ABSTRACT

Title of dissertation: SEEKING CULTURAL FAIRNESS IN A MEASURE OF RELATIONAL REASONING

Denis Dumas, Doctor of Philosophy, 2016

Dissertation directed by: Patricia A. Alexander
Department of Human Development and Quantitative Methodology

Relational reasoning, or the ability to identify meaningful patterns within any stream of information, is a fundamental cognitive ability associated with academic success across a variety of domains of learning and levels of schooling. However, the measurement of this construct has been historically problematic. For example, while the construct is typically described as multidimensional—including the identification of multiple types of higher-order patterns—it is most often measured in terms of a single type of pattern: analogy. For that reason, the Test of Relational Reasoning (TORR) was conceived and developed to include three other types of patterns that appear to be meaningful in the educational context: anomaly, antinomy, and antithesis. Moreover, as a way to focus on fluid relational reasoning ability, the TORR was developed to include, except for the directions, entirely visuo-spatial stimuli, which were designed to be as novel as possible for the participant. By focusing on fluid intellectual processing, the TORR was also developed to be fairly administered to undergraduate students—regardless of the particular gender, language, and ethnic groups they belong to. However, although some psychometric investigations of the TORR have been conducted, its actual fairness across those demographic groups has yet to be empirically demonstrated.

Therefore, a systematic investigation of differential-item-functioning (DIF) across demographic groups on TORR items was conducted. A large (N = 1,379) sample, representative of the University of Maryland on key demographic variables, was collected, and the resulting data was analyzed using a multi-group, multidimensional item-response theory model comparison procedure. Using this procedure, no significant DIF was found on any of the TORR items across any of the demographic groups of interest. This null finding is interpreted as evidence of the cultural-fairness of the TORR, and potential test-development choices that may have contributed to that cultural-fairness are discussed. For example, the choice to make the TORR an untimed measure, to use novel stimuli, and to avoid stereotype threat in test administration, may have contributed to its cultural-fairness. Future steps for psychometric research on the TORR, and substantive research utilizing the TORR, are also presented and discussed.

SEEKING CULTURAL FAIRNESS IN A MEASURE OF RELATIONAL REASONING

by

Denis Dumas

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2016

Advisory Committee:

Professor Patricia Alexander, Chair
Professor Kevin Dunbar
Professor Gregory Hancock
Associate Professor Jeffrey Harring
Associate Professor Linda Schmidt

Acknowledgments

As far as I know, no child grows up wanting to be an educational psychologist, and especially not an applied psychometrician. So, maybe I am betraying my much younger self (whose career goal of becoming a dinosaur veterinarian never worked out) when I say that an educational psychologist is absolutely what I do want to be. Perhaps miraculously, in the five years I have spent in the HDQM department, I’ve developed into a researcher who, although novice, is ready to contribute positively to the field of educational psychology. For that development, I have so many people to thank. First and foremost, I need to thank my PhD adviser, Dr. Patricia Alexander, whose attentive guidance molded me more, and in more ways, than any other teacher I have had. Next, Dr. Kevin Dunbar, who exposed me to a great diversity of ideas, and who afforded me the summer research time to explore them. Also Dr. Greg Hancock, who advised me through my master’s program in measurement and statistics, and gave invaluable feedback while I was on the job market. Next, to Dr. Jeff Harring, who raised my confidence in programming and whose calm demeanor helped me avoid hyperventilation on the day of the comprehensive exam. Further, to Dr. Linda Schmidt, who taught me how valuable cross-disciplinary collaboration can be. Thank you also to all of my fellow graduate student co-authors, especially those who have brainstormed with me to develop new directions for our research: Emily Grossnickle, Sophie Jablansky, and Dan McNeish. Working with you all has been a blast. Finally, to my wife Allison, who knew me long before I had heard of educational psychology, and to whom all of my work both personal and professional is dedicated. Without you, I would be homeless.

