).

Both cylinders are emptied, and water is poured into the narrow cylinder up to the 11th mark. How high would this water rise if it were poured into the empty wide cylinder?

a) To the $7\frac{1}{2}$ mark
b) To the 9th mark
c) To the 8th mark
d) To the $7\frac{1}{3}$ mark
e) To the 11th mark

Simply comprehending item 3, proved extremely difficult for multiple students. As an exemplar from the cognitive interviews:

Participant: [Thinks for two minutes after reading the question]

Me: So, what do you think?

Participant: Mmm, I don't know which cylinder is this cylinder

Me: Which one do you think it might be?

Participant: Hmm, it means pour the water again this cylinder? I don't know. But it just two cylinders here.

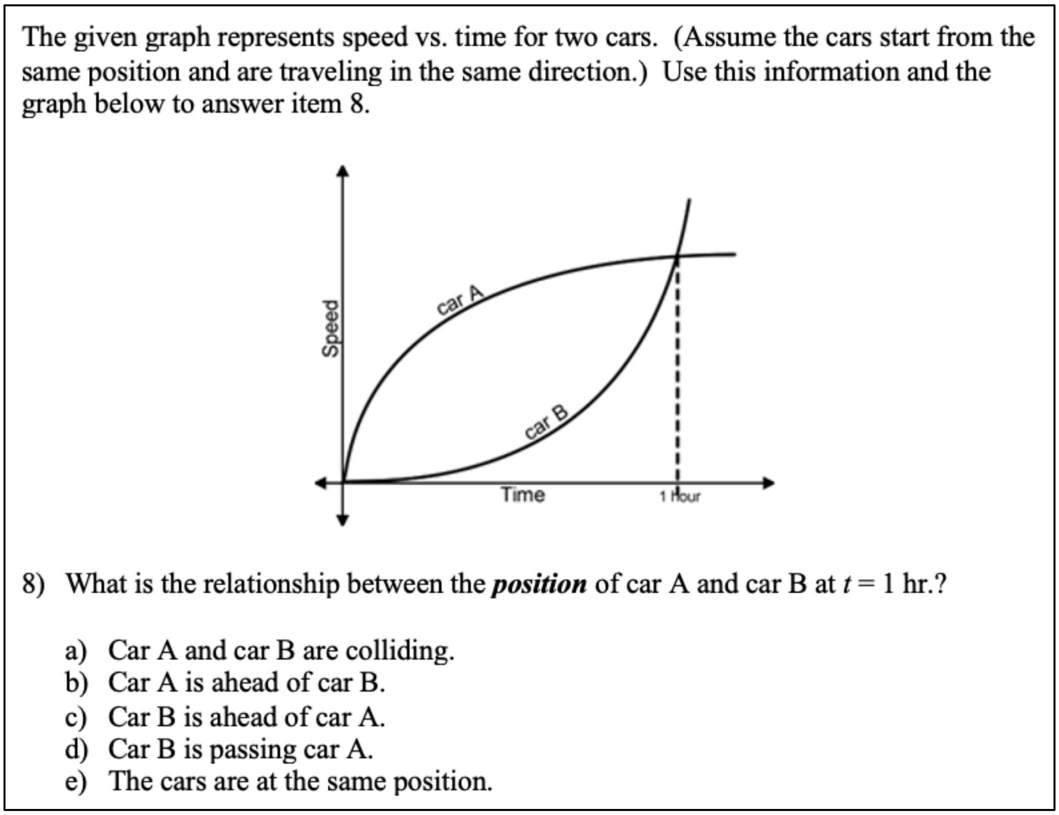
Me: Mmhmm

Participant: [Thinks for about 10 seconds] Can I just go on to the next question?

Other test items were not as difficult to comprehend but included some vocabulary which occasionally presented a hurdle for students trying to select the correct answer. For example, on item 8 one participant correctly understood that car A travelled further than car B (see Figure 6), but this participant did not select the right answer because the vocabulary in the answers caused some confusion.

**Figure 6**

*Item 8 classified as problematic*



The given graph represents speed vs. time for two cars. (Assume the cars start from the same position and are traveling in the same direction.) Use this information and the graph below to answer item 8.

Speed
car A
car B
Time
1 Hour

8) What is the relationship between the ***position*** of car A and car B at $t = 1$ hr.?

a) Car A and car B are colliding.
b) Car A is ahead of car B.
c) Car B is ahead of car A.
d) Car B is passing car A.
e) The cars are at the same position.

Here is an excerpt from the cognitive interview demonstrating the confusion:

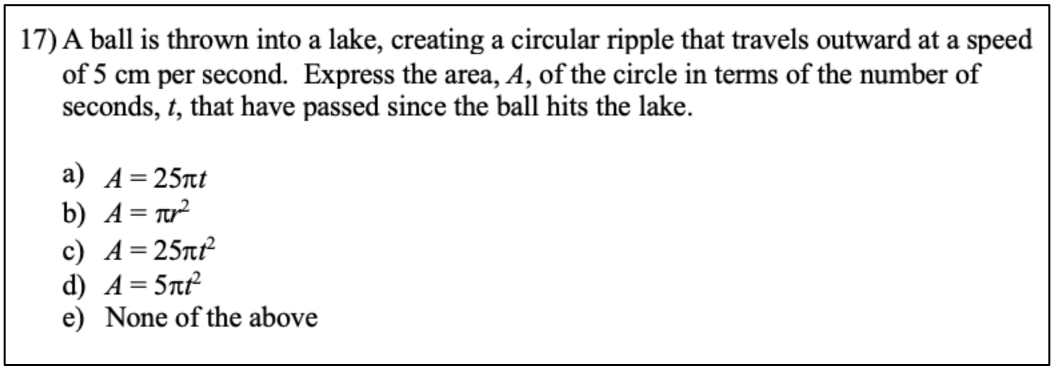
> Participant: I think this should be C. The speed times time equals the distance. From the graph, I can know the distance of A is farther than B because this arrow is bigger than this arrow. So, the distance of A is uh is longer than B. So just if this is a way. Just A is here and B is here then they can uh arrive in the same position in the final because A uh … [selects C]

Despite stating that car A is farther than car B and the distance of car A is longer than car B, this participant selected the answer stating, "Car B is ahead of car A" demonstrating a misunderstanding or misreading of the answers.

Lastly, some of the interactions between participants and items that were *not* coded as disconnects still demonstrated some issues worth discussion. For example, one participant on item 17 showed how not knowing a word like "ripple" in a long word problem about area caused unease, even though the participant selected the right answer in the end (See Figure 7).

**Figure 7**

*Item 17 classified as problematic*



17) A ball is thrown into a lake, creating a circular ripple that travels outward at a speed of 5 cm per second. Express the area, $A$, of the circle in terms of the number of seconds, $t$, that have passed since the ball hits the lake.

a) $A = 25\pi t$
b) $A = \pi r^2$
c) $A = 25\pi t^2$
d) $A = 5\pi t^2$
e) None of the above

The effect of this self-reported unease on test scores is hard to measure exactly. However, the existence of increased unease due to vocabulary is worth noting. As an example, from the cognitive interviews, the following participant selected the correct answer, but qualified the selection by stating, “OK. So maybe this is the wrong, uhh, right answer because I just guess that this word means the perimeter of the circle”:

Participant: I guess this one means uh, this one this one. I don’t know. I guess. I think it’s maybe this one…area…[works out mathematics for 22 seconds]…maybe I got it wrong…I’m not sure

Me: Mmmhmm

Participant: So, I think I need to understand what that means. So, I can just figure out what that means [points at “ripple”]. Because I guess.

Me: OK, if you had to guess what would you put?

Participant: OK, I guess it’s C or this one. It can be this one [selects correct answer].

Me: Why guessing C?

Participant: Because maybe this is this one. I just need to guess. OK. So maybe this is wrong, uhh, right answer because I just guess that this word means the perimeter of the circle.

Though this participant selected the correct answer and understood the math, I coded this as a demonstrated disconnect because this participant was still left unsure of his/her answer. Though the participant ultimately received selected the correct answer, these type instances which increase the unease of test-takers should be considered. I will discuss these types of instances as an area for further study in the discussion section forthcoming.

### ***Is there a pattern of confusion or misconception that matches the types of test items as classified by Jamal Abedi and colleagues’ linguistic complexity framework?***

Test items functioned as expected in most cases; however, I flagged several items for being problematic. Using the Abedi et al. (2002; 2012; 2001) framework to look for patterns related to linguistic complexity, I revealed some patterns. Using the Abedi framework, I correctly predicted nine items that would not be problematic, but only correctly predicted two of the five most problematic test items. Essentially, using Abedi’s framework, I successfully predicted items that would not be problematic, and successfully predicted some, but not all items that would be problematic. With Abedi’s framework as the predictor and interview data as the criterion, the findings show mixed results. On one hand, using Abedi's framework, I predicted non-problematic items with 100% accuracy (e.g., all of the predicted non-problematic items are non-problematic; see Table 7). Using the framework, I tended to have more false positives (e.g., falsely identifying problematic items) than false negatives. Unfortunately, using the framework, I only clearly predicted two of five problematic items.

**Table 7**

*Confusion Matrix Showing False Positive and False Negative Rates With Abedi's Framework Results as the Predictor and Cognitive Interview Results as the Criterion*

| | Predictor: Problematic | Predictor: Rarely Problematic | Predictor: Non-Problematic | |
|---|---|---|---|---|
| Criterion: Problematic | 2 | 3 | 0 | 2/5 (40%)[1] |
| Criterion Rarely Problematic | 2 | 2 | 0 | 2/4 (50%) |
| Criterion: Non-Problematic | 1 | 6 | 9 | 9/16 (56.25%) |
| | 2/5 (40%) | 2/11 (18%) | 9/9 (100%)[2] | 25 Total Items |

*Note.*

[1] The framework only predicted 40% of problematic items

[2] The framework predicted non-problematic items with 100% accuracy

I examined the three problematic items not flagged by Abedi's framework. These unique items demonstrated small specific issues. In one straightforward item, test-takers were asked: “Which one of the following formulas defines the area, $A$, of a square in terms of its perimeter, $p$?” Some participants struggled to interpret “square” within the context of this question. Those participants assumed the “square” in question was asking for an exponent (i.e., “squared”). This simple confusion caused intermediate mathematics students to miss the question.

In contrast, test items that functioned correctly most of the time presented mathematics symbols that dictated how to solve the problem. Items became more likely to be problematic when more context was introduced to make the problem more authentic. Test items that functioned correctly were straightforward, with mathematics symbols directing how to solve the problem, for example, “Given the function $h(x) = 3x-1$ and $g(x)=x^2$, evaluate $g(h(2))$. As another example, the lone item that was flagged by the Abedi framework as problematic, but did not appear to be problematic in cognitive interviews was item 11 (See Appendix A), which could be solved by ignoring the English words completely. The item included straightforward math symbols (i.e., a formula) embedded within a paragraph of text.

Items became more likely to be problematic when more context was introduced to make the problem more authentic. For example, item 18, which was problematic for some participants, started with the phrase “The wildlife game commission poured five cans of fish (each can contained approximately 100 fish) into a farmer’s lake” followed by the formula that included 500 (i.e., the amount of fish) as a constant in the formula. Most participants were able to figure out that the constant in the formula and the five cans of fish in the first sentence were relaying the same information, but three participants skipped answering the item despite having the correct initial idea about how to solve it. This case represents how language ability may not always actually affect comprehension, but rather it may affect confidence and undermine the ability to persist in the face of stressful testing situations.

I expected that test items that ranked as complex on Abedi et al.'s framework (2002) would also present difficulties in real life during cognitive interviews. However, some of the more complex ways that an item might be difficult (such as those using conditional clauses or relative clauses) were not as predictive as simply the number of words in the question (Abedi, 2002). I also expected the number of problematic items to be greater than four based on the pre-phase DIF.

## Phase Two Results

### *Research Question 3: Is there preliminary evidence for the validity of the new language-adaptive concept assessment as evidenced by participant explanations of their selected answers?*

This phase of the dissertation was more of a production phase than a research phase. However, test development was complex and required piloting the test among the target population. As such, research question three (Is there preliminary evidence for the validity of the new language-adaptive concept assessment?) relates directly to gathering initial data about whether the newly developed assessment seems reliable and valid.

I examined validity evidence based on response processes. I coded each cognitive interview for any unique or new type of issue not seen in phase one. I was looking for any differences in the way the item functioned. For example, one participant on item two stated:

> Participant: See, right here. I was looking for the “next” button and I immediately went over here to the bottom-right and I couldn’t find it. Just for. A quick second.

This example revealed a unique type of computer-based issue not found elsewhere. Of the 500 instances probed (i.e., 25 items by 20 participants), only three (0.6%) revealed some sort of issue that stood apart from the typical issues found on items 3, 4, 7, 8, 10, 17, 18. The computer-based test appeared to function as expected in almost all cases. Further differences of functioning could be probed in the following phase with a larger sample size. I also asked probing questions about the translated items as to see if they appeared to operate in much the same way. Not a single participant mentioned a translation issue on any of the items during phase two. Hence, based on evidence from cognitive interviews, I deemed the LAT test as ready for a more robust efficacy study in phase three.

## Phase Three Results

### *Research Question 4: Is the new language-adaptive concept assessment reliable and valid?*

Creating a computer-based version of the test items with a toggled translation would likely affect the validity. Hence, I took steps to demonstrate the reliability and validity of the LAT.

**Reliability.** First, I calculated the test-retest reliability. Test-retest reliability is an index of score stability over time. Using the testRetest function from R package psych (Revelle, 2017), I calculated a test-retest statistic of .91 over a period of time ranging from two weeks to eight weeks depending on the participant. This result indicated "excellent reliability" because it was over .90 (Koo & Li, 2016).

Next, I tested the internal reliability of the test items. I calculated coefficient alpha to evaluate whether the test items seemed to be measuring the same construct (Cronbach, 1975). The Cronbach's alpha measure for the LAT was .88, and the Cronbach's alpha measure for the Non-LAT test was .90. Such results indicate that the lower bound reliability for both versions of the test were high among the population studied, especially for a multidimensional test like the PCA. More importantly, the reliability coefficient for both versions of the test were similar, indicating no major reliability differences between test versions.

**Validity.** To assess the validity of the LAT, I organized my evidence according to the Standards' five sources of validity evidence.

1. Evidence based on test content (e.g., test blueprint, expert panel review)
2. Evidence based on response process (e.g., cognitive interview)
3. Evidence based on internal structure (e.g., factor analysis)
4. Evidence based on relations to other variables (e.g., convergent/divergent validity, predictive/concurrent validity, change across time, difference between groups)
5. Evidence based on related consequences (e.g., intended consequences, unintended consequences)

**Evidence based on test content.** The precalculus content on LAT version of the test is the same on the non-LAT version of the test. Hence, I designed the LAT test to avoid the need to re-argue for validity of the exam based on evidence based on test content. The content is the same as long as the translation was not made in error. Hence, ensuring a valid translation requires a careful process as well as piloting of the items on the target audience. As for the process of translation, a professional dominant Spanish speaking translator translated the Spanish items and then a different dominant Spanish speaker back translated the items to ensure no meaning was lost. Additionally, for the Mandarin Chinese items, one dominant Mandarin Chinese speaker translated the items then a different dominant Mandarin Chinese speaker back translated the items.

For the purpose of testing the validity of the exam, I was more interested in testing how the delivery of the content may have changed the way test-takers interacted with the test. Hence, the focus of my evidence in this section highlights the response process and the results (including evidence of the internal structure and evidence of the validity based on relations to other variables).

**Evidence based on response process.** I collected evidence based on the response process from the phase two cognitive interviews. As stated earlier, of the 500 instances probed (i.e., 25 items by 20 participants), only three (0.6%) revealed some sort of issue that stood apart from the typical issues found on items 3, 4, 7, 8, 10, 17, 18. The computer-based test appeared to function as expected in almost all cases. Additionally, I checked for issues with the translations of the items during the response process. During phase two, I asked 10 dominant Spanish speakers and 10 dominant Mandarin Chinese speakers, “are there any issues with the translation?” for each question that they used the toggle button. There were zero issues according to the participants during this set of interviews.

**Evidence based on internal structure.** Validity evidence based on internal structure is primarily related to dimensionality, measurement invariance, and reliability (Rios & Wells, 2014). Evidence of reliability has already been discussed previously in the under the section header for reliability. As has evidence for support that the test can be used to measure a single factor of interest.

**Evidence based on relations to other variables.** In theory, the overall test scores for the LAT should correlate to the two measures of math ability that I collected for each participant. If these two variables are closely related, that provides evidence that the test is validly measuring the purported construct. First, I listed of test score means and standard deviations by test score ability to show that self-reported math ability aligned with mean test scores for all participants. In general, the test score means aligned with self-reported math ability except for participants who reported very weak or weak math ability with reported means of 8 and 7.54, respectively (See Table 8).

**Table 8**

*Mean Test Scores by Self-Reported Math Ability for All Participants*

| Self-Reported Math Ability | *n* | Mean LAT Score (Standard Deviation in Parentheses) |
|---|---|---|
| Very Weak | 4 | 8.00 (5.60) |
| Weak | 13 | 7.54 (5.36) |
| Slightly Weak | 21 | 10.20 (4.59) |
| Slightly Strong | 18 | 12.70 (4.87) |
| Strong | 18 | 16.30 (5.22) |
| Very Strong | 4 | 20.80 (4.42) |

**Evidence based on related consequences.** Gathering this type of validity evidence in a short-term research setting presents unique difficulties. This particular type of evidence relates to how a test and test scores are used in the wider world. Collecting longitudinal evidence of this type is beyond the scope of this dissertation. However, as a proxy, I did conduct follow-up interviews with some participants to ask about their test experience. For a more in-depth view of the results see research question six results, but initial the in-depth interviews did not indicate any negative consequences on the test-takers. The participants perceived the LAT to be the same as the non-LAT version of the test in many cases, others considered the test to be fun or novel, while others considered the accommodation to be unnecessary. However, no in-depth interview responses raised red-flags regarding the negative or unexpected unintended consequences.

**Reliability and Validity Conclusions.** In general, using multiple pieces of evidence, the LAT appears to be a reliable and valid measure of precalculus conceptual knowledge. Previous research on the PCA speaks to evidence that the content of the exam is valid. Cognitive interviews on pilot exams indicated little to no clear response process issues. Measures of reliability indicated good reliability over time and test scores measures correlate to outside measures as expected for the most part. However, some threats to validity may exist that are unmeasured. Stronger correlation between measures of math ability and overall test scores would increase the strength of the validity argument.

### ***Research Question 5: Do Chinese and Spanish speakers perform better on the new language-adaptive concept assessment compared to the original English version and the non-adaptive translated version?***

**Descriptive Statistics.** To begin exploring the data to answer this research question, I first calculated some descriptive statistics and tested the assumptions required to run inferential statistical tests like ANOVAs. In terms of descriptive statistics, the mean overall scores for each subgroup and each type of test did not differ greatly (see

Table 9).

**Table 9**

*Means and Standard Deviations for Each Type of Test and Dominant Language Group*

| | English (*n* = 31) | Spanish (*n* = 24) | Mandarin Chinese (*n* = 23) |
|---|---|---|---|
| Original | 14.9 (6.86) | 8.92 (4.94) | 12.6 (5.11) |
| LAT | 14.3 (6.73) | 8.54 (4.20) | 13.1 (5.16) |

Participants who identified as dominant English speakers scored on average 14.9 points (*SD* = 6.86) on the original exam and 14.3 (*SD* = 6.73) on the LAT exam. Participants who identified as dominant Spanish speakers scored on average 8.92 points (*SD* =4.94) on the original exam and 8.54 (*SD* = 4.20) on the LAT exam. Participants who identified as dominant Mandarin Chinese speakers scored on average 12.6 points (*SD* = 5.11) on the original exam and 13.1 (*SD* = 5.16) on the LAT exam. A cursory evaluation shows that the differences in mean are well below the standard deviation, indicating a high likelihood that such differences would not be found statistically significant by inferential statistical tests.

Additionally, and of note, dominant Spanish speakers' mean scores on the exams (8.92 and 8.54) did seem to be about one standard deviation lower than mean scores for dominant English speakers or dominant Mandarin Chinese speakers. Though these results do not directly relate to any of the research questions, they do show that the research design included a large variety of mathematics test skills and abilities, as already evidence in the other measures that I collected on math ability levels. I also make a note of these mean differences because a two-way mixed ANOVA, as described in the methods section, may capture a significant mean difference between language groups. Though, again, this finding does not contribute to the main research questions driving this study.

**Mixed ANOVAs.** To ensure the accuracy of these cursory findings, I conducted two different two-way mixed ANOVAs with test type as the within-subjects factor (i.e., non-LAT version or LAT version) and dominant language as the between-subjects factor (i.e., English vs. Spanish in the first ANVOA, and English vs. Mandarin Chinese in the second ANOVA). I tested for either an interaction between the dominant language and test type or a main effect signifying a significant difference between test type means. As noted, there may be a significant main effect when comparing mean scores on the between-subject factor (i.e., dominant language). However, finding a significant difference in test scores among different language groups lies outside of the scope of my research questions which, again, focus on how the test type impacts scores.

**Testing Assumptions.** To begin conducting the two different two-way mixed ANOVAs, I checked for the robustness of the assumptions required for such a statistical test. Running such an ANOVA requires testing for outliers in each group (i.e., for each test type and each language), testing for a normal distribution in each group, checking for homogeneity of variance for between-subjects factor (e.g., language), testing for homogeneity of covariances of the between-subjects factor, and testing of adjusting for sphericity.

First, I checked for outliers using the *rstatix* package in the R programming language to check for datapoints that lie above the 3rd quartile plus 1.5 times the interquartile range or below the 1st quartile plus 1.5 times the interquartile range (Kassambara, 2013). I found no such outliers. Secondly, I tested for a normal distribution among test scores, which can often prove problematic when tests are either too easy for the test-takers, creating a ceiling effect, or too hard for test-takers, creating a floor effect. To test for normality, I conducted a Shapiro-Wilks test for each combination of factor levels. If data are normally distributed, the *p*-value should be above .05. Each combination of the factors showed a normal distribution ($p > .05$) (see Table 10).

**Table 10**

*Shapiro-Wilks Test for Normality Results for Each Combination of Factor Levels*

| Factor Level Combinations | Shapiro-Wilks Test Statistic | *p*-value |
|---|---|---|
| Chinese x Non-LAT | .97 | .778 |
| English x Non-LAT | .95 | .147 |
| Spanish x Non-LAT | .95 | .209 |
| Chinese x LAT | .92 | .076 |
| English x LAT | .96 | .226 |
| Spanish x LAT | .95 | .228 |

Then, I checked for the homogeneity of variance for each type of test using the Levene's test with the null hypothesis being that the variances were homogeneous. According to the results, for the LAT, we can reject the null hypothesis indicating that there is a statistically significant difference in the homogeneity of the variances (see Table 11). So, instead of running typical mixed ANOVAs, I ran more robust ANOVAs such as two-way between-within subjects ANOVAs on the trimmed means using the R package *WRS2* (Mair & Wilcox, 2020).

**Table 11**

*Levene's Test for Homogeneity of Variance Results for Each Test Type*

| Test Type | *df* | *df* | Levene's Test Statistic | *p*-value |
|---|---|---|---|---|
| Non-LAT | 2 | 75 | 2.73 | .072 |
| LAT | 2 | 75 | 3.17 | .048* |

Additionally, I checked for the homogeneity of covariances of the between-subjects factor (e.g., test type) using Box's M-test (Box, 1949). If the test was not statistically significant with a *p*-value > .001 then I could accept the null hypothesis that the covariances are homogeneous. The results showed that the box' M- test statistic equaled 9.08 with a *p*-value of .011 which is greater than .001. So, I failed to reject the null hypothesis and assumed equal covariances.

The last assumption to check was sphericity. However, the R package used to run the ANOVA, *WRS2* (Mair & Wilcox, 2020), automatically applies Mauchly's test for sphericity and automatically applies the Greenhouse-Geisser correction, if found. Hence, I could proceed without testing and reporting that this assumption was met.

**ANOVA Results**. For this mixed ANOVA comparing English dominant speakers and Spanish dominant speakers, the interaction between test type and language was not statistically significant $F(1,33.73) = 0.04$, $p = .845$ (See Table 12), so the impact of language or test type on test scores did not depend on the level of the other factor (i.e. the impact of test type on test scores does not depend on the level of language and vice versa). Since, I hypothesized that the LAT test type would have a positive effect on Spanish dominant speakers without significantly affecting English dominant speakers (i.e., the interaction hypothesis), I hypothesized finding an interaction. This outcome therefore does not support the hypothesis that the LAT test supports Spanish dominant speakers while not affecting English dominant speakers.

With no significant interaction effect, the main effects could be interpreted. For this mixed ANOVA comparing English-dominant speakers and Spanish dominant speakers, the language factor exhibited a significant main effect $F(1,28.06) = 14.10$, $p < .001$ (See Table 12). This supports the finding that participants who identified as dominant Spanish speakers scored significantly lower than dominant English speakers.

**Table 12**

*Two-way Mixed ANOVA Results with Test Type as the Within-Subjects Factor (i.e., Original or LAT Version) and Dominant Language as the Between-Subjects Factor (i.e., English or Spanish)*

| Factors | *df*1 | *df*2 | *F* | *p*-value |
|---|---|---|---|---|
| Language | 1 | 28.06 | 14.10 | < .001* |
| Test Type | 1 | 33.73 | 0.21 | .653 |
| Language * Test Type | 1 | 33.73 | 0.04 | .845 |

To probe this effect further, I performed a pairwise paired t-test for language using the Bonferroni correction to adjust for Type I error. Findings revealed that participants who identified as dominant Spanish speakers scored significantly lower than and dominant English speakers on the non-LAT test ($p < .001$) and the LAT test ($p < .001$). Such findings, as explained previously are not directly relevant to the research questions because they reveal potential differences in mathematics ability by language regardless of the test type, not differences in test delivery type.

I then conducted the same mixed ANOVA comparing English-dominant spearkers against Mandarine Chinese dominant speakers, so my within-subject factor was test type (with LAT and non-LAT as the two levels) and my between-subject factor was dominant language (with Mandarin Chinese speaker and English speaker as the two levels). As stated earlier, according to the interaction hypothesis, I expected to see an interaction between test type and language where language moderated the effect caused by test type. In this model there were no significant interactions or significant main effects (See Table 13). It is important to note that the interaction between test type and language was not statistically significant $F(1,28) = 2.48$, $p = .127$, so the was impact of language or test type on test scores did not depend on the level of the other factor (i.e. the impact of language on test scores does not depend on the level of test types and vice versa). Hence, this outcome does not support the hypothesis that the LAT test supports Mandarin Chinese dominant speakers while not affecting English dominant speakers.

With no significant interaction effect, the main effects could be interpreted. However, test type showed no main effect $F(1,28) = 0.67$, $p < .420$. Hence, indicated that the LAT and non-LAT delivery showed no significant difference among these data either.

**Table 13**

*Two-way Mixed ANOVA Results with Test Type as the Within-Subjects Factor (i.e., Original or LAT Version) and Dominant Language as the Between-Subjects Factor (i.e., English or Mandarin Chinese)*

| | *df*1 | *df*2 | *F* | *p*-value |
|---|---|---|---|---|
| Language | 1 | 25.81 | 0.71 | .407 |
| Test Type | 1 | 28 | 0.67 | .420 |
| Language * Test Type | 1 | 28 | 2.48 | .127 |

However, I conducted further analyses to control for potential covariates, such as economic background, gender, and math ability. The next section of results discusses findings related to the linear mixed-effects regression models, which account for the repeated measures

**Linear Mixed-Effect Model Results.** I conducted two linear mixed-effects regression models, one on only Spanish speaker data and one on only Mandarin Chinese speaker data, to account for covariates such as gender, socioeconomic status, and math ability using the *lme4* R package. For Spanish speakers, math ability score showed a significant relationship with test score, as expected, but no other variable, including test type (e.g., LAT or non-LAT) appeared to have a significant relationship with test score (See Table 14).

**Table 14**

*Results From Linear Mixed-Effects Regression Model Using Only Spanish Speaker Data*

| | Estimate | Std. Error | *df* | *t* statistic | *p*-value |
|---|---|---|---|---|---|
| Intercept | 2.09 | 1.16 | 22.56 | 1.79 | .086 |
| Math Ability | 1.05 | 0.15 | 20 | 7.11 | < .001* |
| Gender | 1.34 | 1.16 | 20 | 1.16 | .261 |
| Socioeconomic status | 0.12 | 0.12 | 20 | 0.29 | .771 |
| Test Type | -0.38 | 0.57 | 23 | -0.66 | .518 |

For Chinese speakers, the math ability score also had a significantly positive relationship with overall test score, *t*(19) = 10.50, *p*-value < .001 (See Table 15), but that was the only significant independent variable in the model. Test type (e.g., LAT or non-LAT) did not have a significant relationship to the dependent variable, overall total score, *p*-value = .443.

**Table 15**

*Results From Linear Mixed-Effects Regression Model Using Only Mandarin Chinese Speaker Data*

| | Estimate | Std. Error | *df* | *t* statistic | *p*-value |
|---|---|---|---|---|---|
| Intercept | -1.37 | 1.60 | 20.44 | -0.85 | .403 |
| Math Ability | 1.51 | 0.14 | 19 | 10.50 | < .001* |
| Gender | 1.61 | 0.82 | 19 | 1.96 | .065 |
| Socioeconomic Status | -0.11 | 0.32 | 19 | -0.35 | .730 |
| Test Type | 0.48 | 0.61 | 22 | 0.78 | .443 |

I also ran regression models using only the scores from the complex items for both Spanish and Mandarin Chinese speakers. In this case, I used the total score from the linguistically complex items and the dependent variable in the regression. The findings were similar, however, for Spanish speakers, math ability was not significantly related to total score *t*(20) = 1.57, *p* = .132, which is a surprising result (See Table 16).

**Table 16**

*Results From Linear Mixed-Effects Regression Model Using Only Spanish Speaker Data and Only the Scores From Complex Items as the Dependent Variable*

| | Estimate | Std. Error | *df* | *t* statistic | *p*-value |
|---|---|---|---|---|---|
| Intercept | 1.48 | 0.83 | 22.68 | 1.79 | .087 |
| Math Ability | 0.16 | 0.10 | 20 | 1.57 | .132 |
| Gender | 1.45 | 0.82 | 20 | 1.77 | .092 |
| Socioeconomic Status | -0.05 | 0.29 | 20 | -0.17 | .864 |
| Test Type | -0.04 | 0.41 | 23 | -0.10 | .921 |

For Mandarin Chinese speakers, however, math ability scores were significantly related to total score on the complex items *t*(19) = 5.70, *p* < .001 (See Table 17). Again, no other predictor in the model proved to be statistically significant.

**Table 17**

*Results From Linear Mixed-Effects Regression Model Using Only Mandarin Chinese Speaker Data and Only the Scores From Complex Items as the Dependent Variable*

| | Estimate | Std. Error | *df* | *t* statistic | *p*-value |
|---|---|---|---|---|---|
| Intercept | -1.59 | 1.13 | 22.56 | -1.40 | .176 |
| Math Ability | 0.59 | 0.10 | 19 | 5.70 | < .001* |
| Gender | 0.93 | 0.58 | 19 | 1.59 | .128 |
| Socioeconomic Status | -0.13 | 0.23 | 19 | -0.60 | .557 |
| Test Type | 0.09 | 0.38 | 22 | 0.231 | .820 |

Even accounting for possible covariates such as math ability, gender, and socioeconomic status, I found no significant effect of test type on overall test scores.

Hence, for my next analysis, I focused on item-by-item analysis instead of overall test score because some items appeared to have more linguistic issues than others.

**Item-Level Analysis.** With no findings revealing differences between the LAT and the original version, the next task was to look at individual items. Of particular interest were differences of performance on the items flagged most problematic in phase one, such as items 3, 4, 8, 17, and 18, or the rarely problematic test items such as items 7, 10, 20, and 24. To establish differences in item performance, I compared the mean scores for each language group on each item on each test type then performed multiple paired-sample *t*-tests to check for a significant difference adjusting the necessary *p*-value using a Bonferroni adjustment to reestablish a more conservative alpha of .002. First, among Mandarin Chinese speakers (*n* = 23), there were no statistically significant mean differences among any of the items (See Table 18).

**Table 18**

*Item-by-Item Comparison of Means by Test Type for Mandarin Chinese Speakers*

(*n* = 23)

| Item | Non-LAT Score Mean | LAT Score Mean | *t* | *p*-value |
|---|---|---|---|---|
| Q1 | .83 | .78 | 0.57 | .575 |
| Q2 | .72 | .74 | 0.37 | .714 |
| Q3 | .52 | .57 | -0.37 | .714 |
| Q4 | .39 | .52 | -1.00 | .328 |
| Q5 | .61 | .74 | -1.82 | .083 |
| Q6 | .78 | .74 | 0.57 | .575 |
| Q7 | .35 | .30 | 0.44 | .665 |
| Q8 | .39 | .35 | 1.00 | .328 |
| Q9 | .83 | .78 | 0.44 | .665 |
| Q10 | .22 | .09 | 1.14 | .266 |
| Q11 | .48 | .65 | -1.70 | .104 |
| Q12 | .83 | .78 | 1.00 | .328 |
| Q13 | .09 | .22 | -1.82 | .083 |
| Q14 | .13 | .17 | -0.44 | .665 |
| Q15 | .52 | .48 | 0.37 | .714 |
| Q16 | .96 | .96 | 0.00 | 1.000 |
| Q17 | .52 | .61 | -0.81 | .426 |
| Q18 | .35 | .48 | -0.90 | .377 |
| Q19 | .48 | .61 | -1.00 | .328 |
| Q20 | .52 | .48 | 0.33 | .747 |
| Q21 | .65 | .65 | 0.00 | 1.000 |
| Q22 | .39 | .39 | 0.00 | 1.000 |
| Q23 | .13 | .13 | 0.00 | 1.000 |
| Q24 | .57 | .52 | 0.37 | .714 |
| Q25 | .48 | .52 | -0.37 | .714 |

*Note.* Darkly shaded items were flagged as problematic items, medium shaded items were flagged as rarely problematic, unshaded items were not flagged

Hence, I failed to reject the hull hypothesis stating that the means between the non-LAT and the LAT were statistically the same. So, for Chinese speakers, the any difference in means could be attributed to random noise.

Additionally, I ran pairwise *t*-tests for each item for Spanish speakers with a similar result. For Spanish speakers (*n* = 24), no single test item demonstrated a statistically significant difference in means between the non-LAT test and the LAT test, when adjusting the necessary *p*-value using a Bonferroni adjustment to reestablish a more conservative alpha of .002 (See Table 19).

**Table 19**

*Item-by-Item Comparison of Means by Test Type for Spanish speakers* (*n* = 24)

| Item | Non-LAT Score Mean | LAT Score Mean | *t* | *p*-value |
|---|---|---|---|---|
| Q1 | .71 | .63 | 0.81 | .426 |
| Q2 | .38 | .33 | 0.57 | .575 |
| Q3 | .33 | .33 | 0.00 | 1.000 |
| Q4 | .42 | .42 | 0.00 | 1.000 |
| Q5 | .42 | .42 | 0.00 | 1.000 |
| Q6 | .58 | .38 | 2.01 | .057 |
| Q7 | .29 | .21 | 0.70 | .491 |
| Q8 | .08 | .04 | 1.00 | .328 |
| Q9 | .29 | .29 | 0.00 | 1.000 |
| Q10 | .17 | .29 | -1.14 | .266 |
| Q11 | .38 | .29 | 0.81 | .426 |
| Q12 | .54 | .46 | 1.00 | .328 |
| Q13 | .25 | .25 | 0.00 | 1.000 |
| Q14 | .00 | .13 | -1.81 | .083 |
| Q15 | .25 | .25 | 0.00 | 1.000 |
| Q16 | .75 | .83 | -1.00 | .328 |
| Q17 | .21 | .17 | 0.37 | .714 |
| Q18 | .38 | .54 | -1.45 | .162 |
| Q19 | .46 | .42 | 0.37 | .714 |
| Q20 | .38 | .42 | -0.30 | .770 |
| Q21 | .50 | .50 | 0.00 | 1.000 |
| Q22 | .33 | .17 | 1.70 | .104 |
| Q23 | .04 | .04 | 0.00 | 1.000 |
| Q24 | .50 | .29 | 2.46 | .022 |
| Q25 | .29 | .46 | -1.16 | .257 |

*Note.* Darkly shaded items were flagged as problematic items, medium shaded items were flagged as rarely problematic, unshaded items were not flagged

Statistical tests showed no significant score differences at the item-level for either Mandarin Chinese or Spanish speaking students when taking the non-LAT test or taking the LAT test. Even among the items that I previously flagged as problematic or rarely problematic, pairwise *t*-test results showed no significant difference in performance among the test-takers. Hence, statistically speaking, any difference in performance for both Mandarin Chinese and Spanish speaking test-takers can be attributed to random noise as opposed to differences in test delivery methods.

To further explore the data, I fit a 2-parameter IRT model to the responses of Spanish speakers and Mandarin Chinese speakers, respectively. In other words, I removed the dominant English speakers from the response data, then fit the IRT model to either the Spanish speakers or the Mandarin Chinese speakers to examine item-level differences for my populations of interest. For each group of participants, I calibrated the item parameters separately for the LAT test data and the non-LAT test data. The calibration was standardized based on the examinees' ability parameter ($\theta$) distribution. And because both calibrations used the identical examinee sample, the item parameters are considered directly comparable (i.e., placed on the same scale). Hypothetically, the difficulty parameter of test items would decrease for the LAT items compared to the Non-LAT items, if the LAT version reduced construct-irrelevant features.

For Spanish speakers on the non-LAT test, slopes ranged from 0.097 on item 22 to 1.278 on item 5, indicating that item 22 was the least discriminating item and item 5 was the most (see Table 20). The difficulty parameter ranged from easy items like item 16 which had a difficulty parameter of 1.388 to difficult items like item 23 which had a difficulty parameter of -1.977.

**Table 20**

*IRT Estimates for Spanish Speakers on the Non-LAT and LAT Versions of the Test*

| | Non-LAT Test | | LAT Test | |
|---|---|---|---|---|
| Item | Slope | Difficulty Parameter | Slope | Difficulty Parameter |
| 1 | 1.09 | 1.14 | 1.14 | 0.86 |
| 2 | 1.25 | -0.20 | 1.24 | -0.29 |
| 3 | 1.19 | -0.39 | 1.12 | -0.31 |
| 4 | 1.19 | -0.03 | 1.19 | 0.05 |
| 5 | 1.28 | -0.01 | 1.22 | 0.06 |
| 6 | 1.22 | 0.67 | 1.17 | -0.13 |
| 7 | 1.02 | -0.58 | 1.18 | -0.87 |
| 8 | 1.18 | -1.69 | 1.20 | -1.85 |
| 9 | 1.18 | -0.57 | 1.11 | -0.49 |
| 10 | 1.16 | -1.19 | 1.14 | -0.48 |
| 11 | 1.12 | -0.21 | 1.19 | -0.48 |
| 12 | 1.21 | 0.50 | 1.14 | 0.20 |
| 13 | 1.18 | -0.77 | 1.15 | -0.67 |
| 14 | NA | NA | 1.16 | -1.31 |
| 15 | 1.20 | -0.77 | 1.17 | -0.67 |
| 16 | 1.22 | 1.39 | 1.22 | 1.81 |
| 17 | 1.21 | -0.98 | 1.18 | -1.08 |
| 18 | 1.14 | -0.21 | 1.13 | 0.52 |
| 19 | 1.19 | 0.15 | 1.15 | 0.04 |
| 20 | 1.18 | -0.21 | 1.22 | 0.06 |
| 21 | 1.10 | 0.29 | 1.00 | 0.32 |
| 22 | 0.10 | -0.40 | 1.16 | -1.08 |
| 23 | 1.18 | -1.98 | 1.15 | -1.85 |
| 24 | 1.11 | 0.30 | 1.15 | -0.48 |
| 25 | 1.19 | -0.57 | 1.21 | 0.22 |

The intervention test, in the righthand column of Table 20, showed similar patterns among items. The three most discriminating items were items 2, 5, 16, and 20. For the non-LAT test, the four most discriminating items were items 2, 5, 6, 12, and 16. Among difficulty parameters, some items seemed much easier on the LAT test, including items 10, 16, 18, and 25. Other items appeared to increase in difficulty such as items 6, 22, and 24.

For Mandarin Chinese speakers on the non-LAT test, slopes ranged from 1.015 on item 19 to 1.323 on item 5, indicating that item 19 was the least discriminating item and item 5 was the most (see Table 21). The difficulty parameter ranged from easy items like item 16 which had a difficulty parameter of 2.176 to difficult items like item 13 which had a difficulty parameter of -1.917.

**Table 21**

*IRT Estimates for Mandarin Chinese Speakers on the Non-LAT and LAT Versions of the Test*

| | Non-LAT Test | | LAT Test | |
|---|---|---|---|---|
| Item | Slope | Difficulty Parameter | Slope | Difficulty Parameter |
| 1 | 1.10 | 1.40 | 1.14 | 1.15 |
| 2 | 1.25 | 0.44 | 1.12 | 0.19 |
| 3 | 1.17 | 0.07 | 1.23 | 0.18 |
| 4 | 1.21 | -0.46 | 1.16 | 0.01 |
| 5 | 1.32 | 0.44 | 1.34 | 0.97 |
| 6 | 1.26 | 1.23 | 1.25 | 0.96 |
| 7 | 1.15 | -0.64 | 1.16 | -0.89 |
| 8 | 1.25 | -0.47 | 1.26 | -0.73 |
| 9 | 1.06 | 1.39 | 0.97 | 1.12 |
| 10 | 1.18 | -1.23 | 1.02 | -1.90 |
| 11 | 1.15 | -0.10 | 1.17 | 0.55 |
| 12 | 1.28 | 1.45 | 1.24 | 1.17 |
| 13 | 1.14 | -1.92 | 1.11 | -1.26 |
| 14 | 1.16 | -1.67 | 1.16 | -1.49 |
| 15 | 1.10 | 0.07 | 1.17 | -0.17 |
| 16 | 1.14 | 2.18 | 1.13 | 2.15 |
| 17 | 1.21 | 0.07 | 1.21 | 0.37 |
| 18 | 1.16 | -0.64 | 1.19 | -0.17 |
| 19 | 1.02 | -0.10 | 1.10 | 0.37 |
| 20 | 1.21 | 0.07 | 1.22 | -0.18 |
| 21 | 1.14 | 0.61 | 1.25 | 0.56 |
| 22 | 1.15 | -0.46 | 1.05 | -0.50 |
| 23 | 1.18 | -1.68 | 1.16 | -1.71 |
| 24 | 1.21 | 0.25 | 1.30 | -0.01 |
| 25 | 1.08 | -0.10 | 1.13 | 0.01 |

For the intervention test, in the righthand column of Table 21, the slopes ranged from 0.970 on item 9 too 1.344 on item 5 indicating that item 9 was the least discriminating item on the LAT test among Mandarin Chinese speakers and item 5 was the most discriminating. The difficulty parameter on the intervention test, ranged from easy items like item 16 which had a difficulty parameter of 2.148 to difficult items like item 10 which had a difficulty parameter of -1.900. Among difficulty parameters, some items seemed much easier on the LAT test, including items 4, 5, 11,13, 18, and 19. Only one item increased in difficulty greatly, item 10.

**Logistic Regression Models.** Additionally, I examined the items considering other variables that I collected about the test-takers background. To control for other potential confounding variables such as math ability, gender, and socioeconomic status, I conducted a series of repeated-measures generalized mixed-effects logistic regressions for each item using the *lme4* R package in order to see if the test type was a significant variable in any of the models (Bates et al., 2014). After conducting 23 logistic regressions, because the models for items 8 and items 24 did not converge, no model indicated that test type significantly affected scores. Of particular interest, were the test items that appeared to cause the most linguistic issues in phase one, items 3, 4, 8, 17, and 18.

For items 3, 4, 17, and 18, test type (e.g., LAT or non-LAT) was not a significant predictor with a p-values of .797, .493, .781, and .120, respectively. The model for item 8 did not converge. Hence, providing further evidence that the LAT test type does not affect scoring on the item level even when controlling for some demographic variables among the test-takers.

### ***Research Question 6: What perceptions do test-takers have about the new language-adaptive concept assessment?***

A vital argument for validity includes taking measures of the consequential validity of a test. Consequential validity is a term that captures the social consequences or unanticipated side-effects of a test. Messick (1989a), who coined the term, linked potential consequential validity issues explicitly with construct-irrelevant difficulty, “I wanted to draw attention to unanticipated side-effects of legitimate test use, especially if unanticipated adverse effects are traceable to sources of test invalidity such as construct underrepresentation and construct irrelevant difficulty” (p. 40). Hence, evaluating the perceptions of test-takers is an important research question to evaluate.

After the first round of coding, described in the methods section, I uncovered five major groupings for the first round of coding to help parse the information into research question focused content. 1. Perception related to the context of the study, 2. Perception related tests in general, 3. Perception related to LAT – Positive, 4. Perception related to LAT – Negative, 5. The tests are the same. For example, the perception related to the test in general included sub-codes like: a) like not being timed, b) straight-forward, c) word problems are stressful, d) not hard.

After conducting the second and third round of coding with a focus on the phrases coded “Perception related to LAT - Positive,” “Perception related to LAT – Negative,” and “The tests are the same.” I came up with four major themes: “The tests are the same,” “This accommodation is not for me,” “This accommodation is fun/novel,” “The accommodation reduces stress.”

**The Tests Are the Same.** In general, findings showed that typically participants felt like the two versions of the test were the same. For example, I applied the “same” code 14 times during the interviews. Typical examples included the following exchange:

Participant: I think they're the same.

Me: Do you notice any differences?

Participant: No, I couldn't remember exactly what difference was.

In other words, many participants did not explicitly notice or discuss the most prominent feature of the LAT test despite a practice test question at the beginning of the test directing students to try the toggle button.

One student mentioned that the tests were the same, but also shared that the student noticed differences. This type of behavior was attributable to thinking of the tests as essentially the same because the toggle button did not apply to him. These types of responses were coded into both the “the tests are the same” and the “the accommodation is not for me” categories. Here’s a good example:

Me: That makes sense. And you didn't notice any differences between the tests, right?

Participant: Yeah. I think they're basically 99% they are the same, right? Yeah.

Me: Did you notice or did you ever use the button that you could press to translate the test?