Table of Contents

List of Tables

List of Figures

CHAPTER 1:

INTRODUCTION

Statement of Problem

Relational reasoning has been characterized as the ability to discern meaningful patterns within any informational stream (Alexander & The Disciplined Reading and Learning Research Laboratory [DRLRL], 2012, Bassok, Dunbar, & Holyoak, 2012; Crone, Wendelken, Leijenhorst, Honomichl, Christoff, & Bunge, 2009; Dumas, Alexander, & Grossnickle, 2013). Moreover, this ability to detect a meaningful pattern within seemingly unrelated information, as well as to derive overarching patterns from sets of relations from different domains, is fundamental to human cognitive functioning (e.g., Krawzcyk, 2012) and learning (e.g. Richland, Zur, & Holyoak 2007). Relational reasoning has been empirically linked to academic achievement in a variety of domains such as reading (Ehri, Satlow, & Gaskins, 2009), chemistry (Bellochie & Ritchie, 2011; Trey & Khan, 2008), mathematics (DeWolf, Bassok, & Holyoak, 2015), engineering (Dumas & Schmidt, 2015), and medicine (Dumas, Alexander, Baker, Jablansky, & Dunbar, 2014). Given its significance to human learning and performance, relational reasoning and its effective measurement is of burgeoning interest to researchers in fields such as education, psychology, and neuroscience (Crone et al., 2009; Dumas et al., 2013; Krawzcyk, 2012; Kumar, Cervesato, & Gonzalez, 2014).

Because relational reasoning broadly encompasses the mapping of patterns between and among pieces of information, it can be described as a general cognitive ability that manifests differently depending on the relations within the information at hand (Chi, 2013). To account for this diversity of possible relations, Alexander and colleagues have posited at least four forms of relational reasoning: analogy, anomaly, antinomy, and antithesis (e.g., Alexander et al., 2012;

Alexander, Dumas, Grossnickle, List, & Firetto, 2015; Dumas et al., 2013). Specifically, if a higher-order pattern of similarity can be mapped between concepts, the form of relational reasoning at work is termed *analogy* (Holyoak, 2012). By comparison, if an identified relation is perceived as a digression or deviation from a typical pattern, an *anomaly* is present (Chinn & Brewer, 1993). In contrast, if two or more mutually exclusive sets can be formed among pieces of information, reasoning by *antinomy* is taking place (Dumas et al., 2014). Finally, an *antithesis* requires the reversal of salient relations to form an oppositional pattern (Sinatra & Broughton, 2011). Moreover, the pervasiveness of relational reasoning, as evidenced by its presence in a range of academic domains and cognitive tasks (Dunbar, 1995; Richland & McDonough, 2010), suggests that relational reasoning operates whether the information is linguistic, graphic, numeric, or pictorial, and whether the task is more formal or crystallized, or more novel or fluid in form. Therefore, relational reasoning may be conceptualized as a highly generalizable, multidimensional cognitive ability, with wide-ranging applicability across a variety of academic and cultural contexts.

Guiding Postulations and Measurement Trends

This conceptualization of relational reasoning logically leads to three main postulations about the creation of a measure of relational reasoning with maximum utility for research and practice: (a) a measure of relational reasoning should explicitly tap multiple forms of the construct; (b) the amount of crystallized knowledge specific to any academic domain required to respond correctly to the items should be limited, and (c), in order to assess relational reasoning ability reliably across diverse groups of participants, a measure of relational reasoning should function similarly well regardless of participants' demographic or cultural factors (e.g., gender, ethnicity).

The first postulation, explicitly tapping multiple forms of the construct, is necessary because a student could hypothetically be strong at one form of the relational reasoning (e.g., anomaly), but weak at another form (e.g., antithesis). Thus, to fully and accurately assess relational reasoning ability, it would seem essential to explicitly measure multiple manifestations, not being limited to the identification of relational similarity (e.g., analogy) but also including higher-order relations of dissimilarity (e.g., anomaly), opposition (e.g., antithesis), and even exclusivity (e.g., antinomy). Also, without measuring multiple forms of relational reasoning, there would be no way to statistically isolate individuals' overall relational reasoning capacity from their skill at a specific form of the construct, especially some form that may have potentially been embedded within educational experiences, if not directly taught (e.g., analogical reasoning). For example, if a measure of relational reasoning were comprised solely of analogies, scores pertaining to that general construct would be confounded by a participant's specific ability at analogical reasoning. However, when anomalies, antinomies, and antitheses are added to the measure, it becomes statistically possible to assess individuals' broader relational reasoning ability without the confounding effect of any one specific form.