Participant: No. I don't need to use translate. I'm a senior student, so I don't need to use to translate things. Yeah.

Me: Great. So you saw the button you just didn't use it, right?

Participant: Yeah. I don't need to use it. Yeah.

**The Accommodation is Not for Me.** Other participants felt that the toggle button was not necessary for them. Reasons for their responses often reflected the language background of the participant. Participants seemed happy to work through the items in English if they could comprehend what the items asked. Essentially, they needed to reach a baseline of understanding, enough to work through the item. For example, one participant shrugged off the language component of mathematics items:

Me: Yeah. Basically, a lot of the questions were exactly the same, exactly, exactly the same. And then basically, one version of the test had a little button you could click to switch and translate questions.

Participant: Yeah, I do remember seeing that button.

Me: But you didn't use it.

Participant: I was like, "It's math, It's math. It doesn't matter if it's an English or in Spanish."

Other participants set the baseline more explicitly. In one case the participant simply stated that the participant did not need to use the toggle button "because I think I can do all the question in English.” Hence, this participant felt no need to take the extra step of translating the item. For example:

Me: Did you feel like one was easier than the other?

Participant: No, they were the same.

Me: Yeah. One of the tests had a button that you could click to translate the questions into Chinese, into Mandarin. Did you-

Participant: I didn't-

Me: ... notice? You didn't notice?

Participant: I didn't use it?

Me: Oh, you didn't use it.

Participant: No.

Me: Why not?

Participant: Because I think I can do all the question in English.

Me: Awesome. Did you see the button?

Participant: I saw it.

**The Accommodation Reduces Stress.** Only three participants noted that the accommodation provided some sort of stress relief either by reducing time, providing confirmation of understanding, or reducing difficulty. For example:

Participant: Like I said before, sometimes you're like, "What does that mean or am I reading it right? Am I translating in my head right?" And then you see the picture and you put one and one together, but I'm always not going to lie. Every time I see a word problem, I get stressed, still.

Me: Do you remember... Of the two tests, do you remember liking one better than the other?

Participant: I remember I liked the first one more because I think maybe I did not find the button on the second one, but on the first one there was where you can click it and you can actually read it in Spanish, right?

Me: Yep.

Participant: And that was so nice. And I think when I did that test, I remembered how nicely I can read in Spanish and actually think in Spanish. In math, it's much easier. Yeah. And actually what I've noticed in my three years of college and I still do, is that when I count, I only count in Spanish, even if I'm thinking in English.

This participant noted how word problems were stressful, but it was “so nice” to read in Spanish and think in Spanish.

Another participant mentioned switching language when feeling unsure about the meaning or answer of a test item, but also mentioned using the toggle button to relieve time constraints. This participant felt like the slowest participant at the beginning of the exam so thought, “I should, yeah, speed up. So Mandarin, I will understand faster.” See the example of this participant’s expressed thoughts here:

Me: Great. All right, perfect. So, one test had a button that you could translate it to Mandarin, and the other test did not. Did you notice that, or realize that?

Participant: I think because the second test, I didn't check the time, so [inaudible 00:05:09]. When I looked again, and there was only one left. So, I'd think, "I should, yeah, speed up." So Mandarin, I will understand faster.

Me: Oh, so did you change it to Mandarin?

Participant: Yes.

Me: How often did you do that?

Participant: When I'm not sure about the question, I will switch to Mandarin.

This example reveals how the accommodation helped this participant feel surer and more comfortable when speeded.

**This Accommodation is Fun/Novel**. Lastly, only one participant answer seemed to fit a final category which related to enjoying the fun and novel aspect of the accommodation. This participant connected the Spanish items with the participants' joyful summertime experience attending school in Mexico as a visitor while on summer vacation from her school in the United States, noting:

Me: Did you notice that, on one of those exams that I gave you, that there was a button to translate questions into Spanish? Did you notice that?

Participant: Yeah.

Me: Did you use-

Participant: On the first exam, I did do that for a couple of the questions. For the second exam, I think I only did for a couple two. I'm not sure. Spanish is my first language, but I grew up here. My dad is from the south of Mexico, and since summer usually starts end of May here, I travel. I have a house out there. He lives in a little pueblo, so it's very small amount of people there. There's a high school there, and one summer, actually, since I ended school early over here, they were still in school over there till July. So one summer, I was talking to my family and things. They were like, "Oh, you should go shadow at the school so you can get what school's like in Mexico."

This example reveals the complex connection between language and background which cannot be captured by labels like dominant Spanish speaker. In this case, she spoke Spanish at home with family, but English in school, except when going to shadow at the local school during her summers in Mexico.

**Summary**

Findings from all three phases of my study revealed a) that several items on the X test demonstrated linguistic issues; b) the items without issues could be predicted using Abedi et al.'s framework (2002, 2013, 2001), but the problematic items likely need cognitive interviews with multilingual student to discover; c) the LAT can be argued to be reliable and valid, to an acceptable degree; d) the LAT does not appear to have a significant effect on the overall scores or the item scores for multilingual participants; e) most participants felt that the non-LAT version of the test and the LAT version of the test were the same or similar; and f) many participants did not use the toggle button because they understood the mathematics items well enough. In the next section, I discuss some of the implications of the findings and connect the findings to outside literature.

# CHAPTER 6

## CONCLUSIONS, DISCUSSION & FUTURE CONSIDERATIONS

In this study I present the findings from concept to execution for a new type of test delivery informed by applying translanguaging theory to modern computer-based testing platforms. The holistic nature of this study, drawing from multiple methods and multiple stages in the test development process, means that this discussion section has much to offer. The findings of this study speak to the effectiveness of Jamal Abedi et al.'s (2002, 2013, 2001) framework for linguistic complexity. The findings also speak to the magnitude and patterns of linguistic issues found in well-regarded, well-researched tests used nationwide. Additionally, the findings reveal lessons learned and best practices for test development. They also reveal ways in which a new test delivery approach went well, and ways it did not. These speak to findings from the practical aspect of test design.

Most importantly, however, this study speaks to what happened when I applied theories which add complexity and nuance to language, like translanguaging, to standardized formats which often try to reduce complexity for decision makers. The result, predictably, sometimes makes inroads and other times fails to effectively enact theory into usable practice. That messy result is fine. It is expected. It is also part of the process to rethink and reimagine standardized testing practices for culturally diverse learners. Hence, in this section I present discussion about both the implications of the detailed stages of the research project, but also the overall implications of an undertaking like this research project. For the sake of organization, I will discuss the findings for each phase then discuss the overall implication of this study on the field.

### Discussion of Phase One

In phase one, I evaluated the content of the test items and how study participants interacted with the test items. My focus was to discover whether the test items on the PCA revealed linguistic issues and, if so, the magnitude of those issues. Building on earlier literature, I predicted that even a well-regarded test like the PCA would include items with linguistic issues for certain students (Luykx et al., 2007; Winter et al., 2006). To measure the magnitude, I used two tools. For one, I used Jamal Abedi and colleague's (2002, 2013, 2001) framework for linguistic issues in testing to evaluate the linguistic content of the test items. For another, I used cognitive interviews with linguistically diverse students to evaluate how students with diverse linguistic backgrounds interact with test items building off the work of others in the field (Benítez & Padilla, 2014). I also combined results from the linguistic analysis using Abedi et al.'s framework and the cognitive interviews to compare where overlap existed providing some novel empirical research regarding the framework. In other words, I discovered some cases where the framework succeeded to predict linguistic issues, and also where it failed to predict linguistic issues among self-identified dominant Spanish and dominant Mandarin Chinese speakers.

Two practical implications follow from the results. For one, using Abedi et al.'s (2002, 2013, 2001) framework for linguistic complexity could be an excellent tool for test item developers because Abedi’s framework represents a clear list of articulated issues. While there are already many methods to detect or adjust bias at the disposal of test developers some are controversial (Freedle, 2003), others are expensive (Thurlow et al., 2006), and yet others require post-hoc adjustments (Holland & Wainer, 2012). This type of framework, which can be applied during test development with little cost in time or money, could be of immediate use to test developers or teachers to create fairer tests. More importantly, following such a framework involves thinking about linguistic bias concurrently while developing items, which is something researchers have posited could overcome some of the inherent issues with the test development process (Solano-Flores et al., 2002a). However, the framework clearly misses some issues; while helpful, the framework is not holistic. While it cannot replace current processes for establishing validity (e.g., testing out items in experimental sections of the GRE and running DIF), it can provide a simple framework to help test developers think about and avoid some linguistic issues concurrently with item development.

In fact, Abedi’s framework could and should be used in conjunction with other frameworks and readability tools for identifying linguistic or item design issues. These could include such frameworks and tool such as the universal design framework, a framework to design items that emphases universal accessibility (Johnstone et al., 2008), and tools like Coh-Metrix measures for readability (Graesser et al., 2004, 2011). Testing the use of all these frameworks in conjunction during the test development process could be a direction for further study.

For another practical implication, the magnitude of linguistic issues in such a well-regarded test is of concern. This is yet more evidence that current test development processes are failing to account for the CIV that disadvantages students who do not speak the test language natively (Abedi & Linquanti, 2012). Using cognitive interviews, I coded five of the 25 items as problematic. For the code “problematic,” I operationalized it as meaning that some students were not able to show their mathematics ability because of misunderstanding or misinterpreting the item. This evidence demonstrates that a quote from AERA in 1985, “Any test is to some degree a test of English language proficiency and may be [an] invalid [and] unreliable measure of English language learners' academic proficiencies," continues to hold true (AERA, 1985, as cited in Solano-Flores et al., 2001, p. 50). If five of 25 mathematics items cause misunderstanding for some non-dominant English speakers, then 20% of the test items are fundamentally unfair. This magnitude, on such a well-researched and widely distributed test, not only supports research saying that linguistic issues are prevalent in testing (Abedi, 2002), but also supports arguments on a more systematic scale saying that standardized assessments are inherently and structurally blind to the needs of minorities (Solano-Flores et al., 2002b; Solano-Flores & Trumbull, 2003). As stated earlier in the literature review when discussing structural issues with standardized tests, researchers like Wayne Au question whose standards test developers base their tests (2010).

Regardless, finding 20% of test items fundamentally unfair falls under a fairness issue as defined by the *Standards.* As stated in chapter one, under the definition of fairness according to the *Standards*, fairness can be thought of as a measurement principle (i.e., test should not demonstrate bias) and as an accessibility principle (i.e., test-takers should be given an opportunity to demonstrate their level of the construct) (2014). Items like the five shown to be problematic, in essence, do not provide fair access for test-takers hence leading to a higher likelihood to demonstrate bias.

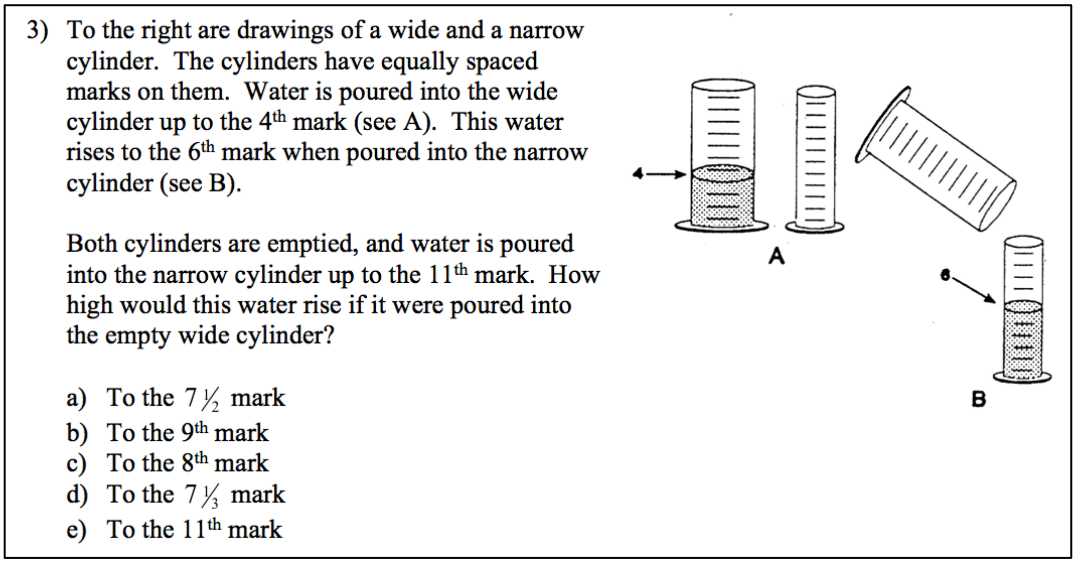
Additionally, I coded another four of the 25 items as "rarely problematic" as opposed to "non-problematic." These not-as-problematic items likely represent less of an issue. However, overlooking these items as completely fine does not do justice to the micro-burdens these types of items may have on students during a stressful test. The corpora of research on test anxiety demonstrates a clear correlation between increased anxiety and decreased test performance (Aydin, 2009; Madsen, 1982). Hence, if these rarely problematic items are increasing anxiety than, in turn, they are increasing the likelihood of optimal test performance for test takers.

In terms of how these findings may relate to wider trends, these results raise interesting implications about the movement to add more real-life context into mathematics items. For instance, the Common Core State Standards in Mathematics (CCSS-M) requires students to "solve real life and mathematical problems using numerical and algebraic expressions and equations" or "solve real world and mathematical problems involving area, surface area, and volume" (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). These types of standards strive to improve the meaning behind math, but they also expose mathematics items to more complexity, more length, more difficult vocabulary, and more content which may favor one group or another. Sometimes, adding a real-life element to mathematics items leads to more difficult items for non-dominant English speakers.

Through this study, I expose such issues empirically. Item three, for example, presents a ratio problem for students (e.g., if 6:4, then 11:?). However, the problem is placed in a real-world context where water is poured from a wide cylinder to narrow cylinder (see Figure 8). Such a context, pouring water from one container to another, is simple and probably relatable; However, the vocabulary is complicated, the word count is high, and the language includes passive voice, conditional clauses, relative clauses, complex question phrases, and abstract or impersonal representations. As Abedi et al.'s framework predicts (Abedi, 2002), participants reported confusion about the item. Most participants had to read through the paragraphs multiple times to ascertain what the item was even asking them to do. Words like "cylinder" provided a hurdle. Many students were able to figure out that cylinder related to the shape in the picture to the right of the question, but adding additional difficulty seemingly created a disadvantage to such students.

**Figure 8**

*Item Three Adds Real World Context to An Item that Tests Conceptual Knowledge of Ratios*



3) To the right are drawings of a wide and a narrow cylinder. The cylinders have equally spaced marks on them. Water is poured into the wide cylinder up to the 4th mark (see A). This water rises to the 6th mark when poured into the narrow cylinder (see B).

Both cylinders are emptied, and water is poured into the narrow cylinder up to the 11th mark. How high would this water rise if it were poured into the empty wide cylinder?

a) To the $7\frac{1}{2}$ mark
b) To the 9th mark
c) To the 8th mark
d) To the $7\frac{1}{3}$ mark
e) To the 11th mark

For test developers, careful consideration of the consequences of adding more context must be taken into account during the test development process. Adding real-life context to a mathematics problem may conform better to the CCSS-M (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010), but it may also disadvantage or present micro-burdens for multilingual test-takers (Kyttälä & Björn, 2014). As one participant stated when taking the test, "Every time I see a word problem, I get stressed, still."

## Discussion of Phases Two and Three

The discussion of phases two and three are combined because phase two was the development phase and phase three was the study of the efficacy of that development.

Firstly, during the development of this dissertation one of the premier journals in the field of educational measurement published a study written by Abedi et al. which overlapped significantly with this dissertation (2020). Abedi et al. studied five different language-based accommodations as given to approximately 3,000 Grade 9 ELL and non-ELL students. One of those accommodations included presenting a fully translated Spanish version of the mathematics items to test-takers who indicated their preference. Abedi et al. found that such an accommodation had significantly negative effects on test scores. Multilingual learners performed worse on tests using this type of accommodation. Abedi et al. (2020) concluded:

> Our results also show that some ELL accommodations, including Spanish Math Test and Bilingual Glossary, had significantly lower mean scores than the control group. This may be due to the lack of alignment between language of instruction and language of assessment and issues related to translations. If students are instructed in English, then the Spanish translation and the Spanish glossary may not produce valid results no matter how proficient they are in Spanish. (p. 47-28)

They argued that the language of instruction played a more significant role than they originally expected during the process of designing accommodations, which is an idea discussed and supported by some researchers (Komba & Bosco, 2015; Palmer & Lynch, 2008). In fact, the only accommodation that appeared to have a positive effect was an accommodation that simply reduced the linguistic complexity of test items but continued to present such item exclusively in English.

The test delivery method that I used in this dissertation, the LAT, did not completely overlap with the study presented by Abedi et al. because the LAT incorporated more nuance when accounting for multilingual backgrounds. Multilingual students could toggle between languages, using either their dominant language, or English, or both when reading and working through a test item. In other words, they could use both their dominant language and their likely language of instruction, if those were not the same. Yet, findings proved to be similar. The LAT did not have a significant effect on the scores of multilingual learners, at the level of the whole test or individual item scores.

The mixed-methods nature of this dissertation provided both inferential evidence that the LAT had no significant effect *and* some answers when exploring why. For one, most participants appeared to feel that the LAT was the same as a typical test. Most participants simply did not use the toggle button. Some participants felt that their English was or should be good enough, some felt like translations would not make a difference for math, and others felt that they could understand the item, so translations were not needed. In most cases, the affordances of the LAT had no effect because they were not used. Further study is needed to parse out exactly why, but certainly pressing the toggle button or reading a translated version of an item does not come naturally to most participants while working through a test item.

It appeared that the toggle button was also too awkward of a method, or the time or effort costs were too high, for switching between languages quickly and naturally. This awkwardness or effort to switch fluidly stands in contrast to how multilingual participants use their full linguistic repertoire when thinking and being as theorized by translanguaging (García & Wei, 2014). This was always a concern in the design. Creating a method of test delivery allowing for the “dynamic semiotic and linguistic practices” of multilingual students always carried the risk of treating multilingualism as  “a double monolingualism,” a phrase used by Prinsloo and Krause (2019) to describe approaches to multilingualism that are not translanguaging (p. 1). Based on the in-depth interviews, the LAT test delivery format did not appear to encourage dynamic linguistic practices.

## Thoughts on Applying Translanguaging Theory to Standardized Testing

Applying the translanguaging to applied contexts is especially difficult (Ascenzi-Moreno, 2018). On one hand, creating a toggled test allows test-takers to use their full linguistic repertoire. Test-takers have a choice to use standardized English or their self-identified dominant language or both, toggling back-and-forth, for each test item. However, creating a natural feeling flow that captures the way multilingual test-takers think and use their linguistic repertoire may not be possible on a test that resembles something standardized. Switching between one fully English and one fully Spanish or Mandarin Chinese version of a test item by clicking a toggle button is not close to the fluidity and complexity with which multilingual students think through their linguistic understanding (García & Wei, 2014). In the theory of translanguaging, such wholesale switching resembles more theoretical code-switching more than translanguaging. Additionally, the act of clicking a toggle button to do so, is clunky and unnatural. Hence, so few test-takers in the throes of a difficult test took the time or click the button. Most focused on the task at hand and moved on.

However, the test delivery design still contributes to the field because applying translanguaging to a standardized artifact, like a standardized test, will never be a perfect fit (Ascenzi-Moreno, 2018). There is a fundamental disconnect between applying a standardized measuring stick to the decidedly diverse and unstandardized populations that enter universities in the United States who take a placement exam (W. Au, 2010; Castagno & Brayboy, 2008). Not only do mathematics skills and abilities exist, which the test seeks to differentiate, but so do the ways in which students read, interpret, and interact with the test items themselves (Almond et al., 2010; Winter et al., 2006).

## Limitations

Studies relating to standardized testing often struggle to stimulate enough motivation, or stress, to simulate a true testing environment. Though, I attempted to create a familiar testing environment by proctoring the test online as multiple students took the test side-by-side, many participants may have lacked enough motivation to work with the test items the same way they would in a real testing environment.

Additionally, to counteract the relatively low number of participants that I would be able to recruit, I conducted a counter-balanced within-subjects design to compare the two versions of the test. Such a design means that participants received the same test items twice with only 2-8 weeks in between. Though research shows that is typically enough time for participants to forget the items (Streiner & Norman, 2008), it may have influenced the way participants interacted with the test.

There was also a unique challenge in this study which may affect how these findings can be generalized. During the middle of the test development process, the COVID-19 virus began to spread around the world. During this time, many schools transitioned to online or hybrid classes, which often included online examinations. The features of such designs came up in conversation several times during my in-depth interviews as participants compared the LAT or the non-LAT version of the PCA with other tests they had taken recently online. Often, these tests were proctored by professors who used anti-cheating software that required students to keep their attention on the screen, limit their computer abilities, and limit the sounds within their test environment. Taking a test like the LAT around the same time as experiencing these unique changes to their own testing experiences may have had some effect on their perceptions.

Additionally, the broad nature of this study sacrificed some of its potential depth. For example, the in-depth interviews that I conducted at the end of phase three would have improved with more participants and a second round of in-depth interviews after conducting an analysis of the quantitative results of the test. However, time constraints lead to conducting the interviews immediately after proctoring the LAT which meant that not much time had passed between the test and the interview. This was good for attaining clearer memories from participants, but this also meant that I had not yet scored the participants exams. So instead of asking more pointed questions about specific test items, I asked more broad interview questions about their perceptions of the test.

The study also may not always effectively parse linguistic differences between students in a nuanced way. As stated earlier, culture, as a concept is incredibly complex and linked to language (Rogoff, 2003; Wertsch, 1993). I predicted that participants who self-identified as dominant speakers of Spanish or Mandarin Chinese would or could use their identified dominant language to overcome linguistic difficulties in test items. Instead, I found that participants varied in the ways they used language in an academic context. Creating a better measure of English language ability or a better way to learn the linguistic background of each participant would have improved the study. Regardless, I had to place test-takers into large bins based on their self-identified dominant language for the quantitative analyses regardless of their linguistic ability or nuance linguistic or cultural background.

Lastly, via this study I assumed that testing the LAT efficacy on Spanish and Chinese speakers could be generalized to a larger population of linguistic minorities, but this may not apply broadly. I chose to work with Spanish and Chinese speakers because they represent the two more commonly spoken non-English languages in the United States (U.S. Census Bureau, 2015). However, these two languages only represent a small fraction of the linguistic diversity of test-takers in the United States.

## Recommendations for Further Study

Considering the recent findings offered by Abedi et al. (2020) in conjunction with the findings from this study, I recommend further study regarding ways and methods to reduce the linguistic complexity of items. Well-regarded and well-validated tests like the PCA have existing linguistic issues, but counteracting these linguistic issues can be extremely difficult (W. Au, 2010). Based on the literature and on my own experience conducting this study (Abedi, 2002, 2009; Abedi et al., 2020), one potential solution to the issue seems to simplify the language as much as possible to increase understanding. Researchers have test this type of solution empirically in the past (Abedi et al., 2006, 2020), but more research addressing how to apply this solution during the test development process could reduce the types of linguistic issues seem on tests.

Researchers have tested carefully prescribed methods of pedagogy in classrooms around the world (Palinscar & Brown, 1984). However, more research can be done to see how carefully prescribed test development processes can reduce linguistic issues in test development as well (Ascenzi-Moreno, 2018). One process, for example could apply Abedi et al.'s (2001) framework of linguistic complexity, Johnstone et al.'s framework of universal design in testing (Johnstone et al., 2008), and automated readability measures like Coh-Metrix measures (Graesser et al., 2011) to examine test items from multiple perspectives. Personal experience, also while working at ACT Inc., revealed that these types of processes are not part of the test development process. ACT Inc., for example, relies on DIF and cognitive interviews during test processes (Collins, 2014; Holland &

Wainer, 2012), but more work on applying these frameworks could lead to even fewer instance of linguistic issues on items. Simply put, researchers have shown that reducing linguistic complexity reduces CIV for multilingual students (Abedi et al., 2020; Martiniello, 2009), yet applied methods for reducing linguistic complexity could use more research.

Lastly, directions for further research could include more research on the consequential validity of testing issues related to linguistic complexity. There are two potential directions for further research regarding this issue. Novel test development methods must include research on the consequential validity (Messick, 1989b). In particular, researchers could review both the unintended and the unintended consequences, both negative and positive, of using these type of tests. Beardsely and Collins (2012) provide one example of this type of study applied to a particular value-added measurement system.

Further research could also be done on the consequential validity of current testing practices through the lens of linguistic issues. Following the footsteps of previous research on linguistic bias in testing (Aguirre-Muñoz, 2000; Chen & Henning, 1985; Haladyna & Downing, 2004), this type of study would reframe linguistic bias through the lens of theory on validity. In particular, this type of study using consequential validity as a frame would provide a way to examine the applied, real-world linguistic effects of tests. Though this dissertation contributed to the field by evaluating a novel type of test delivery, more applied research must be done to apply the strong theoretical research currently redefining multilingualism, like translanguaging, to the ways in which we deliver, and also score and use tests.

# REFERENCES

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educational Assessment*, *8*(3), 231–257. https://doi.org/10.1207/S15326977EA0803_02

Abedi, J. (2009). Computer Testing as a Form of Accommodation for English Language Learners. *Educational Assessment*, *14*(3–4), 195–211. https://doi.org/10.1080/10627190903448851

Abedi, J., Courtney, M., Leon, S., Kao, J., & Azzam, T. (2006). English Language Learners and Math Achievement: A Study of Opportunity to Learn and Language Accommodation. Technical Report 702. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*. https://files.eric.ed.gov/fulltext/ED495848.pdf

Abedi, J., & Ewers, N. (2013). Accommodations for English language learners and students with disabilities: A research-based decision algorithm. *Smarter Balanced Assessment Consortium*. https://portal.smarterbalanced.org/library/en/accommodations-for-english-language-learners-and-students-with-disabilities-a-research-based-decision-algorithm.pdf

Abedi, J., & Linquanti, R. (2012). Issues and opportunities in improving the quality of large scale assessment systems for English language learners. *Understanding Language: Language, Literacy, and Learning in the Content Areas*. https://ell.stanford.edu/publication/issues-and-opportunities-improving-quality-large-scale-assessment-systems-ells

Abedi, J., & Lord, C. (2001). The Language Factor in Mathematics Tests. *Applied Measurement in Education*, *14*(3), 219–234. https://doi.org/10.1207/S15324818AME1403_2

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance*. Center for the Study of Evaluation, University of California, Los Angeles. http://cresst.org/wp-content/uploads/TECH478.pdf

Abedi, J., Zhang, Y., Rowe, S. E., & Lee, H. (2020). Examining Effectiveness and Validity of Accommodations for English Language Learners in Mathematics: An Evidence-Based Computer Accommodation Decision System. *Educational*

*Measurement: Issues and Practice*, *39*(4), 41–52. https://doi.org/10.1111/emip.12328

Aguirre-Muñoz, Z. (2000). The impact of language background characteristics on complex performance assessments: Linguistic accommodation strategies for English language learners. *Unpublished Doctoral Dissertation, University of California, Los Angeles*.

Almond, P., Winter, P., Cameto, R., Russell, M., Sato, E., Clarke-Midura, J., Torres, C., Haertel, G., Dolan, R., & Beddow, P. (2010). Technology-enabled and universally designed assessment: Considering access in measuring the achievement of students with disabilities—A foundation for research. *The Journal of Technology, Learning and Assessment*, *10*(5). https://ejournals.bc.edu/index.php/jtla/article/view/1605

American Educational Researchers Association (AERA), American Psychological Association (APA), & National Council for Measurement in Education (NCME). (2014). *The standards for educational and psychological testing*. https://www.testingstandards.net/open-access-files.html

Amrein-Beardsley, A., & Collins, C. (2012). The SAS education value-added assessment system (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and unintended consequences. *Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas*, *20*, 1–28.

Ascenzi-Moreno, L. (2018). Translanguaging and responsive assessment adaptations: Emergent bilingual readers through the lens of possibility. *Language Arts*, *95*(6), 355–369.

Au, K. H., & Kawakami, A. J. (1985). Research currents: Talk story and learning to read. *Language Arts*, *62*(4), 406–411.

Au, K., & Jordan, C. (1981). Teaching reading to Hawaiian children: Finding a culturally appropriate solution. In H. T. Trueba, G. P. Guthrie, & K. Au (Eds.), *Culture and the bilingual classroom: Studies in classroom ethnography* (pp. 139–152). Random House.

Au, W. (2010). *Unequal by design: High-stakes testing and the standardization of inequality*. Routledge.

Aydin, S. (2009). Test anxiety among foreign language learners: A review of literature. *Journal of Language and Linguistic Studies*, *5*(1).

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *ArXiv Preprint ArXiv:1406.5823*. https://doi.org/10.18637/jss.v067.i01

Benítez, I., Padilla, J. L., van de Vijver, F., & Cuevas, A. (2018). What Cognitive Interviews Tell Us about Bias in Cross-cultural Research: An Illustration Using Quality-of-life Items. *Field Methods*, *30*(4), 277–294. https://doi.org/10.1177/1525822x18783961

Benítez, I., & Padilla, J.-L. (2014). Analysis of nonequivalent assessments across different linguistic groups using a mixed methods approach: Understanding the causes of differential item functioning by cognitive interviewing. *Journal of Mixed Methods Research*, *8*(1), 52–68. https://doi.org/10.1177/1558689813488245

Berardi, G., Burns, D., Duran, P., Gonzalez-Plaza, R., Kinley, S., Robbins, L., Williams, T., & Woods, W. (2003). The tribal environment and natural resources management approach to Indian education and student assessment. *Journal of American Indian Education*, *42*(1), 58–74.

Boeije, H. (2002). A purposeful approach to the constant comparative method in the analysis of qualitative interviews. *Quality and Quantity*, *36*(4), 391–409. https://doi.org/10.1023/A:1020909529486

Booth-Kewley, S., Larson, G. E., & Miyoshi, D. K. (2007). Social desirability effects on computerized and paper-and-pencil questionnaires. *Computers in Human Behavior*, *23*(1), 463–477. https://doi.org/10.1016/j.chb.2004.10.020

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Bourdieu, P. (1984). *Distinction: A social critique of the judgement of taste*. Routledge & Kegan Paul.

Box, G. E. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, *36*(3/4), 317–346. https://doi.org/10.1093/biomet/36.3-4.317

Bronfenbrenner, U. (2009). *The ecology of human development: Experiments by nature and design*. Harvard university press.

Camara, W. (2014). *Standards for Educational and Psychological Testing: Historical Notes.* http://www.aera.net/Portals/38/docs/Outreach/Standards_Hill_Briefing_Slides_FINAL.pdf?timestamp=1410876719244

Carlbring, P., Brunt, S., Bohman, S., Austin, D., Richards, J., Öst, L.-G., & Andersson, G. (2007). Internet vs. Paper and pencil administration of questionnaires commonly used in panic/agoraphobia research. *Computers in Human Behavior*, *23*(3), 1421–1434. https://doi.org/10.1016/j.chb.2005.05.002

Carlson, M. (1997). Obstacles for College Algebra Students in Understanding Functions: What Do High-Performing Students Really Know?. *AMATYC Review*, *19*(1), 48–59.

Carlson, M. (1998). A cross-sectional investigation of the development of the function concept. In A. H. Shoenfeld, J. Kaput, E. Dubinsky, & T. Dick (Eds.), *Research in Collegiate Mathematics Education III* (pp. 114–162). American Mathematical Society and CBMS. https://bookstore.ams.org/cbmath-7

Carlson, M., & Bloom, I. (2005). The cyclic nature of problem solving: An emergent multidimensional problem-solving framework. *Educational Studies in Mathematics*, *58*(1), 45–75. https://doi.org/10.1007/s10649-005-0808-x

Carlson, M., Oehrtman, M., & Engelke, N. (2010). The precalculus concept assessment: A tool for assessing students' reasoning abilities and understandings. *Cognition and Instruction*, *28*(2), 113–145. https://doi.org/10.1080/07370001003676587

Carter, D. (2008). Achievement as resistance: The development of a critical race achievement ideology among Black achievers. *Harvard Educational Review*, *78*(3), 466–497. https://doi.org/10.17763/haer.78.3.83138829847hw844

Casaletto, K. B., Umlauf, A., Marquine, M., Beaumont, J. L., Mungas, D., Gershon, R., Slotkin, J., Akshoomoff, N., & Heaton, R. K. (2016). Demographically corrected normative standards for the Spanish language version of the NIH toolbox cognition battery. *Journal of the International Neuropsychological Society*, *22*(3), 364–374. https://doi.org/10.1017/S135561771500137X

Castagno, A. E., & Brayboy, B. M. J. (2008). Culturally responsive schooling for Indigenous youth: A review of the literature. *Review of Educational Research*, *78*(4), 941–993. https://doi.org/10.3102/0034654308323036

Chalmers, R. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, *2*(2), 155–163.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31–44. https://doi.org/10.1111/j.1745-3992.1998.tb00619.x

Coles-Ritchie, M., & Charles, W. (2011). Indigenizing assessment using community funds of knowledge: A critical action research study. *Journal of American Indian Education*, 26–41.

Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research*, *12*(3), 229–238. https://doi.org/10.1023/A:1023254226592

Collins, D. (2014). *Cognitive interviewing practice*. Sage.

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston.

Crenshaw, K. (1990). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, *43*, 1241. https://heinonline.org/HOL/LandingPage?handle=hein.journals/stflr43&div=52&id=&page=

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research*. Sage publications.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston.

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist*, *30*(1), 1. https://doi.org/10.1037/0003-066X.30.1.1

Dasen, P. R. (1984). The cross-cultural study of intelligence: Piaget and the Baoule. *International Journal of Psychology*, *19*(1–4), 407–434. https://doi.org/10.1080/00207598408247539

Dendrinos, B. (2013). Social meanings in global-glocal language proficiency exams. In D. Tsagari, S. Papadima-Sophocleous, & S. Loannou-Georgiou (Eds.), *International experience in language testing and assessment* (pp. 33–58). Peter Lang.

Denzin, N. K. (2012). Triangulation 2.0. *Journal of Mixed Methods Research*, *6*(2), 80–88. https://doi.org/10.1177/1558689812437186

DeVellis, R. F. (2016). *Scale development: Theory and applications* (Vol. 26). Sage publications.

Dumas, D. G., & McNeish, D. M. (2017). Dynamic measurement modeling: Using nonlinear growth models to estimate student learning capacity. *Educational Researcher*, *46*(6), 284–292. https://doi.org/10.3102/0013189X17725747

Duncan, T. G., Parent, L. del R., Chen, W.-H., Ferrara, S., Johnson, E., Oppler, S., & Shieh, Y.-Y. (2005). Study of a dual-language test booklet in eighth-grade mathematics. *Applied Measurement in Education*, *18*(2), 129–161. https://doi.org/10.1207/s15324818ame1802_1

Elliott, S. N., Kurz, A., Beddow, P., & Frey, J. (2009). Cognitive load theory: Instruction-based research with applications for designing tests. *Proceedings of the National Association of School Psychologists' Annual Convention*, *24*, 1–22.

Elliott, S. N., & Marquart, A. M. (2004). Extended time as a testing accommodation: Its effects and perceived consequences. *Exceptional Children*, *70*(3), 349–367. https://doi.org/10.1177/001440290407000306

Engelke, N. (2007). *Students' understanding of related rates problems in calculus* [PhD Thesis]. Arizona State University.

Engelke, N., Oehrtman, M., & Carlson, M. (2005). Composition of functions: Precalculus students' understandings. *Proceedings of the Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*.

Evers, J. C. (2015). Elaborating on thick analysis: About thoroughness and creativity in qualitative analysis. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, *17*(1). https://doi.org/10.17169/fqs-17.1.2369

Feuchtwang, S. (1990). *Racism: Territoriality and ethnocentricity* (A. X. Cambridge & S. Feuchtwang, Eds.). Avebury.

Freedle, R. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. *Harvard Educational Review*, *73*(1), 1–43. https://doi.org/10.17763/haer.73.1.8465k88616hn4757

Gallagher, A., Bennett, R. E., & Cahalan, C. (2000). Detecting construct-irrelevant variance in an open-ended, computerized mathematics task. *ETS Research Report Series*, *2000*(2), i–32.

García, O., & Wei, L. (2014). *Translanguaging: Language, Bilingualism, and Education*. Palgrave Pivot. https://doi.org/10.1057/9781137385765

Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement*, *6*(1), 109–127.

Gipps, C. (2002). *Beyond testing: Towards a theory of educational assessment*. Routledge.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, *18*(8), 519. https://doi.org/10.1037/h0049294

Glick, J., Sharp, D., Cole, M., & Gay, J. (1969). Linguistic structure and transposition. *Science*, *164*(3875), 90–91. https://doi.org/10.1126/science.164.3875.90

Gould, S. J. (1996). *The mismeasure of man*. WW Norton & Company.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*(5), 223–234.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202.

Greenfield, P. M. (1997). Culture as process: Empirical methods for cultural psychology. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology: Theory and method (2nd Ed), Vol 1: Theory and method.* (pp. 301–346). Allyn & Bacon.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2004). *Adapting educational and psychological tests for cross-cultural assessment*. Psychology Press.

Haney, W. (1981). Validity, vaudeville, and values: A short history of social concerns over standardized testing. *American Psychologist*, *36*(10), 1021. https://doi.org/10.1037/0003-066X.36.10.1021

Harmon, D. A. (2012). Culturally responsive teaching though a historical lens: Will history repeat itself? *Interdisciplinary Journal of Teaching and Learning*, *2*(1), 12–22.

Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evidence-Based Nursing*, *18*(3), 66–67.

Heath, S. B. (1983). *Ways with words: Language, life and work in communities and classrooms*. Cambridge University Press.

Hesse-Biber, S. N. (2010). *Mixed methods research: Merging theory with practice*. Guilford Press.

Holland, P. W., & Wainer, H. (2012). *Differential item functioning*. Routledge.

Hood, S. (1998). Culturally responsive performance-based assessment: Conceptual and psychometric considerations. *Journal of Negro Education*, *67*(3), 187–196. https://doi.org/10.2307/2668188

Hornberger, N. H., & Link, H. (2012). Translanguaging and transnational literacies in multilingual classrooms: A biliteracy lens. *International Journal of Bilingual Education and Bilingualism*, *15*(3), 261–278. https://doi.org/10.1080/13670050.2012.658016

Howe, K. R. (1988). Against the quantitative-qualitative incompatibility thesis or dogmas die hard. *Educational Researcher*, *17*(8), 10–16. https://doi.org/10.3102/0013189X017008010

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, *33*(7), 14–26. https://doi.org/10.3102/0013189X033007014

Johnstone, C. J., Thompson, S. J., Bottsford-Miller, N. A., & Thurlow, M. L. (2008). Universal design and multimethod approaches to item review. *Educational Measurement: Issues and Practice*, *27*(1), 25–36.

Jones, L. V., & Thissen, D. (2006). A history and overview of psychometrics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 1–27). North Holland.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342.

Kane, M. T. (2006). Validation. In R. Brennan, *Educational Measurement* (pp. 17–64). Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kassambara, A. (2013). *rstatix: Pipe-friendly framework for basic statistical tests* (0.7.0) [Computer software]. https://CRAN.R-project.org/package=rstatix

Kiplinger, V. L., Haug, C. A., & Abedi, J. (2000). *Measuring math – not reading – on a math assessment: A language accommodations study of English language learners and other special populations.* Annual Meeting of the American Educational Research Association, New Orleans. https://eric.ed.gov/?id=ED441813

Komba, S. C., & Bosco, S. (2015). *Do students' backgrounds in the language of instruction influence secondary school academic performance?*

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kosinski, M., & Rust, J. (2011). *The development of Concerto: An open-source online adaptive testing platform*. International Association for Computerized Adaptive Testing, Pacific Grove, CA.

Kruse, A. J. (2016). Cultural Bias in testing: A review of literature and implications for music education. *Update: Applications of Research in Music Education*, *35*(1), 23–31. https://doi.org/10.1177/8755123315576212

Kyttälä, M., & Björn, P. M. (2014). The role of literacy skills in adolescents' mathematics word problem performance: Controlling for visuo-spatial ability and mathematics anxiety. *Learning and Individual Differences*, *29*, 59–66.

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

Lambert, W. E. (1973). *Culture and language as factors in learning and education.* Annual Learning Symposium on "Cultural Factors in Learning," Bellingham, WA.

Lee, C. D. (1998). Culturally responsive pedagogy and performance-based assessment. *Journal of Negro Education*, *67*(3), 268–279. https://doi.org/10.2307/2668195

Lévy-Bruhl, L. (1926). *How natives think*. Princeton University Press.

Lewis, G., Jones, B., & Baker, C. (2012). Translanguaging: Developing its conceptualisation and contextualisation. *Educational Research and Evaluation*, *18*(7), 655–670. https://doi.org/10.1080/13803611.2012.718490

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*(8), 437–448. https://doi.org/10.3102/0013189X07311286

Liu, K. K., Anderson, M. E., Swierzbin, B., & Thurlow, M. L. (1999). *Bilingual Accommodations for Limited English Proficient Students on Statewide Reading Tests: Phase 1. State Assessment Series, Minnesota Report 20.* https://eric.ed.gov/?id=ED441305

Lopez, A. A., Turkan, S., & Guzman-Orth, D. (2017). Conceptualizing the use of translanguaging in initial content assessments for newly arrived emergent bilingual students. *ETS Research Report Series*, *2017*(1), 1–12. https://doi.org/10.1002/ets2.12140

Lun, V. M.-C., Fischer, R., & Ward, C. (2010). Exploring cultural differences in critical thinking: Is it about my thinking style or the language I speak? *Learning and Individual Differences*, *20*(6), 604–616. https://doi.org/10.1016/j.lindif.2010.07.001

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.

Luyt, R. (2012). A framework for mixing methods in quantitative measurement development, validation, and revision: A case study. *Journal of Mixed Methods Research*, *6*(4), 294–316. https://doi.org/10.1177/1558689811427912

Madsen, H. S. (1982). Determining the debilitative impact of test anxiety. *Language Learning*, *32*(1), 133–143.

Mair, P., & Wilcox, R. (2020). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods*, *52*, 464–488.

Maller, S. J. (2003). Best practices in detecting bias in nonverbal tests. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 23–47). Kluwer Academic/Plenum Publishers.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment*, *14*(3–4), 160–179.

McCoy, S., Marks, P. V., Carr, C. L., & Mbarika, V. (2004). Electronic versus paper surveys: Analysis of potential psychometric biases. *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*. Hawaii

International Conference on System Sciences, Big Island, HI. https://doi.org/10.1109/HICSS.2004.1265634

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. https://doi.org/10.3102/0013189X018002005

Messick, S. (1989b). Validity. In R. Linne (Ed.), *Educational measurement* (pp. 13–103). New York: Macmillan Publishing.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2018). *Qualitative data analysis: A methods sourcebook*. Sage publications.

Miller, K., Willson, S., Chepp, V., & Ryan, J. M. (2014). Analysis. In K. Miller, Chepp, Valerie, S. Willson, & J. L. Padilla (Eds.), *Cognitive interviewing methodology* (pp. 35–50). John Wiley.

Milner IV, H. R. (2008). Critical race theory and interest convergence as analytic tools in teacher education policies and practices. *Journal of Teacher Education*, *59*(4), 332–346. https://doi.org/10.1177/0022487108321884

Mohatt, G., & Erickson, F. (1981). Cultural differences in teaching styles in an Odawa school: A sociolinguistic approach. In H. T. Trueba, G. P. Guthrie, & K. H. Au (Eds.), *Culture and the bilingual classroom: Studies in classroom ethnography* (pp. 105–119). Newbury House.

Nagai, Y. (2001). Developing assessment and evaluation strategies for vernacular elementary school classrooms: A collaborative study in Papua New Guinea. *Anthropology & Education Quarterly*, *32*(1), 80–103. https://www.jstor.org/stable/3196212

National Council for Measurement in Education (NCME). (2020). *Position statement on testing English learners*. https://higherlogicdownload.s3.amazonaws.com/NCME/c53581e4-9882-4137-987b-4475f6cb502a/UploadedImages/English_learners_Statement_sept_2020.pdf

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics*. National Governors Association Center for Best Practices, Council of Chief State School Officers. http://www.corestandards.org/read-the-standards/

National Research Council. (2004). *Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessment* (J. A. Koenig & L. F. Bachman, Eds.). The National Academies Press.

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. https://doi.org/10.1002/tea.21026

Oehrtman, M., Carlson, M., & Thompson, P. W. (2008). Foundational reasoning abilities that promote coherence in students' function understanding. In M. Carlson & C. Rasmussen (Eds.), *Making the connection: Research and teaching in undergraduate mathematics education* (pp. 27–42). Mathematical Association of America.

Otheguy, R., García, O., & Reid, W. (2015). Clarifying translanguaging and deconstructing named languages: A perspective from linguistics. *Applied Linguistics Review*, *6*(3), 281–307. https://doi.org/10.1515/applirev-2015-0014

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture. *Instructional Science*, *32*(1), 1–8. https://doi.org/10.1023/B:TRUC.0000021806.17516.d0

Palinscar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, *1*(2), 117–175.

Palmer, D., & Lynch, A. W. (2008). A bilingual education for a monolingual test? The pressure to prepare for TAKS and its influence on choices for language of instruction in Texas elementary bilingual classrooms. *Language Policy*, *7*(3), 217–235.

Papay, J. P., Murnane, R. J., & Willett, J. B. (2011). *How performance information affects human-capital investment decisions: The impact of test-score labels on educational outcomes*. National Bureau of Economic Research.

Papay, J. P., Murnane, R. J., & Willett, J. B. (2016). The impact of test score labels on human-capital investment decisions. *Journal of Human Resources*, *51*(2), 357–388. https://doi.org/10.3368/jhr.51.2.0713-5837R

Patz, R. J. (2015). From the president: Educational measurement remains a pillar in federal and state education policy. *NCME Newsletter*, 1–3.

Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice*, *30*(3), 10–28. https://doi.org/10.1111/j.1745-3992.2011.00207.x

Prinsloo, M., & Krause, L.-S. (2019). Translanguaging, place and complexity. *Language and Education*, *33*(2), 159–173. https://doi.org/10.1080/09500782.2018.1516778

Puhan, G., Boughton, K., & Kim, S. (2007). Examining Differences in Examinee Performance in Paper and Pencil and Computerized Testing. *Journal of Technology, Learning, and Assessment*, *6*(3).