The second postulation guiding assessment of relational reasoning in this investigation points to the need to ensure that respondents' existing knowledge in one domain (e.g., reading, mathematics, or history) does not unduly influence measurement of the target construct. Thus, it would be essential to construct a measure that either accounted for any domain-specific influences or that was sufficiently generic in its content so as to avoid such influences. Although it may be impossible to isolate relational reasoning entirely from prior knowledge or experiences, it should be the goal of an assessment to reduce those influences to whatever extent possible. To accomplish this goal, a commonly utilized test development strategy has been to construct items that incorporate mainly novel graphical arrays, rather than formally taught linguistic or symbolic systems such as the alphabet or mathematical operators (Cattell, 1940; McCallum, 2003; Naglieri & Insko, 1986).

The third postulation relates specifically to the cultural-fairness of a measure of relational reasoning. Within a certain defined population, an ideal measure would tap relational reasoning equally reliably, and with similar underlying latent structure, regardless of participants' demographic or cultural background (Tomes, 2012). For example, if a measure of relational reasoning were designed to assess the construct in university undergraduates, then that measure should operate similarly well whether those students are male or female. If this goal of measure development was not appropriately met, it would drastically limit the usefulness of the measure, because inferences drawn from test scores may not be valid for certain groups or samples drawn from its target population (Reynolds & Ramsay, 2003).

Importantly, this third postulation is conceptually linked to the second, because if a test of relational reasoning limits the amount of construct irrelevant variance from extraneous academic or cultural variables, and focuses on fluid intellectual processing, the effect of culture or demographic variables on test scores may be minimized (Borghese & Gronau, 2005; McCallum, 2003). However, evidence suggests that the construction of a fluid measure that does not explicitly tap schooled knowledge is likely a necessary but not sufficient condition for cultural fairness, because cultural background can also affect the way in which an individual reasons fluidly with novel stimuli (Nisbett, 2009; Sternberg, 2004). Therefore, the cultural fairness of a measure of relational reasoning cannot be inferred directly from its fluidity or generality, and must be conceptualized as a separate test development goal.

In a published systematic literature review, Dumas et al. (2013) noted several critical trends in the measurement of relational reasoning. For one, these researchers concluded that, despite the growing interest in and documented role of relational reasoning to learning and cognitive performance, its measurement has been historically problematic. In effect, even though the definitions of relational reasoning that populate the literature speak broadly to individuals' ability to discern patterns; the measures of this construct have focused almost exclusively on one form: analogical reasoning (e.g., Cho, Holyoak, & Cannon, 2007). Thus, the presumed multidimensional character of relational reasoning has not been well represented. In fact, no existing measure of relational reasoning has incorporated multiple forms of the construct, in the same assessment context.

Moreover, many of these measures require domain-specific knowledge and strategies that may be more emphasized in one cultural context than another, at once violating both our second and third postulations. For example, when strong domain-specific skills (e.g., reading) are a prerequisite for the discernment of relational patterns, construct irrelevant variance attributable to that ability affects the measure of relational reasoning. In sum, there is a need for a psychometrically sound assessment of relational reasoning that is not only multidimensional in form (postulation 1), but one that also reduces extraneous crystallized or culturally-specific influences (postulation 2) in order to be reliably used across all members of a given target population, regardless of their demographic background (postulation 3).

The Test of Relational Reasoning

In order to address the gaps just overviewed, the Test of Relational Reasoning (TORR) was conceived (Alexander, 2012) and developed (Alexander et al., 2015). The TORR has 32 visuospatial items, organized in four scales of 8 items and corresponding to the four forms of relational reasoning. Besides the 32 scored TORR items, each scale of the TORR also includes two sample items designed to familiarize participants with the format of the items before they attempt the scored items. It should be noted that a full review of TORR item construction, previous research using the TORR, and detailed information on the latent structure of the relational reasoning construct is available in Chapter 2 of this document. However, some necessary introductory information on the TORR appears here.