Qualls, A. L. (1998). Culturally responsive assessment: Development strategies and validity issues. *Journal of Negro Education*, 296–301. https://doi.org/10.2307/2668197

Rameka, L. K. (2011). Being Māori: Culturally relevant assessment in early childhood education. *Early Years*, *31*(3), 245–256. https://doi.org/10.1080/09575146.2011.614222

Revelle, W. (2017). *Psych: Procedures for personality and psychological research* (1.7.8) [Computer software]. Northwestern University. https://CRAN.R-project.org/package=psych

Reynolds, C. R., Lowe, P. A., & Saenz, A. L. (1999). The problem of bias in psychological assessment. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (3rd ed., pp. 549–595). Wiley.

Reynolds, C. R., & Suzuki, L. A. (2003). Bias in psychological assessment: An empirical review and recommendations. In C. R. Reynolds & M. C. Ramsay (Eds.), *Handbook of psychology* (Vol. 10, pp. 67–93). John Wiley & Sons Inc.

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, *26*(1), 108–116. https://doi.org/10.7334/psicothema2013.260

Ritchie, J., Lewis, J., Elam, R. G., Tennant, R., & Rahim, N. (2013). Selecting samples. In J. Ritchie, J. Lewis, C. McNaughton Nicholls, & R. Ormston (Eds.), *Qualitative research practice: A guide for social science students and researchers* (pp. 111–146). Sage.

Rivera, C., & Stansfield, C. W. (2001). *The effects of linguistic simplification of science test items on performance of limited English proficient and monolingual English-speaking students.* Annual Meeting of the American Educational Research Association, Seattle, WA. https://eric.ed.gov/?id=ED455289

Rogoff, B. (2003). *The cultural nature of human development*. Oxford university press.

Rosner, J. (2000). Disparate outcomes by design: University admissions tests. *Berkeley La Raza Law Journal*, *12*, 377–386.

Saakvitne, K. W., Tennen, H., & Affleck, G. (1998). Exploring thriving in the context of clinical trauma theory: Constructivist self development theory. *Journal of Social Issues*, *54*(2), 279–299. https://doi.org/10.1111/j.1540-4560.1998.tb01219.x

Sadker, D., & Zittleman, K. R. (2009). *Still failing at fairness: How gender bias cheats girls and boys in school and what we can do about it*. Simon and Schuster.

Saldaña, J. (2021). *The coding manual for qualitative researchers*. Sage.

Sandelowski, M. (2008). Theoretical saturation. In L. M. Given (Ed.), *The Sage encyclopedia of qualitative methods* (Vol. 1, pp. 875–876). Sage.

Sapir, E. (2004). *Language: An introduction to the study of speech*. Courier Corporation.

Semken, S., & Freeman, C. B. (2008). Sense of place in the practice and assessment of place-based science teaching. *Science Education*, *92*(6), 1042–1057. https://doi.org/10.1002/sce.20279

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, *75*(4), 457–490. https://doi.org/10.3102/00346543075004457

Slee, J. (2010). A systemic approach to culturally responsive assessment practices and evaluation. *Higher Education Quarterly*, *64*(3), 246–260. https://doi.org/10.1111/j.1468-2273.2010.00464.x

Sleeter, C., & Grant, C. (1987). An analysis of multicultural education in the United States. *Harvard Educational Review*, *57*(4), 421–445. https://doi.org/10.17763/haer.57.4.v810xr0v3224x316

Smits, C. H., de Vries, W. M., & Beekman, A. T. (2005). The CIDI as an instrument for diagnosing depression in older Turkish and Moroccan labour migrants: An exploratory study into equivalence. *International Journal of Geriatric Psychiatry*, *20*(5), 436–445. https://doi.org/10.1002/gps.1303

Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English language learners. *Teachers College Record*, *108*(11), 2354. https://doi.org/10.1111/j.1467-9620.2006.00785.x

Solano-Flores, G., Lara, J., Sexton, U., & Navarrete, C. (2001). *Testing English Language Learners: A Sampler of Student Responses to Science and Mathematics Test Items.* ERIC. https://eric.ed.gov/?id=ED466496

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, *38*(5), 553–573. https://doi.org/10.1002/tea.1018

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, *32*(2), 3–13. https://doi.org/10.3102/0013189X032002003

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002a). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing*, *2*(2), 107–129. https://doi.org/10.1207/S15327574IJT0202_2

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002b). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing*, *2*(2), 107–129. https://doi.org/10.1207/S15327574IJT0202_2

Sternberg, R. J. (2007). Culture, instruction, and assessment. *Comparative Education*, *43*(1), 5–22. https://doi.org/10.1080/03050060601162370

Strauss, A., & Corbin, J. M. (1997). *Grounded theory in practice*. Sage.

Streiner, D., & Norman, G. (2008). *Health Measurement Scales: A Practical Guide to Their Development and Use*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199231881.001.0001

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1016/0364-0213(88)90023-7

Takahashi, H., Shen, J., & Ogawa, K. (2016). An experimental examination of compensation schemes and level of effort in differentiated tasks. *Journal of Behavioral and Experimental Economics*, *61*, 12–19. https://doi.org/10.1016/j.socec.2016.01.002

Taylor, K. (2016, April 23). Race and the Standardized Testing Wars. *The New York Times*. https://www.nytimes.com/2016/04/24/opinion/sunday/race-and-the-standardized-testing-wars.html

te Nijenhuis, J., Willigers, D., Dragt, J., & van der Flier, H. (2016). The effects of language bias and cultural bias estimated using the method of correlated vectors on a large database of IQ comparisons between native Dutch and ethnic minority immigrants from non-Western countries. *Intelligence*, *54*, 117–135. https://doi.org/10.1016/j.intell.2015.12.003

Thurlow, M. L., Thompson, S. J., & Lazarus, S. S. (2006). Considerations for the administration of tests to special needs students: Accommodations, modifications, and more. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 653–673). Routledge.

Tienken, C. H., & Zhao, Y. (2013). How common standards and standardized testing widen the opportunity gap. In P. Carter & K. G. Welner (Eds.), *Closing the opportunity gap: What America must do to give every child an even chance* (pp. 113–122). https://doi.org/10.1093/acprof:oso/9780199982981.003.0008

Toulmin, S. E. (2003). *The uses of argument*. Cambridge university press.

Ungar, M., & Liebenberg, L. (2011). Assessing resilience across cultures using mixed methods: Construction of the child and youth resilience measure. *Journal of Mixed Methods Research*, *5*(2), 126–149. https://doi.org/10.1177/1558689811400607

U.S. Census Bureau. (2015). *Detailed Languages Spoken at Home and Ability to Speak English for the Population 5 Years and Over: 2009-2013*. https://www.census.gov/data/tables/2013/demo/2009-2013-lang-tables.html

U.S. Department of Education, & National Center for Education Statistics. (2003). *The validity of oral accommodation in testing: NAEP validity studies.* (NCES No. 2003–06). https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=200306

Valdés, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias.* Ablex Publishing.

Van de Vijver, F., & Leung, K. (1997). *Methods and data analysis of comparative research.* Allyn & Bacon.

Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, *54*(2), 119–135. https://doi.org/10.1016/j.erap.2003.12.004

Van Leest, P. F., & Bleichrodt, N. (1990). Testing of college graduates from ethnic minority groups. In N. Bleichrodt & P. J. D. Drenth (Eds.), *Contemporary issues in cross-cultural psychology. Amsterdam: Swets & Zeitlinger*. Swets & Zeitlinger.

Vaught, S. E., & Castagno, A. E. (2008). “I don’t think I’m a racist”: Critical Race Theory, teacher attitudes, and structural racism. *Race Ethnicity and Education*, *11*(2), 95–113. http://dx.doi.org/10.1080/13613320802110217

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard university press.

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., & Steinberg, L. (2001). Computerized adaptive testing: A primer. *Qual Life Res*, *10*(8), 733–734. https://doi.org/10.1023/A:1016834001219

Ward, E. M. G., Semken, S., & Libarkin, J. C. (2014). The design of place-based, culturally informed geoscience assessment. *Journal of Geoscience Education*, *62*(1), 86–103. https://doi.org/10.5408/12-414.1

Wertsch, J. V. (1993). *Voices of the mind*. Harvard University Press.

Whorf, B. L. (1950). An American Indian model of the universe. *International Journal of American Linguistics*, *16*(2), 67–72. https://www.jstor.org/stable/42581334

Whorf, B. L. (2012). *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*. MIT Press.

Williams, J. E., & McCord, D. M. (2006). Equivalence of standard and computerized versions of the Raven Progressive Matrices Test. *Computers in Human Behavior*, *22*(5), 791–800. https://doi.org/10.1016/j.chb.2004.03.005

Winter, P. C., Kopriva, R. J., Chen, C.-S., & Emick, J. E. (2006). Exploring individual and item factors that affect assessment validity for diverse learners: Results from a large-scale cognitive lab. *Learning and Individual Differences*, *16*(4), 267–276. https://doi.org/10.1016/j.lindif.2007.01.001

Wolfe, E. W., & Manalo, J. R. (2004). Composition medium comparability in a direct writing assessment of non-native English speakers. *Language Learning & Technology*, *8*(1), 53–65. http://dx.doi.org/10125/25229

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42–57. https://doi.org/10.1177/0146621607314044

Word, C. O. (1977). Cross cultural methods for survey research in Black urban areas. *Journal of Black Psychology*, *3*(2), 72–87. https://doi.org/10.1177/009579847700300212

Young, J. R. (2003, October). Researchers charge racial bias on the SAT. *Chronicle of Higher Education*, *50*(7). http://eric.ed.gov/?id=EJ677265

Zoom Video Communications Inc. (2020). *Zoom*. https://zoom.us/download

Zuriff, G. E. (2000). Extra examination time for students with learning disabilities: An examination of the maximum potential thesis. *Applied Measurement in Education*, *13*(1), 99–117. https://doi.org/10.1207/s15324818ame1301_5

# APPENDIX A

# FULL VERSION OF THE PCA

# PCA – Precalculus Assessment

1) Given the function *f*, defined by $f(x)=3x^2+2x-4$, find $f(x+a)$.

a) $f(x+a)=3x^2+3a^2+2x+2a-4$

b) $f(x+a)=3x^2+6xa+3a^2+2x-4$

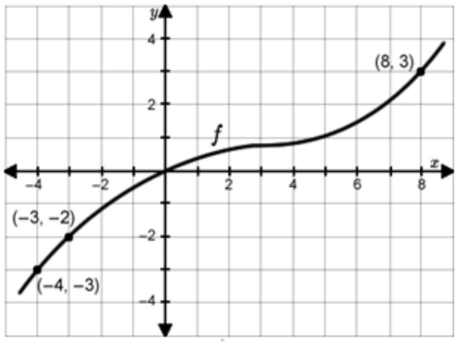
c) $f(x+a)=3(x+a)^2+2(x+a)-4$

d) $f(x+a)=3(x+a)^2+2x-4$

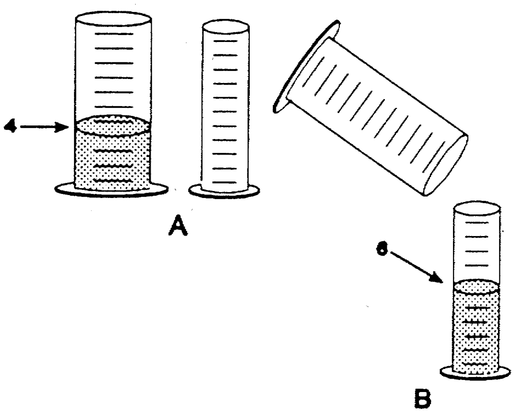
e) $f(x)=3x^2+2x-4+a$

2) Use the graph of *f* to solve $f(x)=-3$ for *x*.

a) (–3,–2)
b) –4
c) (–4,–3)
d) –2
e) –3



3) To the right are drawings of a wide and a narrow cylinder. The cylinders have equally spaced marks on them. Water is poured into the wide cylinder up to the 4th mark (see A). This water rises to the 6th mark when poured into the narrow cylinder (see B).



Both cylinders are emptied, and water is poured into the narrow cylinder up to the 11th mark. How high would this water rise if it were poured into the empty wide cylinder?

a) To the $7\frac{1}{2}$ mark
b) To the 9th mark
c) To the 8th mark
d) To the $7\frac{1}{3}$ mark
e) To the 11th mark

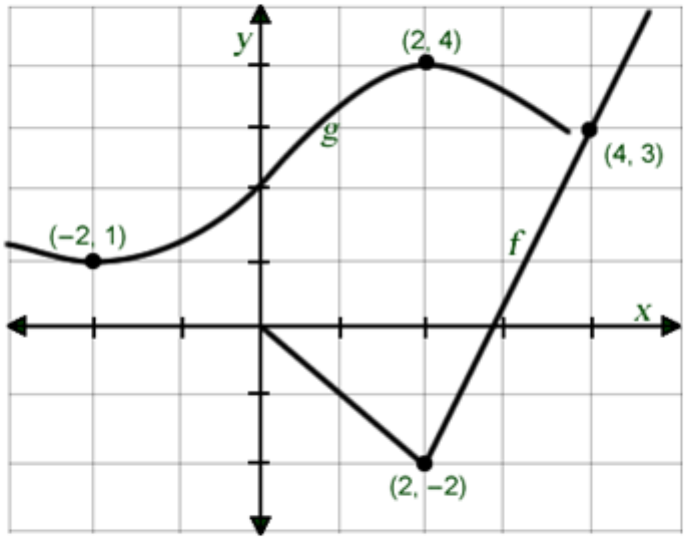
4) Which one of the following formulas defines the area, *A*, of a square in terms of its perimeter, *p*?

a) $A = \frac{p^2}{16}$
b) $A = s^2$
c) $A = \frac{p^2}{4}$
d) $A = 16s^2$
e) $p = 4\sqrt{A}$

Use the graphs of *f* and *g* to answer items 5 and 6.

5) Use the graphs of *f* and *g* to evaluate $g(f(2))$.

a) -2
b) 1
c) 3
d) 4
e) Not defined



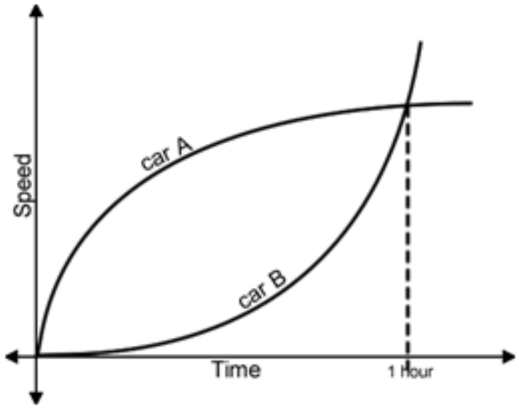
6) Evaluate $f(2) - g(0)$.

a) -4
b) -2
c) 0
d) 2
e) 4

7) The model that describes the number of bacteria in a culture after *t* days has just been updated from $P(t) = 7(2)^t$ to $P(t) = 7(3)^t$. What implications can you draw from this information?

a) The final number of bacteria is 3 times as much of the initial value instead of 2 times as much.
b) The initial number of bacteria is 3 instead of 2.
c) The number of bacteria triples every day instead of doubling every day.
d) The growth rate of the bacteria in the culture is 30% per day instead of 20% per day.
e) None of the above.

The given graph represents speed vs. time for two cars. (Assume the cars start from the same position and are traveling in the same direction.) Use this information and the graph below to answer item 8.

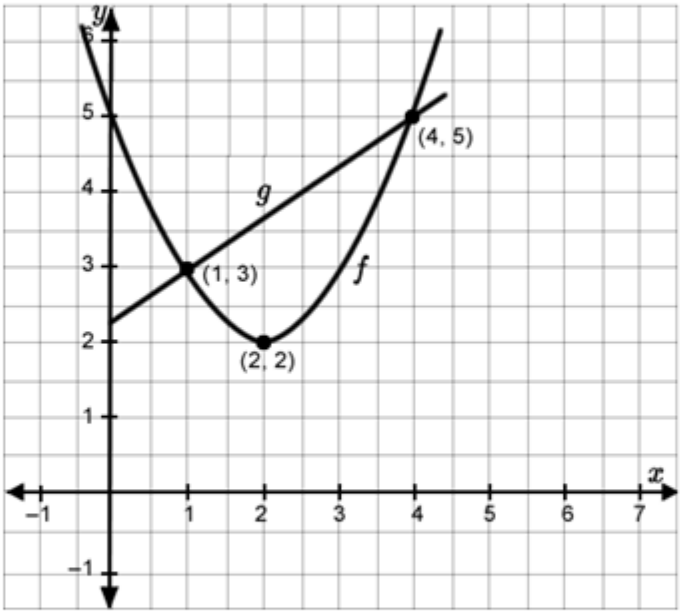


8) What is the relationship between the ***position*** of car A and car B at *t* = 1 hr.?

a) Car A and car B are colliding.
b) Car A is ahead of car B.
c) Car B is ahead of car A.
d) Car B is passing car A.
e) The cars are at the same position.

9) Use the graphs of *f* and *g* to solve $g(x) > f(x)$.

a) $2 < x < 5$
b) $1 < y < 4$
c) $x < 4$
d) $2 < y < 5$
e) $1 < x < 4$



10) A hose is used to fill an empty wading pool. The graph shows volume (in gallons) in the pool as a function of time (in minutes). Which of the following defines a formula for computing the time, *t*, as a function of the volume, *v*?

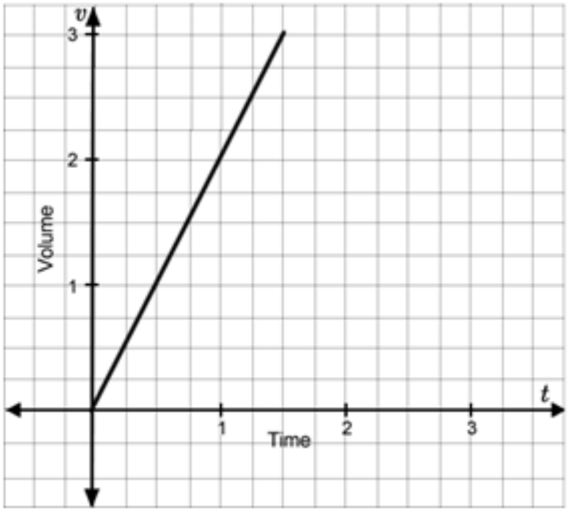
a) $v(t)=\frac{t}{2}$

b) $t(v)=2v$

c) $t(v)=\frac{v}{2}$

d) $v(t)=2t$

e) None of the above



11) The distance, *s* (in feet), traveled by a car moving in a straight line is given by the function, $s(t)=t^2+t$, where *t* is measured in seconds. Find the average velocity for the time period from $t = 1$ to $t = 4$.

a) $5\ {}^{ft}\!/_{sec}$

b) $6\ {}^{ft}\!/_{sec}$

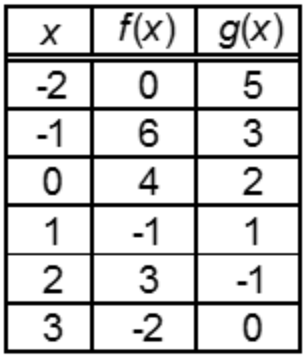
c) $9\ {}^{ft}\!/_{sec}$

d) $10\ {}^{ft}\!/_{sec}$

e) $11\ {}^{ft}\!/_{sec}$

12) Given the table to the right, determine $f\left(g\left(3\right)\right)$.

a) 4
b) -1
c) 0
d) 1
e) 5



| $x$ | $f(x)$ | $g(x)$ |
|---|---|---|
| -2 | 0 | 5 |
| -1 | 6 | 3 |
| 0 | 4 | 2 |
| 1 | -1 | 1 |
| 2 | 3 | -1 |
| 3 | -2 | 0 |

13) Given the table to the right, determine $g^{-1}(-1)$.

a) ½
b) ⅓
c) 1
d) 2
e) 3

14) Given that *f* is defined by $f(t)=100^t$, which of the following is a formula for $f^{-1}$?

a) $f^{-1}(t)=\dfrac{1}{100^t}$

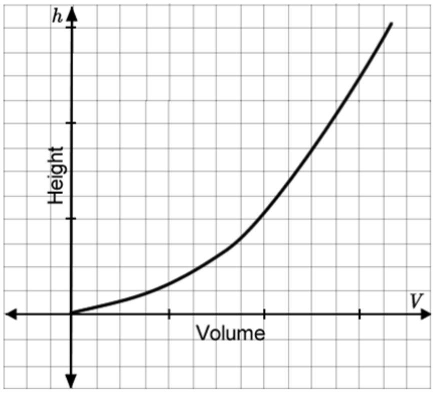
b) $f^{-1}(t)=\dfrac{t}{\ln 100}$

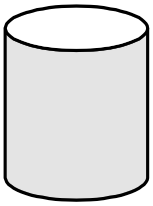
c) $f^{-1}(t)=\dfrac{t}{100^t}$

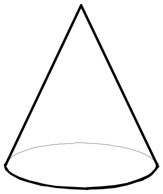
d) $f^{-1}(t)=\dfrac{\ln t}{\ln 100}$

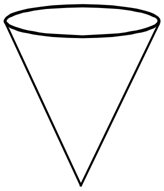
e) $f^{-1}(t)=\dfrac{\ln t}{100}$

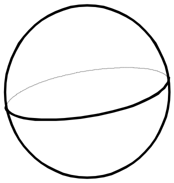
15) The following graph represents the height of water as a function of volume as water is poured into a container. Which container is represented by this graph?

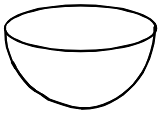


a) 

b) 

c) 

d) 

e) 

16) Given the function $h(x) = 3x - 1$ and $g(x) = x^2$, evaluate $g(h(2))$.

a) 10
b) 11
c) 20
d) 25
e) 36

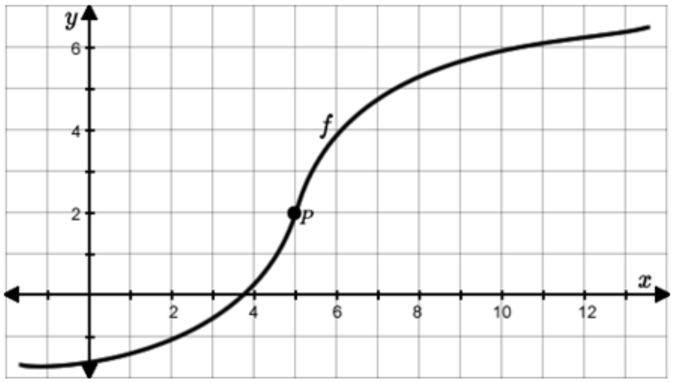
17) A ball is thrown into a lake, creating a circular ripple that travels outward at a speed of 5 cm per second. Express the area, *A*, of the circle in terms of the number of seconds, *t*, that have passed since the ball hits the lake.

a) $A = 25\pi t$
b) $A = \pi r^2$
c) $A = 25\pi t^2$
d) $A = 5\pi t^2$
e) None of the above

18) The wildlife game commission poured 5 cans of fish (each can contained approximately 100 fish) into a farmer's lake. The function *N* defined by $N(t) = \dfrac{600t + 500}{0.5t + 1}$ represents the approximate number of fish in the lake as a function of time (in years). Which one of the following best describes how the number of fish in the lake changes over time?

a) The number of fish gets larger each year, but does not exceed 500.
b) The number of fish gets larger each year, but does not exceed 1200.
c) The number of fish gets smaller every year, but does not get smaller than 500.
d) The number of fish gets larger each year, but does not exceed 600.
e) The number of fish gets smaller every year but does not get smaller then 1200.