The TORR is intended to measure relational reasoning ability among older adolescents or adults, and investigations of the reliability and validity of the TORR with that population have yielded promising results (Alexander et al., 2015). For example, the TORR scores present good reliability ($\alpha = .84$) and appropriate item difficulty values. The TORR has also been subjected to expert validation using retrospective interviews, and has been shown to significantly positively correlate with performance on SAT released items, undergraduate GPA, working memory capacity, and fluid intelligence (Alexander et al., 2015). The TORR was also factor analyzed to confirm that it did indeed contain multiple dimensions corresponding to the forms of relational reasoning around which it was designed. In the domain-specific context, the TORR has also been used to predict innovative ability in the domain of mechanical engineering (Dumas & Schmidt, 2015).

Moreover, the TORR has been fully calibrated for older adolescent and adult population using multidimensional item-response theory (MIRT) based latent variable measurement models (Dumas & Alexander, in press). Through the use of that MIRT model, item parameters (i.e., difficulty, discrimination, and guessing) were estimated, and the residual dependencies among related items (e.g., those on the same scale) were accounted for. Also through MIRT modeling, general relational reasoning ability was estimated separately from the particular forms of the construct, yielding the Relational Reasoning Quotient (RRQ), a measure of generalizable relational reasoning ability not confounded by specific analogy, anomaly, antinomy, or antithesis ability.

Further, throughout the construction of the TORR, every effort was made to limit the amount of crystallized or culturally specific knowledge that is required to correctly respond to the items. Specifically, in order to limit the reading load of the TORR, all items were constructed as graphically presented visuospatial arrays, and the scale and item directions were piloted repeatedly to ensure they were maximally simplistic and comprehensible. In this way, the TORR is designed to tap as little construct irrelevant variance as possible. However, this focus on the novelty and generality of items during test construction does not guarantee that the TORR functions equally well across various demographic or cultural groups within our target population. Therefore, whether or not TORR items function invariantly or differentially across demographic groups remains an empirical question.

Purpose of Study

The purpose of this study was to empirically examine the cultural fairness of the TORR across multiple gender, ethnic, and language groups. To do this, TORR data were analyzed using modern statistical techniques for uncovering differential-item functioning, with the ultimate objective of producing a finished TORR that is unbiased toward any gender, ethnic, or language group for use in school and university settings. It should be noted here that differential-item-functioning (DIF) refers to a situation in which an item's parameters (e.g., difficulty, discrimination, or guessing) are not invariant across demographic groups (Livingston, 2012). Importantly, the presence of DIF does not necessarily imply item bias. Instead, item bias is defined here as DIF caused by construct irrelevant variance from an extraneous culturally-specific construct that is systematically and unevenly distributed across the demographic groups in a target population (Reynolds & Ramsay, 2003; Thissen, Steinberg, & Wainer, 1993). Here, the term *cultural-fairness* is used to describe the goal-state of this investigation, in which none of the TORR items display bias.

These theoretical definitions of DIF and bias, coupled with the purpose of this study, required three conceptually differentiated conditional phases to the investigation. First, the presence of DIF in TORR items across gender, ethnicity, and language groups needed to be tested using a multi-group latent variable measurement model procedure. Secondly, if DIF were detected, it would have to be further scrutinized theoretically and empirically in order to determine whether it constitutes item bias or not. Finally, if bias is determined to be present in one or more TORR items, systematic responses to that bias would be required. Moreover, statistical corrections for bias, such that TORR scores from students in different demographic groups were interpreted using different normative scales, would be considered. It should be noted here that if significant DIF were not uncovered during the first stage of this investigation, than the second and third stages would be unnecessary and therefore would not be pursued.

Taken together, these three conditional phases of this investigation will conceptually address its purpose: namely, to systematically examine the cultural-fairness of the TORR.