19) Using the graph below, explain the behavior of function *f* on the interval from $x = 5$ to $x = 12$.



a) Increasing at an increasing rate.
b) Increasing at a decreasing rate.
c) Increasing at a constant rate.
d) Decreasing at a decreasing rate.
e) Decreasing at an increasing rate.

20) If $S(m)$ represents the salary (per month), in hundreds of dollars, of an employee after $m$ months on the job, what would the function $R(m) = S(m + 12)$ represent?

a) The salary of an employee after $m + 12$ months on the job.
b) The salary of an employee after 12 months on the job.
c) \$12 more than the salary of someone who has worked for $m$ months.
d) An employee who has worked for $m + 12$ months.
e) Not enough information.

21) What is the domain of the following function: $f(x) = \dfrac{\sqrt{x+2}}{x-1}$ ?

a) $(1,\infty)$
b) $x \neq 1$
c) $[-2,1)\cup(1,\infty)$
d) $[-2,\infty)$
e) All real numbers

22) A baseball card increases in value according to the function, $b(t) = \frac{5}{2}t + 100$, where $b$ gives the value of the card in dollars and $t$ is the time (in years) since the card was purchased. Which of the following describe what $\frac{5}{2}$ conveys about the situation?

I. The card's value increases by \$5 every two years.

II. Every year the card's value is 2.5 times greater than the previous year.

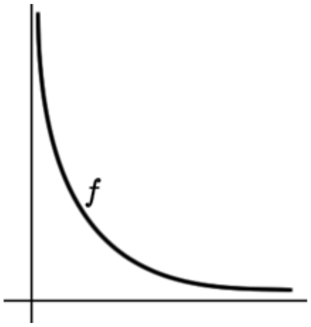
III. The card's value increases by $\frac{5}{2}$ dollars every year.

a) I only
b) II only
c) III only
d) I and III only
e) I, II and III

23) Which of the following best describes the effect of $f^{-1}$, given *f* is a one-to-one function and $f(d) = c$?

a) $f^{-1}$ inverts *f*, so $f^{-1}(d) = \frac{1}{f(d)}$
b) $f^{-1}$ inverts the input to *f*, so. $f^{-1}(d) = \frac{1}{d}$
c) $f^{-1}$ inverts the output to *f*, so $f^{-1}(d) = \frac{1}{c}$
d) $f^{-1}$ inverts $f$, so $f^{-1}(f(d)) = d$
e) a and c

24) A function *f* is defined by the following graph. Which of the following describes the behavior of *f*?



I. As the value of *x* approaches 0, the value of *f* increases.
II. As the value of *x* increases, the value of *f* approaches 0.
III. As the value of *x* approaches 0, the value of *f* approaches 0.

a) I only
b) II only
c) III only
d) I and II
e) II and III

25) Which of the following best describes the behavior of the function *f* defined by,

$$f(x) = \frac{x^2}{x-2}.$$

I. As the value of *x* gets very large, the value of *f* approaches 2.
II. As the value of *x* gets very large, the value of *f* increases.
III. As the value of *x* approaches 2, the value of *f* approaches 0.

a) I only
b) II only
c) III only
d) I and III
e) II and III

# APPENDIX B

# INTITUTIONAL REVIEW BOARD APPROVAL

ASU Knowledge Enterprise Development

EXEMPTION GRANTED

Yi Zheng
Division of Educational Leadership and Innovation - Tempe
-
Yi.Isabel.Zheng@asu.edu

Dear Yi Zheng:

On 3/23/2018 the ASU IRB reviewed the following protocol:

| | |
|---:|---|
| Type of Review: | Initial Study |
| Title: | Developing a Technology-Enhanced Solution to Language Inequality<br>in English-Based Math Tests |
| Investigator: | Yi Zheng |
| IRB ID: | STUDY00007933 |
| Funding: | None |
| Grant Title: | None |
| Grant ID: | None |
| Documents Reviewed: | • Math Test - English Version, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions);<br>• HRP-502a_consent_form_V3_interview.pdf, Category: Consent Form;<br>• Interview_recruitment_script.pdf, Category: Recruitment Materials;<br>• HRP-503a_V4.docx, Category: IRB Protocol;<br>• Draft of Demographics Questions to be asked before the math tests , Category: Measures (Survey questions/Interview questions /interview guides/focus group questions);<br>• HRP-502a_consent_form_V2_test.pdf, Category: Consent Form; |

The IRB determined that the protocol is considered exempt pursuant to Federal Regulations 45CFR46 (1) Educational settings, (2) Tests, surveys, interviews, or observation on 3/23/2018.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Sincerely,

IRB Administrator

cc: Kevin Close
Kevin Close

# APPENDIX C

## PHASE ONE RECRUITMENT VIA ONLINE ADVERTISEMENT

# Native Spanish and Chinese speakers needed for study

Would you like to receive $40 for a few hours of your time by contributing to research on making tests more fair?

Dr. Yi Zheng, professor in the Mary Lou Fulton Teachers College at ASU, is conducting a research study to examine language bias in math tests.

To participate in this study, you must:

- Be a first-year or sophomore ASU student.
- Identify as a native Mandarin Chinese speaker or a native Spanish speaker.
- Have taken a math class in the last year.
- Be 18 years or older.

You will be interviewed in person for one-and-a-half to two hours. Your participation is completely voluntary and data will be made anonymous. If you participate, you will be paid $40. If you have any questions concerning the research study or are interested in participating, please email research assistant, Kevin Close.

# APPENDIX D

# DEMOGRAPHICS INSTRUMENT USED IN PHASES ONE, TWO, AND THREE

Participant ID____________________

What is your gender?

- Male
- Female
- Other
- Do not wish to specify

How old are you (in years)?

What is your ethnicity?

- Asian / Pacific Islander
- Black or African American
- Hispanic or Latino/a
- Native American or American Indian
- White

What is your approximate average household income in your childhood home?

- $0 - $24,999
- $25,000 - $49,999
- $50,000 - $74,999
- $75,000 - $99,999
- $100,000 - $149,999
- $150,000 and up

What is the highest level of math class you have taken?

- Algebra 1
- Algebra 2
- College Algebra
- Trigonometry
- Precalculus
- Calculus
- Calculus II
- Calculus III
- Linear Algebra
- Higher

Rate your current overall math ability

- Very Weak
- Weak
- Slightly Weak
- Slightly Strong
- Strong

- Very Strong

If you took the SAT, what score did you get on the Math Section?

If you took the ACT, what score did you get on the Math Section?

If you took the TOEFL, what score did you get on the following sections?
Listening:
Speaking:
Writing:
Reading:

If you took the IELTS, what score did you get on the following sections?
Listening:
Reading:
Writing:
Speaking:

If you took the Gao Kao (高考), what did you score on the mathematics section?

Rate your current overall language ability in ENGLISH:

- Understand but cannot speak
- Understand and can speak with great difficulty
- Understand and speak but with some difficulty
- Understand and speak comfortably, with little difficulty
- Understand and speak fluently like a native speaker

What is the primary language spoken in your childhood home?

- Chinese
- English
- Spanish
- Other

If English is not the primary language spoken in your childhood home, rate your current overall language ability in the PRIMARY LANGUAGE SPOKEN IN YOUR CHILDHOOD HOME that you checked above:

- Understand but cannot speak
- Understand and can speak with great difficulty
- Understand and speak but with some difficulty
- Understand and speak comfortably, with little difficulty
- Understand and speak fluently like a native speaker

If English is not the primary language spoken in your childhood home, in general, which language do you prefer to use?

- English
- The primary language spoken in your childhood home
- It depends on whom I talk to
- Both

# APPENDIX E

# PHASE ONE CONSENT FORM

**Title of research study:** Developing a technology-enriched solution to language inequality in English-based math tests

**Investigator:** Dr. Yi Zheng

### Why am I being invited to take part in a research study?

We invite you to take part in a research study because you are a student whose first language is not English.

### Why is this research being done?

This research is done to evaluate ways to make tests fairer for student who speak languages other than English.

### How long will the research last?

We expect that individuals will spend 30 minutes to one and a half hour participating in the proposed activities.

### How many people will be studied?

We expect 20-30 people to participate in this interview phase of the research. We expect about 560 people will participate in this research study overall.

### What happens if I say yes, I want to be in this research?

Your interview will be audio recorded and analyzed by researchers. You are free to decide whether you wish to participate in this study.

The interview involves reading through test questions and commenting on how you interpret those questions. You may be rewriting some of those questions into your own words or highlighting parts of questions which seem overly complex.

### What happens if I say yes, but I change my mind later?

You can leave the research at any time it will not be held against you. Your data will then be deleted.

### Is there any way being in this study could be bad for me?

Risk of harm associated with this study is negligible.

### Will being in this study help me in any way?

We cannot promise any benefits to you or others from your taking part in this research. However, possible benefits include advancing knowledge about how to make tests fairer for student who speak languages other than English.

### What happens to the information collected for the research?

Efforts will be made to limit the use and disclosure of your personal information, including research study records, to people who have a need to review this information. We cannot promise complete secrecy. The results of this study may be used in reports, presentations, or publications but your name will not be used.

### Who can I talk to?

If you have questions, concerns, or complaints, talk to the research team by emailing Kevin Close, kevin.close@asu.edu, or Yi Zheng, yi.isabel.zheng@asu.edu.

This research has been reviewed and approved by the Social Behavioral IRB. You may talk to them at (480) 965-6788 or by email at research.integrity@asu.edu if:

- Your questions, concerns, or complaints are not being answered by the research team.
    - You cannot reach the research team.
    - You want to talk to someone besides the research team.
    - You have questions about your rights as a research participant.
    - You want to get information or provide input about this research.

# APPENDIX F

# PHASE ONE SCRIPT

## Script

Thank you for agreeing to participate in this study. To begin, bear with me as I read this script word for word. The study should take about two hours. The first part of the study will be to ask you some questions about yourself. Feel free to skip any question you do not feel comfortable answering.

After answering those questions, I will give you a calculus test to take. However, as you take this test, you will be talking with me about your answers. Think about this as an interesting type of interview. Be aware that I will be recording our conversations the whole time. Is that OK with you?

During this interview you have three directions:

1. Please highlight any words or phrases that you do not understand. This can be vocabulary words or whole phrases
2. Please rewrite each of the math questions in your own words
3. During the process, I want you to talk about what you are thinking. I will also be asking you questions about your thinking.

Remember, rewriting math questions while answering questions is hard. Just do your best and remember that all of this helps us better understand how to make better tests for non-native English speakers in the future.

So again:

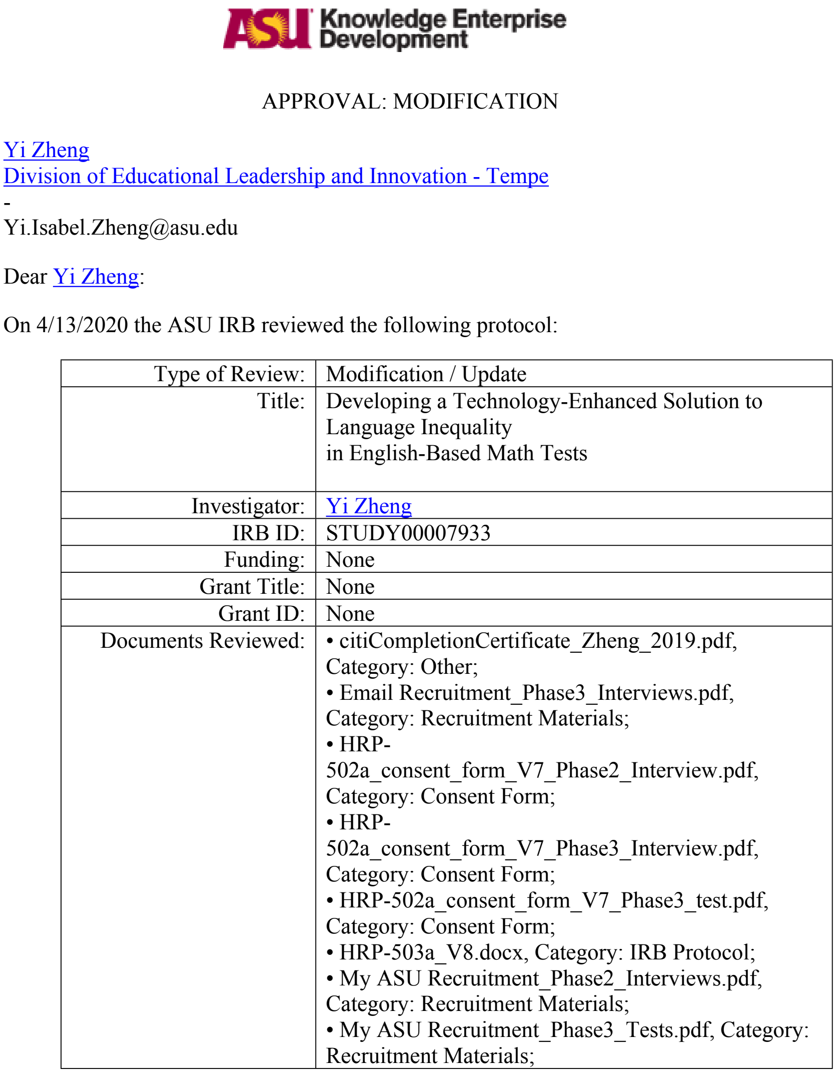
1. Highlight words or phrases that are difficult to understand
2. Rewrite questions in your own words
3. Talk about what you are thinking and answer my questions as best you can.

After everything is over, I'll have you sign for your $40 in cash.

Thank you again!

# APPENDIX G

# INTITUTIONAL REVIEW BOARD MODIFICATION APPROVAL



Knowledge Enterprise Development

APPROVAL: MODIFICATION

Yi Zheng
Division of Educational Leadership and Innovation - Tempe
-
Yi.Isabel.Zheng@asu.edu

Dear Yi Zheng:

On 4/13/2020 the ASU IRB reviewed the following protocol:

| | |
|---|---|
| Type of Review: | Modification / Update |
| Title: | Developing a Technology-Enhanced Solution to Language Inequality<br>in English-Based Math Tests |
| Investigator: | Yi Zheng |
| IRB ID: | STUDY00007933 |
| Funding: | None |
| Grant Title: | None |
| Grant ID: | None |
| Documents Reviewed: | • citiCompletionCertificate_Zheng_2019.pdf, Category: Other;<br>• Email Recruitment_Phase3_Interviews.pdf, Category: Recruitment Materials;<br>• HRP-502a_consent_form_V7_Phase2_Interview.pdf, Category: Consent Form;<br>• HRP-502a_consent_form_V7_Phase3_Interview.pdf, Category: Consent Form;<br>• HRP-502a_consent_form_V7_Phase3_test.pdf, Category: Consent Form;<br>• HRP-503a_V8.docx, Category: IRB Protocol;<br>• My ASU Recruitment_Phase2_Interviews.pdf, Category: Recruitment Materials;<br>• My ASU Recruitment_Phase3_Tests.pdf, Category: Recruitment Materials; |

# APPENDIX H

# PHASE TWO INTERVIEW RECRUITMENT VIA ONLINE ADVERTISEMENT

Title: Native Spanish or Native Chinese speakers needed for study

Would you like to receive $40 for a few hours by contributing to research on making tests fairer?

Dr. Yi Zheng, professor in the Mary Lou Fulton Teachers College at ASU, is conducting a research study to examine language bias in math tests.

To participate in this study, you must:
1) identify as a native Mandarin Chinese speaker or a native Spanish speaker,
2) identify yourself as a non-native English speaker
3) be a prospective first year, sophomore, or junior student
4) have taken a math class in the last year or currently taking a math class
(at ASU or elsewhere),
5) be a current ASU student
6) be 18 years or older

Do not contact us if you do not qualify for ALL of the above requirements. You will participate completely online using Zoom, for about two hours. Your participation is voluntary, and data will be made anonymous. If you participate, you will be paid $40. If you are interested in participating or have any questions concerning the research study, please email research assistant, Kevin Close.

# APPENDIX I

# PHASE TWO INTERVIEW CONSENT FORM

**Title of research study:** Developing a technology-enriched solution to language inequality in English-based math tests

**Investigator:** Dr. Yi Zheng

### Why am I being invited to take part in a research study?

We invite you to take part in a research study because you are a student whose first language is not English.

### Why is this research being done?

This research is done to evaluate ways to make tests fairer for student who speak languages other than English.

### How long will the research last and will there be compensation?

We expect that individuals will spend one to two hours participating in the proposed activities. The compensation will be $40.

### How many people will be studied?

We expect about 200 people will participate in this research study overall.

### What happens if I say yes, I want to be in this research?

You are free to decide whether you wish to participate in this study. The interview will be audio recorded, and the audio will be stored securely.

### What happens if I say yes, but I change my mind later?

You can leave the research at any time it will not be held against you. Your data will then be deleted.

### Is there any way being in this study could be bad for me?

Risk of harm associated with this study is negligible.

### Will being in this study help me in any way?

We cannot promise any benefits to you or others from your taking part in this research. However, possible benefits include advancing knowledge about how to make tests fairer for student who speak languages other than English.

### What happens to the information collected for the research?

Efforts will be made to limit the use and disclosure of your personal information, including research study records, to people who have a need to review this information. Your name and other identifying information will be linked to anonymous user ID on a master list, then that user ID will be used when we analyze the data. This master list will be kept securely apart from the other data and held for 5 only years which makes your information more secure. The results of this study may be used in reports, presentations, or publications but your name will not be used.

### Who can I talk to?

If you have questions, concerns, or complaints, talk to the research team by emailing Kevin Close, kevin.close@asu.edu, or Yi Zheng, yi.isabel.zheng@asu.edu.

This research has been reviewed and approved by the Social Behavioral IRB. You may talk to them at (480) 965-6788 or by email at research.integrity@asu.edu if:

- Your questions, concerns, or complaints are not being answered by the research team.
  - You cannot reach the research team.
  - You want to talk to someone besides the research team.
  - You have questions about your rights as a research participant.
  - You want to get information or provide input about this research.

Signature: Date:

I consent, continue

# APPENDIX J

# PHASE TWO SCRIPT AND QUESTIONS

## Script

Thank you for agreeing to participate in this study. To begin, bear with me as a read this script word for word. The study should take about two hours. The first part of the study will be to ask you some questions about yourself. Feel free to skip any question you do not feel comfortable answering. Additionally, be aware that I will be recording or conversations the whole time.

After answering those questions, I will give you a calculus test to take. However, as you take this test, you will be talking with me about your answers. Think about this as an interesting type of interview. During the process, I want you to talk about what you are thinking. I will also be asking you questions about your thinking.

Remember, speaking while answering questions is hard. Just do your best and remember that all of this helps us better understand how to make better tests for non-native English speakers in the future.

Thank you again!

## Interview Questions

There are no official questions. Probing questions could include:

1. Why did you do that?
2. How confident are you in that answer?
3. How did you remember to do that?
4. Tell me more about that thought?

# APPENDIX K

# SPANISH TRANSLATION OF PCA

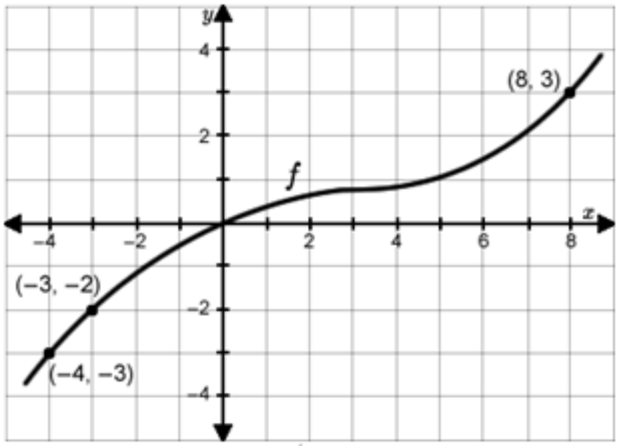
# EPC – Examen de Pre-cálculo

1) Dada la función *f*, definida por la fórmula $f(x) = 3x^2 + 2x - 4$, resuelva $f(x+a)$.

f) $f(x+a) = 3x^2 + 3a^2 + 2x + 2a - 4$
g) $f(x+a) = 3x^2 + 6xa + 3a^2 + 2x - 4$
h) $f(x+a) = 3(x+a)^2 + 2(x+a) - 4$
i) $f(x+a) = 3(x+a)^2 + 2x - 4$
j) $f(x) = 3x^2 + 2x - 4 + a$

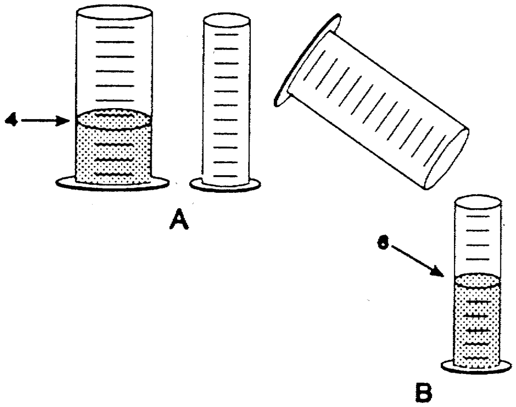
2) Utilice la gráfica lineal de *f* para resolver $f(x) = -3$ y obtener el valor de *x*.

f) (–3,–2)
g) –4
h) (–4,–3)
i) –2
j) –3



3) A la derecha se muestran imágenes de un cilindro ancho y de uno estrecho. Ambos cilindros presentan marcas equidistantes. Se vierte agua en el cilindro ancho hasta alcanzar la 4ª marca (ver A). Esta misma cantidad de agua alcanza la 6ª marca cuando se la coloca en el cilindro estrecho (ver B).

Ambos cilindros se vacían y el agua se vierte en el cilindro estrecho alcanzando la 11ª marca. ¿Hasta dónde subirá el agua si la vertemos en el cilindro ancho vacío?



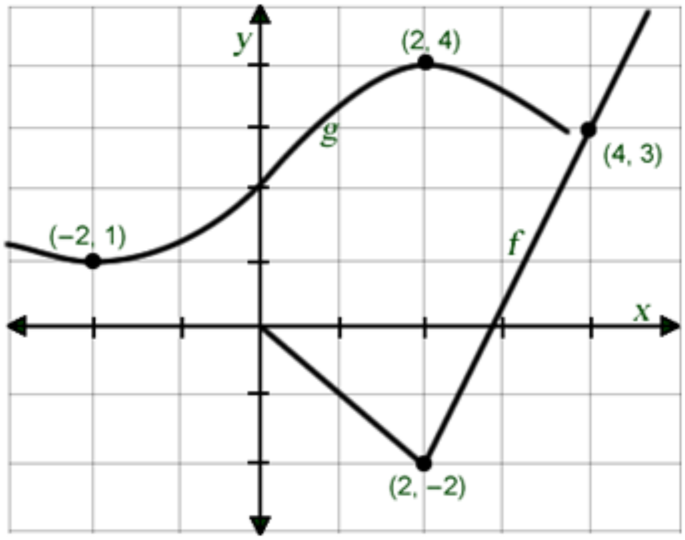
f) Hasta la marca 7 ½
g) Hasta la 9ª marca
h) Hasta la 8ª marca
i) Hasta la marca 7 ⅓
j) Hasta la 11ª marca

4) ¿Cuál de las siguientes fórmulas define el área ($A$) de un cuadrado en función de su perímetro ($p$)?

a) $A = \frac{p^2}{16}$
b) $A = s^2$
c) $A = \frac{p^2}{4}$
d) $A = 16s^2$
e) $p = 4\sqrt{A}$

Responda los puntos 5 y 6 usando las gráficas lineales de $f$ y $g$.

5) Utilice las gráficas lineales de $f$ y $g$ para calcular $g\left(f(2)\right)$.



f) -2
g) 1
h) 3
i) 4
j) No está definido

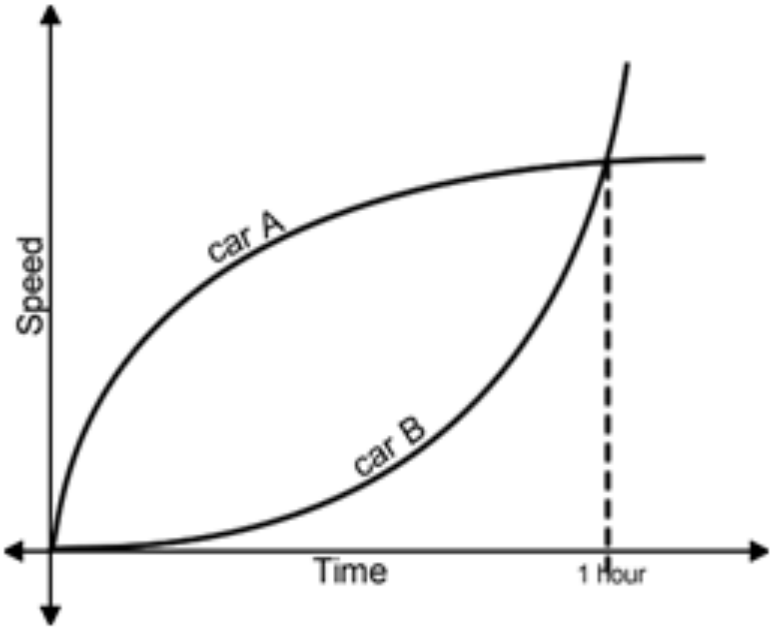
6) Calcule $f(2) - g(0)$.

f) -4
g) -2
h) 0
i) 2
j) 4

7) El modelo que describe la cantidad de bacterias en un cultivo después de *t* días acaba de actualizarse de $P(t) = 7(2)^t$ a $P(t) = 7(3)^t$. ¿Qué conclusión puede obtener de esta información?

f) El número final de bacterias es 3 veces su valor inicial en lugar de 2 veces.
g) El número inicial de bacterias es 3 en lugar de 2.
h) El número de bacterias se triplica todos los días en vez de duplicarse.
i) La tasa de crecimiento de las bacterias en el cultivo es del 30% diario en lugar del 20%.
j) Ninguna de las anteriores.

El siguiente gráfico representa la velocidad vs. el tiempo entre dos automóviles, suponiendo que arrancan desde la misma posición y viajan hacia la misma dirección. Use esta información y el gráfico a continuación para responder el punto 8.

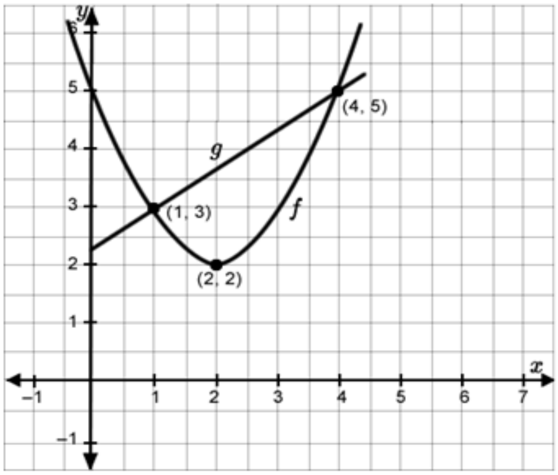


8) ¿Cuál es la relación entre la ***posición*** del automóvil A y la del automóvil B si $t = 1$ hora?

   f) El automóvil A y el automóvil B colisionan.
   g) El automóvil A está delante del automóvil B.
   h) El automóvil B está delante del automóvil A.
   i) El automóvil B pasa al automóvil A.
   j) Los automóviles están en la misma posición.

9) Utilice las gráficas lineales de $f$ y $g$ para resolver $g(x) > f(x)$.

   f) $2 < x < 5$
   g) $1 < y < 4$
   h) $x < 4$
   i) $2 < y < 5$
   j) $1 < x < 4$



10) Se utiliza una manguera para llenar una piscina vacía. El gráfico muestra el volumen (en galones) en la piscina en función del tiempo (en minutos). ¿Cuál de las siguientes opciones representa la fórmula para calcular el tiempo ($t$) como una función del volumen ($v$)?

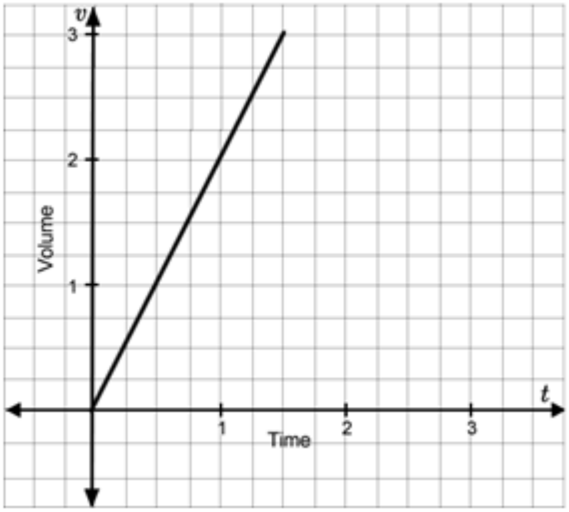
f) $v(t) = \dfrac{t}{2}$

g) $t(v) = 2v$

h) $t(v) = \dfrac{v}{2}$

i) $v(t) = 2t$

j) Ninguna de las anteriores



11) La distancia $s$ (en pies) recorrida por un automóvil que se mueve en línea recta está dada por la función $s(t) = t^2 + t$, donde $t$ se mide en segundos. Encuentre la velocidad promedio del período de tiempo que va de $t = 1$ a $t = 4$.

a) 5 $^{ft}\!/_{sec}$

b) 6 $^{ft}\!/_{sec}$

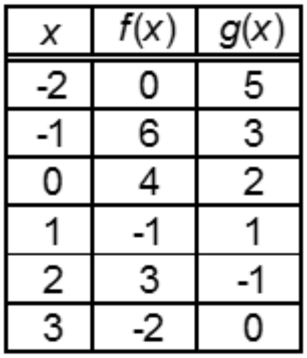
c) 9 $^{ft}\!/_{sec}$

d) 10 $^{ft}\!/_{sec}$

e) 11 $^{ft}\!/_{sec}$

12) Usando la tabla a la derecha, determine $f\left(g\left(3\right)\right)$.

f) 4
g) -1
h) 0
i) 1
j) 5



| $x$ | $f(x)$ | $g(x)$ |
|---|---|---|
| -2 | 0 | 5 |
| -1 | 6 | 3 |
| 0 | 4 | 2 |
| 1 | -1 | 1 |
| 2 | 3 | -1 |
| 3 | -2 | 0 |

13) Usando la tabla a la derecha, determine $g^{-1}(-1)$.

f) ½
g) ⅓
h) 1
i) 2
j) 3

14) Dado que *f* se define como $f(t)=100^t$, ¿cuál de las siguientes opciones representa la fórmula para calcular *f* -1?

f) $f^{-1}(t)=\dfrac{1}{100^t}$

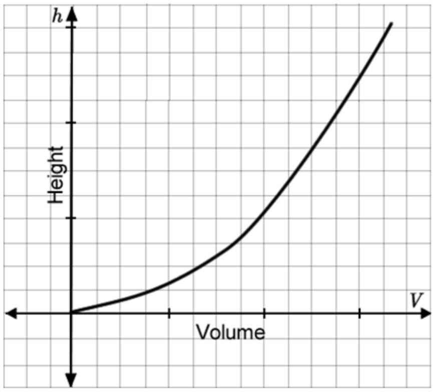
g) $f^{-1}(t)=\dfrac{t}{\ln 100}$

h) $f^{-1}(t)=\dfrac{t}{100^t}$

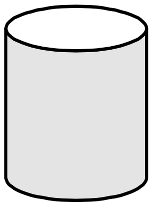
i) $f^{-1}(t)=\dfrac{\ln t}{\ln 100}$

j) $f^{-1}(t)=\dfrac{\ln t}{100}$

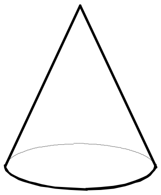
15) El siguiente gráfico representa la altura del agua como una función del volumen a medida que se vierte el agua en un recipiente. ¿Cuál recipiente está representado en esta gráfica?



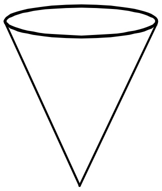
a)



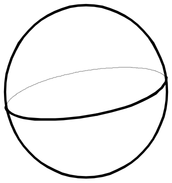
b)



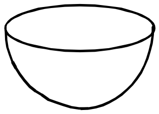
c)



d)



e)



16) Dada la función $h(x) = 3x - 1$ y $g(x) = x^2$, calcule $g(h(2))$ .

f) 10
g) 11
h) 20
i) 25
j) 36

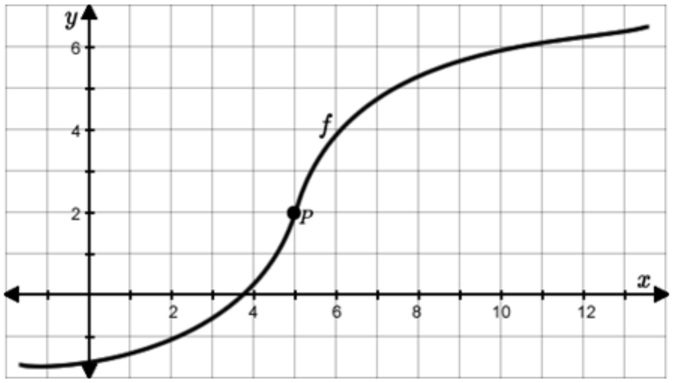
17) Se arroja una pelota a un lago generando una onda circular que se desplaza hacia afuera a una velocidad de 5 cm por segundo. Especifique el área (*A*) del círculo en función de la cantidad de segundos (*t*) que pasaron desde que la pelota tocó el lago.

f) $A = 25\pi t$
g) $A = \pi r^2$
h) $A = 25\pi t^2$
i) $A = 5\pi t^2$
j) Ninguna de las anteriores

18) La comisión de caza de animales silvestres vertió 5 latas de pescado (cada una contenía aproximadamente 100 peces) en el lago de un granjero. La función *N* definida por $N(t) = \dfrac{600t + 500}{0.5t + 1}$ representa la cantidad aproximada de peces en el lago en función del tiempo (en años). ¿Cuál de las siguientes opciones describe mejor el cambio en el número de peces en el lago a lo largo del tiempo?

f) El número de peces aumenta cada año, pero no supera los 500.
g) El número de peces aumenta cada año, pero no supera los 1200.
h) El número de peces disminuye cada año, pero no lo hace a menos de 500.
i) El número de peces aumenta cada año, pero no supera los 600.
j) El número de peces se disminuye cada año, pero no lo hace a menos de 1200.

19) Usando el siguiente gráfico, explique el comportamiento de la función *f* en el intervalo comprendido entre $x = 5$ a $x = 12$.



f) Aumenta a un ritmo creciente.
g) Aumenta a un ritmo decreciente.
h) Aumenta a un ritmo constante.
i) Disminuye a un ritmo decreciente.
j) Disminuye a un ritmo creciente.

20) Si *S*(*m*) representa el salario (mensual), en cientos de dólares, de un empleado luego de *m* meses en el trabajo, ¿qué representaría la función *R* (*m*) = *S* (*m* + 12)?

a) El salario de un empleado luego de *m* + 12 meses en el trabajo
b) El salario de un empleado luego de 12 meses en el trabajo
c) $12 más que el salario de alguien que trabajó durante *m* meses
d) Un empleado que trabajó durante *m* + 12 meses
e) No hay información suficiente.

21) ¿Cuál es el dominio de la siguiente función: $f(x)=\dfrac{\sqrt{x+2}}{x-1}$ ?

f) $(1,\infty)$
g) $x \neq 1$
h) $[-2,1)\cup(1,\infty)$
i) $[-2,\infty)$
j) Todos los números reales

22) Una tarjeta de béisbol aumenta su valor según la función $b\ (t) = \frac{5}{2}\ t + 100$, donde *b* representa el valor de la tarjeta en dólares y *t* es el tiempo (en años) desde que se adquirió la tarjeta. ¿Cuál de las siguientes opciones describe lo que $\frac{5}{2}$ refleja sobre la situación?

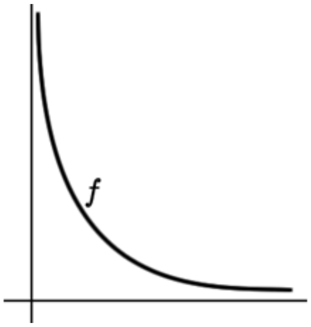
I. El valor de la tarjeta aumenta $5 cada dos años.
II. Cada año el valor de la tarjeta es 2,5 veces mayor que el año anterior.
III. El valor de la tarjeta aumenta en $\frac{5}{2}$ dólares cada año.

f) La opción I únicamente
g) La opción II únicamente
h) La opción III únicamente
i) I y III únicamente
j) I, II y III

23) ¿Cuál de las siguientes opciones describe el efecto de $f^{-1}$, cuando *f* es una función individual y $f(d) = c$?

f) $f^{-1}$ invierte *f*, entonces $f^{-1}(d) = \frac{1}{f(d)}$
g) $f^{-1}$ invierte la entrada a *f*, entonces $f^{-1}(d) = \frac{1}{d}$
h) $f^{-1}$ invierte la salida a *f*, entonces $f^{-1}(d) = \frac{1}{c}$
i) $f^{-1}$ invierte *f*, entonces $f^{-1}(f(d)) = d$
j) a y c

24) Definimos una función *f* mediante el siguiente gráfico. ¿Cuál de las siguientes opciones describe el comportamiento de *f*?



I. A medida que el valor de $x$ se aproxima a 0, el valor de $f$ aumenta.
II. A medida que el valor de $x$ aumenta, el valor de $f$ se aproxima a 0.
III. A medida que el valor de $x$ se aproxima a 0, el valor de $f$ se aproxima a 0.

a) La opción I únicamente
b) La opción II únicamente
c) La opción III únicamente
d) I y II
e) II y III

25) ¿Cuál de las siguientes opciones describe mejor el comportamiento de la función $f$ definida por $f(x)=\frac{x^2}{x-2}$?

I. A medida que el valor de $x$ aumenta, el valor de $f$ se aproxima a 2.
II. A medida que el valor de $x$ aumenta, el valor de $f$ aumenta.
III. A medida que el valor de $x$ se aproxima a 2, el valor de $f$ se aproxima a 0.

f) La opción I únicamente
g) La opción II únicamente
h) La opción III únicamente
i) I y III
j) II y III

# APPENDIX L

# MANDARIN CHINESE TRANSLATION OF PCA

1) 已知函数 $f(x) = 3x^2 + 2x - 4$, 以下哪个选项等价于$f(x+a)$?

k) $f(x+a) = 3x^2 + 3a^2 + 2x + 2a - 4$

l) $f(x+a) = 3x^2 + 6xa + 3a^2 + 2x - 4$

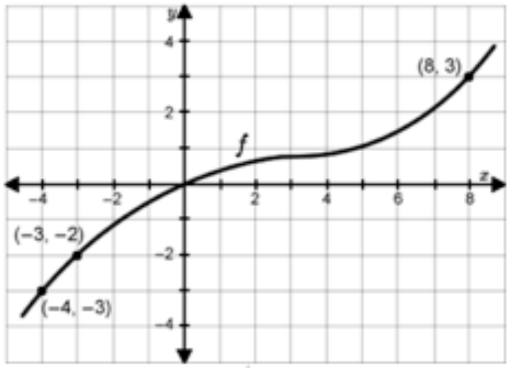
m) $f(x+a) = 3(x+a)^2 + 2(x+a) - 4$

n) $f(x+a) = 3(x+a)^2 + 2x - 4$

o) $f(x) = 3x^2 + 2x - 4 + a$

2) 请用右图中函数f的曲线来求解f(x) = −3时的x值。

k) (–3,–2)
l) –4
m)(–4,–3)
n) –2
o) –3



3）右图中有一个宽的圆柱体容器和一个窄的圆柱体容器。图中的圆柱体容器上刻有等距离的标记。如图A所示，将水注入到宽的圆柱体容器中直到水位上升到第4个标记为止。然后将宽的圆柱体容器中的全部水倒入到窄的圆柱体容器中时，水位上升到第6个标记 （如图B所示）。

将两个圆柱体容器倒空，然后将水注入到窄的圆柱体容器中直到水位上升到第11个标记为止。若将窄的圆柱体容器中的全部水倒入到空的宽圆柱体容器中，水位将上升到宽圆柱体容器中的哪个位置上‘

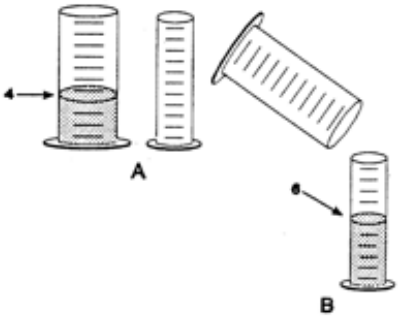
a) 第7 ½ 个标记

b) 第9个标记

c) 第8 个标记

d) 第7 $^{1}/_{3}$个标记

e) 第11个标记



4）以下哪个选项正确地以一个正方形的周长$p$定义该正方形的面积$A$？

a) $A = \frac{p^2}{16}$

b) $A = s^2$

c) $A = \frac{p^2}{4}$

d) $A = 16s^2$

e) $p = 4\sqrt{A}$

请用右图中函数f 和函数 g 的曲线来回答以下第5题和第6题。

5）请用图中函数f 和函数 g 的曲线来求解 g(f(2)).

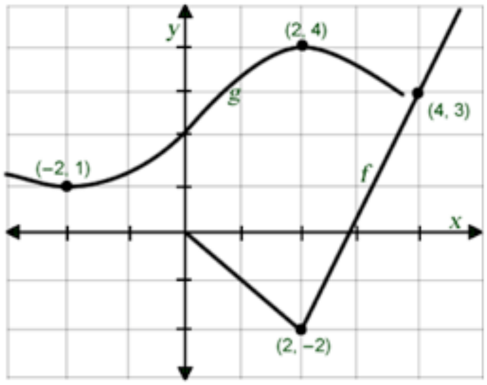
k) -2

l) 1

m) 3

n) 4

o) 未定义的



6）求解 $f(2) - g(0)$.

p) -4

q) -2

r) 0

s) 2

t) 4

7)用来描述细菌在培养皿中t天之后的细菌数量的函数模型从 $P(t) = 7(2)^t$ 被更新到 $P(t) = 7(3)^t$。根据这个信息，以下哪一个陈述是正确的？

a) 培养皿中最终的细菌数量是原始细菌数量的3倍，而不是原来的2倍。

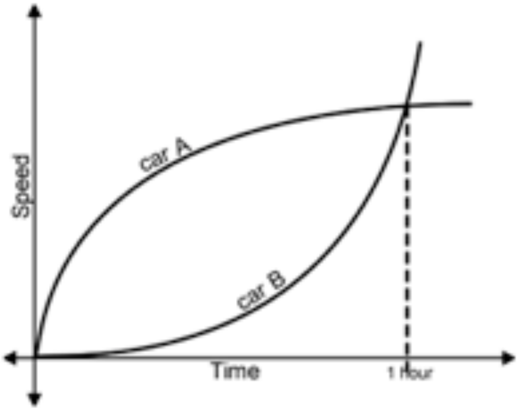
b) 培养皿中原始的细菌数量是3个而不是2个。

c) 培养皿中细菌数量每天翻3倍而不是每天翻两倍。

d) 培养皿中的细菌的增长速率是每天增长30%， 而不是每天增长20%

e)以上陈述均不正确

下图描绘了车A 和车 B 两辆车行驶时间和行驶速度的关系（假设两辆车均从同一个地点出发并驶向同一个方向）请用以下信息和图中曲线来回答第八题。



8）在时间t=1小时的时候，两辆车的位置关系是什么样子的？

a) 车A和车B正在相撞

b) 车A在车B的前方

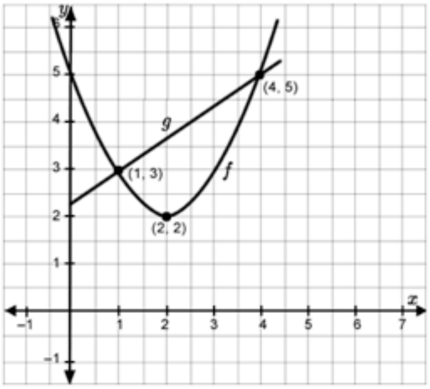
c) 车B 在车 A 的前方

c) 车B 正在超越车 A

d) 两辆车在相同的位置

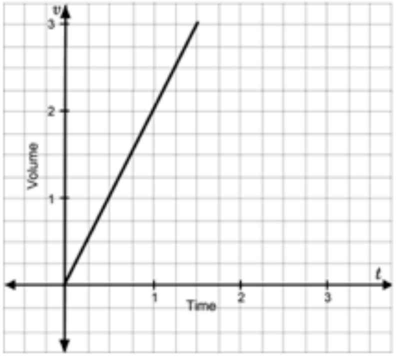
9） 请用下图中函数 $f$ 和函数 $g$ 的曲线来求解 $\mathrm{g(x)} > f(x)$.

k) $2 < x < 5$
l) $1 < y < 4$
m) $x < 4$
n) $2 < y < 5$
o) $1 < x < 4$



10) 用一个水管来住满一个空的水池。下图描绘了注水时间（分钟）和水量（加仑）的关系式。以下哪个选项用水量v来定义计算时间t的公式？

a) $v(t) = \frac{t}{2}$
b) $t(v) = 2v$
c) $t(v) = \frac{v}{2}$
d) $v(t) = 2t$
e)以上选项均不正确

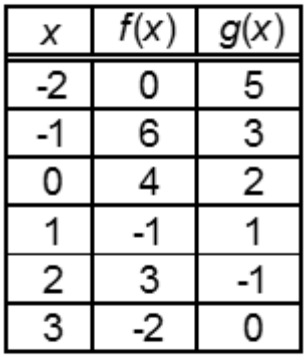


11）已知函数$s(t) = t^2 + t$描述了一辆车直线行驶时的距离s（英尺）和时间t（秒钟）的关系。请计算车在时间段从 t = 1 到 t = 4 的平均行驶速度.

a) 5 $^{\text{英尺}}/_{\text{秒}}$
b) 6 $^{\text{英尺}}/_{\text{秒}}$
c) 9 $^{\text{英尺}}/_{\text{秒}}$
d) 10 $^{\text{英尺}}/_{\text{秒}}$
e) 11 $^{\text{英尺}}/_{\text{秒}}$

12) 请用右图中的函数关系来求解 $f(g(3))$.

k) 4
l) -1
m) 0
n) 1
o) 5



| $x$ | $f(x)$ | $g(x)$ |
|---|---|---|
| -2 | 0 | 5 |
| -1 | 6 | 3 |
| 0 | 4 | 2 |
| 1 | -1 | 1 |
| 2 | 3 | -1 |
| 3 | -2 | 0 |

13) 请用右图中的函数关系来求解 $g^{-1}(-1)$.

k) ½
l) ⅓
m) 1
n) 2
o) 3

14)已知函数 f(t) = $100^t$, 以下哪个公式正确表达了 $f^{-1}$ ?

k) $f^{-1}(t) = \dfrac{1}{100^t}$

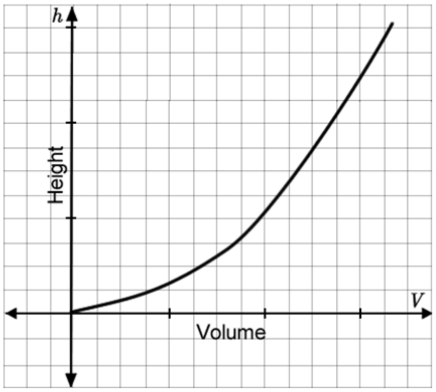
l) $f^{-1}(t) = \dfrac{t}{\ln 100}$

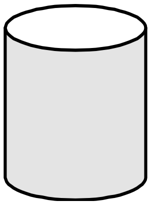
m) $f^{-1}(t) = \dfrac{t}{100^t}$

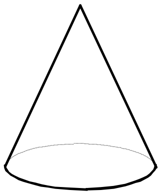
n) $f^{-1}(t) = \dfrac{\ln t}{\ln 100}$

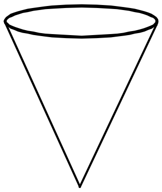
o) $f^{-1}(t) = \dfrac{\ln t}{100}$

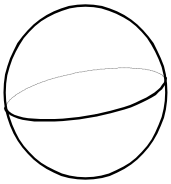
15)下图描述了将水逐渐注入一个未知形状的容器时，水的高度和水的体积之间的关系。请问以下哪个形状的容器符合图中描绘的关系？

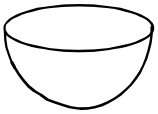


a) 

b) 

c) 

d) 

e) 

16）已知函数 $h(x) = 3x - 1$ and $g(x) = x^2$ , 求解 $g(h(2))$.

k) 10
l) 11
m) 20
n) 25
o) 36

17) 当小球被扔入一条河中的时候，河面上会产生以每秒钟5厘米的速度向外围扩散的圆形波纹。请用自小球击中湖面的瞬间开始所经过的时间t（秒钟）来表达圆形波纹区域内的面积A。

a) $A = 25\pi t$
b) $A = \pi t^2$
c) $A = 25\pi t^2$
d) $A = 5\pi t^2$
e)以上表达式均不正确

18）野生动物狩猎协会向农民的湖中倾倒了5罐鱼 （每罐大约装载了100条鱼）。函数N(t) = $\frac{600t+500}{0.5t+1}$ 以时间t（年）的方程代表了湖中鱼的大概数量。以下哪个陈述最恰当地描述了湖中鱼的数量如何随时间变化？

a) 湖中鱼的数量每年都在变大，但不会超过500条。

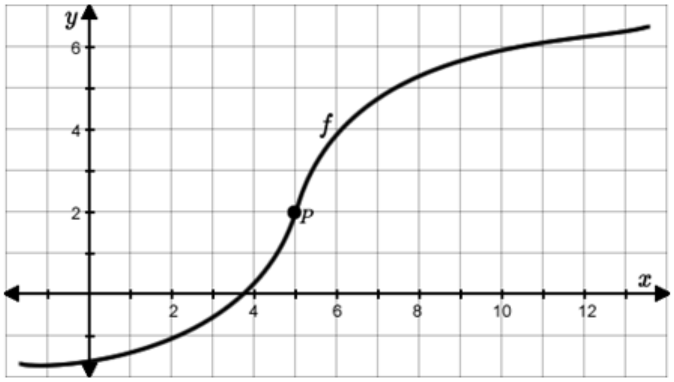
b) 湖中鱼的数量每年都在变大，但不会超过 1200条。

c) 湖中鱼的数量每年都在变小，但不会小于500条。

d) 湖中鱼的数量每年都在变大，但不会超过600条。

e) 湖中鱼的数量每年都在变小，但不会小于1200条。

19）如图，以下哪个选项最恰当地描述了函数 $f$ 在定义域区间$\mathrm{x} = 5$ 到 $x = 12$ 的变化趋势。



a) 以逐渐加快的速度增大

b) 以逐渐减慢的速度增大

c) 以不变的速度逐渐增大

d) 以逐渐减慢的速度减小

e) 以逐渐加快的速度减小

20）已知函数 $\mathrm{S(m)}$ 代表了一个员工在工作了m月后的工资（几百美元／每月），那么函数 $R(m) = S(m + 12)$ 代表什么？

a) 一个员工在工作了 $m + 12$ 个月之后的工资。

b) 一个员工在工作了12个月之后的工资。

c) 比一个工作了m月的员工的工资高12美元。

d) 一个工作了 $m + 12$ 个月的员工。

e) 给出的信息不足

21）求已知函数 $f(x) = \frac{\sqrt{x+2}}{x-1}$ 的定义域？

a) $(1, \infty)$
b) $x \neq 1$
c) $[-2,1) \cup (1, \infty)$
d) $[-2, \infty)$
e) 全部实数

22）一个棒球卡的价值$b$ (美元) 从该卡被购买后随时间 $t$ (年) 按函数 $b(t) = \frac{5}{2}t + 100$ 来逐渐增值。以下哪个陈述正确描述了 $\frac{5}{2}$ 在此函数中代表的含义？

I. 该卡的价值每两年增加5美元。

II. 每年该卡的价值比前一年大2.5倍。

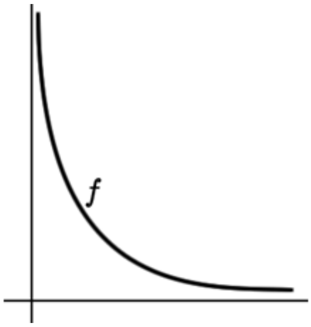
III. 该卡的价值每年增加 $\frac{5}{2}$ 美元。

a) 只有Ⅰ
b) 只有Ⅱ
c) 只有Ⅲ
d) 只有Ⅰ和Ⅲ
e) Ⅰ，Ⅱ，和Ⅲ

23）已知函数 $f$ 是一个双射函数（也称一一对应）且 $f(d) = c$, 以下哪个选项最佳地描述了 $f^{-1}$?

a) $f^{-1}$ 倒置函数 $f$ ，所以 $f^{-1}(d) = \frac{1}{f(d)}$
b) $f^{-1}$ 倒置函数 $f$ 定义域的值，所以 $f^{-1}(d) = \frac{1}{d}$
c) $f^{-1}$ 倒置函数 $f$ 的值域的值，所以 $f^{-1}(d) = \frac{1}{c}$
d) $f^{-1}$ 倒置函数 $f$ ，所以 $f^{-1}(f(d)) = d$
e) 选项a 和 c

24) 如图描绘了函数 $f$. 以下哪个陈述正确描述了函数 $f$ 的趋势？



I. 当x 的值趋近0的时候，函数 $f$ 的值增大

II. 当 x 的值增大时，函数 $f$ 的值趋近0

II. 当x 的值趋近0时，函数 $f$ 的值趋近0

a) 只有Ⅰ
b) 只有Ⅱ
c) 只有Ⅲ
d) Ⅰ和Ⅱ
e) Ⅱ和Ⅲ

25) 以下哪个陈述最恰当地描述了函数 $f(x) = \frac{x^2}{x-2}$。

I. 当x的值变得很大的时候，函数 $f$ 的值趋近2。

II. 当x的值变得很大的时候，函数 $f$ 的值增大。

III. 当x的值趋近2时，函数 $f$ 的值趋近0。

a) 只有Ⅰ
b) 只有Ⅱ
c) 只有Ⅲ
d) Ⅰ和Ⅲ
e) Ⅱ和Ⅲ

# APPENDIX M

# PHASE THREE RECRUITMENT VIA ONLINE ADVERTISEMENT

Title: Native Spanish or Native Chinese speakers needed for study

Would you like to receive $50 for a few hours by contributing to research on making tests fairer?

Dr. Yi Zheng, professor in the Mary Lou Fulton Teachers College at ASU, is conducting a research study to examine language bias in math tests.

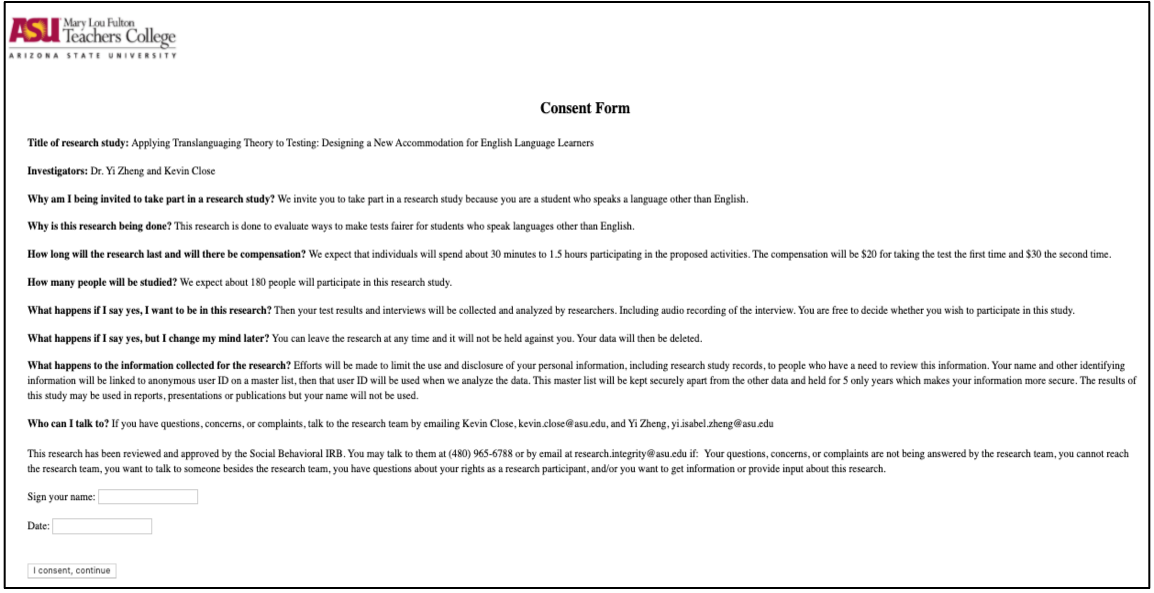
To participate in this study, you must:
1) identify as a native Mandarin Chinese speaker or a native Spanish speaker,
2) identify yourself as a non-native English speaker
3) be a prospective first year, sophomore, or junior student
4) have taken a math class in the last year or currently taking a math class
(at ASU or elsewhere),
5) be a current ASU student
6) be 18 years or older

Do not contact us if you do not qualify for ALL of the above requirements. You will participate completely online, for about an hour, two times over the course of a month (so two hours total). Your participation is voluntary, and data will be made anonymous. If you participate, you will be paid $50 ($20 for the first hour and $30 for the second hour). If you are interested in participating or have any questions concerning the research study, please email research assistant, Kevin Close.

# APPENDIX N

# PHASE THREE TASK EXPLANATION

Phase three is a test given online. The first page asks for consent and has an area for a signature and a date. It looks like this (Text sized to fit one screenshot):



**Consent Form**

**Title of research study:** Applying Translanguaging Theory to Testing: Designing a New Accommodation for English Language Learners

**Investigators:** Dr. Yi Zheng and Kevin Close

**Why am I being invited to take part in a research study?** We invite you to take part in a research study because you are a student who speaks a language other than English.

**Why is this research being done?** This research is done to evaluate ways to make tests fairer for students who speak languages other than English.

**How long will the research last and will there be compensation?** We expect that individuals will spend about 30 minutes to 1.5 hours participating in the proposed activities. The compensation will be $20 for taking the test the first time and $30 the second time.

**How many people will be studied?** We expect about 180 people will participate in this research study.

**What happens if I say yes, I want to be in this research?** Then your test results and interviews will be collected and analyzed by researchers. Including audio recording of the interview. You are free to decide whether you wish to participate in this study.

**What happens if I say yes, but I change my mind later?** You can leave the research at any time and it will not be held against you. Your data will then be deleted.

**What happens to the information collected for the research?** Efforts will be made to limit the use and disclosure of your personal information, including research study records, to people who have a need to review this information. Your name and other identifying information will be linked to anonymous user ID on a master list, then that user ID will be used when we analyze the data. This master list will be kept securely apart from the other data and held for 5 only years which makes your information more secure. The results of this study may be used in reports, presentations or publications but your name will not be used.

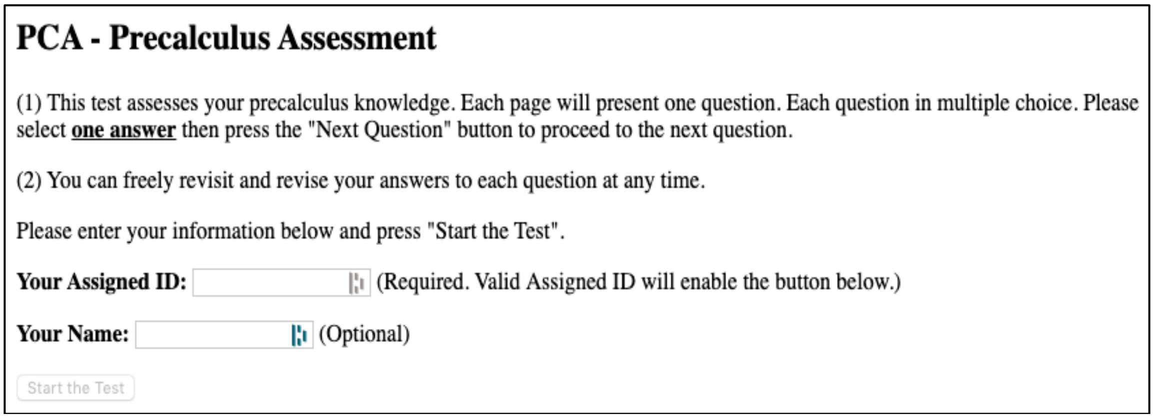
**Who can I talk to?** If you have questions, concerns, or complaints, talk to the research team by emailing Kevin Close, kevin.close@asu.edu, and Yi Zheng, yi.isabel.zheng@asu.edu

This research has been reviewed and approved by the Social Behavioral IRB. You may talk to them at (480) 965-6788 or by email at research.integrity@asu.edu if: Your questions, concerns, or complaints are not being answered by the research team, you cannot reach the research team, you want to talk to someone besides the research team, you have questions about your rights as a research participant, and/or you want to get information or provide input about this research.

Sign your name:

Date:

Page 2 looks like this:



**PCA - Precalculus Assessment**

(1) This test assesses your precalculus knowledge. Each page will present one question. Each question in multiple choice. Please select **one answer** then press the "Next Question" button to proceed to the next question.

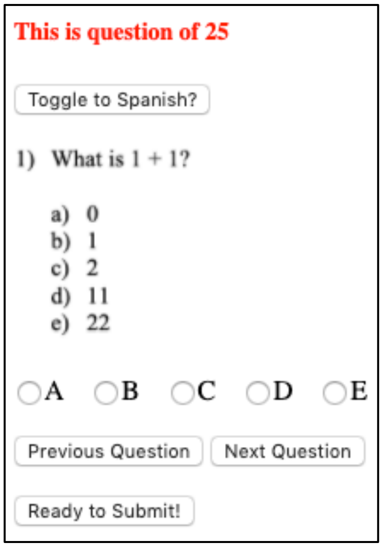
(2) You can freely revisit and revise your answers to each question at any time.

Please enter your information below and press "Start the Test".

**Your Assigned ID:** (Required. Valid Assigned ID will enable the button below.)

**Your Name:** (Optional)

Pages 3 and beyond looks like this:



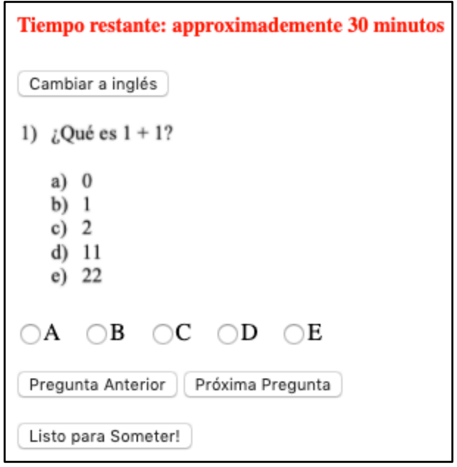
This is question of 25

1) What is 1 + 1?

a) 0
b) 1
c) 2
d) 11
e) 22

○A ○B ○C ○D ○E

Or students can toggle to another language and it will look like this:



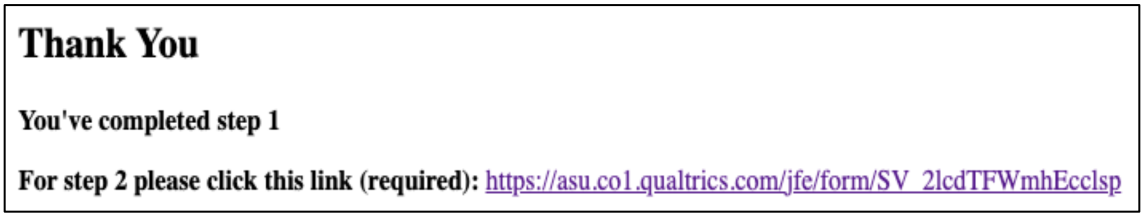
Tiempo restante: approximademente 30 minutos

1) ¿Qué es 1 + 1?

a) 0
b) 1
c) 2
d) 11
e) 22

○A ○B ○C ○D ○E

This question is the example question. There are 25 official questions that follow (there is a Spanish and a Mandarin Chinese version). The last page looks like this:



**Thank You**

**You've completed step 1**

**For step 2 please click this link (required):** https://asu.co1.qualtrics.com/jfe/form/SV_2lcdTFWmhEcclsp

The link leads to the demographics questions (See Appendix D).

# APPENDIX O

# PHASE THREE CONSENT FORM

**Title of research study:** Developing a technology-enriched solution to language inequality in English-based math tests

**Investigator:** Dr. Yi Zheng

## Why am I being invited to take part in a research study?

We invite you to take part in a research study because you are a student whose first language is not English.

## Why is this research being done?

This research is done to evaluate ways to make tests fairer for student who speak languages other than English.

## How long will the research last and will there be compensation?

We expect that individuals will spend one to two hours participating in the proposed activities. The compensation will be $20 for taking the test the first time and $30 the second time.

## How many people will be studied?

We expect about 200 people will participate in this research study overall.

## What happens if I say yes, I want to be in this research?

You will take a math test on your own computer at your preferred location. Your computer needs to be connected to the internet.

## What happens if I say yes, but I change my mind later?

You can leave the research at any time it will not be held against you. Your data will then be deleted.

## Is there any way being in this study could be bad for me?

Risk of harm associated with this study is negligible.

## Will being in this study help me in any way?

We cannot promise any benefits to you or others from your taking part in this research. However, possible benefits include advancing knowledge about how to make tests fairer for student who speak languages other than English.

## What happens to the information collected for the research?

Efforts will be made to limit the use and disclosure of your personal information, including research study records, to people who have a need to review this information. Your name and other identifying information will be linked to anonymous user ID on a master list, then that user ID will be used when we analyze the data. This master list will be kept securely apart from the other data and held for 5 only years which makes your information more secure. The results of this study may be used in reports, presentations, or publications but your name will not be used.

### Who can I talk to?

If you have questions, concerns, or complaints, talk to the research team by emailing Kevin Close, kevin.close@asu.edu, or Yi Zheng, yi.isabel.zheng@asu.edu.

This research has been reviewed and approved by the Social Behavioral IRB. You may talk to them at (480) 965-6788 or by email at research.integrity@asu.edu if:

- Your questions, concerns, or complaints are not being answered by the research team.
  - You cannot reach the research team.
  - You want to talk to someone besides the research team.
  - You have questions about your rights as a research participant.
  - You want to get information or provide input about this research.

Signature: Date:

I consent, continue

# APPENDIX P

# PHASE THREE INTERVIEW RECRUITMENT VIA EMAIL

_________,

Thank you so much for participating in our research project on making tests fairer. You have been selected for a follow up interview about your experience taking the tests. This is completely voluntary, but we can offer another $20 and the interview should take less than an hour. The interview will be held on Zoom. Let me know if you have any questions.

Thank you,
Kevin Close

# APPENDIX Q

# PHASE THREE SEMI-STRUCTURED INTERVIEW INSTRUMENT

Script:

Welcome, how are you today?

Before we get started, I want to ask your permission to record, may I record our interview?

START RECORDING

Great, this is: [state their ID], bear with me for a second, as I read this script:

You've already signed a consent form to take part in the previous phases of this study, so you know what the study is about (improving the fairness of standardized tests). However, I want to go over a few things just to emphasize.

1. This interview will take less than 45 minutes. For that you will be paid $20.
2. The questions will be about the tests that you took. I'm just curious to see how you felt about them.
3. Though this interview will be recorded, the audio and video will be stored securely. The video will likely only be seen by me or possibly by my advisor. That's it. The audio may also be heard by a transcription company employee who are required to keep it confidential.
4. You'll see that I use ID numbers instead of names as much as possible. This is to keep your information as anonymous as possible. There is only one master list that links your name with you ID. This will be held separately and securely from the other data and destroyed after 5 years. The results of this study may be used in papers, results, etc, but your name will never be used.
5. If there are any questions that you don't want to answer or if you wish to leave the study entirely, just let me know.
6. If you have any concerns or complaints, you can always email me or email the ASU's IRB office
7. Lastly, during the interview, you'll see me typing notes, that's normal.

Does that all sound OK? Do you consent to the interview?

Great!

Any questions before we get started?

Main Questions:

1. How do you typically feel about math tests?
2. How do you typically feel about online tests?
3. Which of the last two tests that you took did you prefer?
4. Which do you feel was easier?
5. How did you feel about the tests you just took?
6. Did you have a favorite or least favorite question?
7. How could I make the tests better?
8. How could I make the tests fairer?

9. What language do you prefer speaking?

10. In what language did you learn math?

11. In what language do you prefer to take math tests?

Topics to Probe:

1. Anxiety (e.g., you mentioned anxiety, tell me more about what you mean?)
2. Bias or fairness (e.g., Why do you say that this test feels less biased?)
3. Motivation (e.g., you mentioned that you kept going because of this. Why?)
4. Identity (e.g., you said you think this is good for others, but not for you. Tell me what you mean? Who else would it not be good for?)

Possible Probing Questions

1. Why?
2. Can you tell me more?
3. What did you mean by that?
4. Why did you say it that way (or why did you choose that word to describe it)?
5. Give an example of test problem that made you feel that way.

Ideas adapted from:
Lovett, B. J., & Leja, A. M. (2013). Students' perceptions of testing accommodations: What we know, what we need to know, and why it matters. *Journal of Applied School Psychology*, *29*(1), 72-89.