Research Questions

This study's three conceptual phases logically lead to three specific and consecutive research questions. Therefore, in order to meet the stated purpose of this investigation, the following three research questions were posed:

1. Do any of the TORR items display significant differential-item-functioning (DIF) across gender, ethnic, or language groups?

2. If significant DIF is uncovered, does that DIF indicate item bias?
3. If item bias is determined to exist, what test revisions or statistical corrections can be meaningfully applied to remedy that bias?

Key Terms

For organizational clarity, key terms are organized conceptually, rather than alphabetically.

Constructs Being Measured

Relational reasoning. The ability to discern meaningful patterns within any stream of information (Alexander et al., 2012; Bassok, Dunbar, & Holyoak, 2012).

Analogy. A structural similarity between two or more concepts, objects, or situations. (Goswami, 2013; Holyoak, 2012)

Anomaly. An abnormality, digression, or deviation from an established pattern (Trickett, Trafton, & Schunn, 2009).

Antinomy. A mutual exclusivity among sets of information, (Dumas et al., 2014; Gardner, 1995).

Antithesis. A directly oppositional relations between two ideas, concepts, or objects (Baker, Friedman, & Leslie, 2010).

Statistical Concepts

Latent variable measurement models. A class of statistical models, including both confirmatory factor analysis (CFA) and multidimensional item-response theory (MIRT), that posit an underlying ability or trait causing a participant’s response pattern on a measure. In this conceptualization, test items are considered observed indicators of the ability or trait the test is measuring. Following the IRT tradition utilized in this study, item parameters (i.e., difficulty, discrimination and guessing) are generated using latent variable measurement models (Hancock & Mueller, 2013; Reckase, 2009). It should be noted here that the differences and similarities between the CFA and IRT approaches to latent variable measurement modelling are detailed in Chapter 2 of this document. The following terms specifically relate to IRT models:

Item parameters. Statistics about a given test item that are estimated using a latent variable measurement model, and used to describe key elements of that item's functioning. Specifically, item parameters can be conceptualized using the following item-response function, which models the probability of an item being answered correctly, given a participants' ability level and that item's parameters (Embretson & Reise, 2013). As seen in the following equation:

$$P(x_{ij} = 1 | \theta_i) = c_j + \frac{1 - c_j}{1 + e^{-a_j(\theta_i - b_j)}} \tag{1}$$

In this model, theta (θ) represents a participant's level of ability, and *e* is the irrational constant 2.718. Each of the other parameters will now be further described.

***Discrimination parameter* (*a*).** A measure of an item's effectiveness at distinguishing participants with different levels of latent ability. Also, a measure of how strongly a given item relates to the underlying ability it is intended to measure.

***Difficulty parameter* (*b*).** The amount of latent ability, on a z-score metric, required for a participant to have a particular probability of responding correctly to a given item. This probability is termed the inflection point, and is calculated separately for each item using the formula $\frac{1+c}{2}$. In a model without a guessing parameter (c), the inflection point is always the probability of .50 getting the item correct. However, when the guessing parameter is in the model, that inflection point must be adjusted. In this way, the difficulty parameter for each item is the amount of latent ability, on a z-score metric, needed to have a greater than $\frac{1+c}{2}$ probability of getting the item correct.

***Guessing parameter* (*c*)**. The probability of a participant responding correctly to a particular item given they have extremely low ability.

Test Development Terms

Differential-item-functioning (DIF). A situation in which item parameters (e.g., difficulty, discrimination, or guessing) is not invariant across groups of respondents (Holland & Wainer, 1993; Livingston, 2012).

Item bias. A situation in which DIF is caused by construct irrelevant variance arising from an extraneous culturally-specific trait that is systematically and unevenly distributed across the demographic groups in a target population (Reynolds & Ramsay, 2003).

Cultural-fairness. A goal-state, in which a test exhibits no bias towards any demographic groups included in the target population (Saklofske, Vijver, Oakland, Mpofu, & Suzuki, 2015).

CHAPTER 2:

REVIEW OF THE LITERATURE

In this chapter, literature relevant to this research endeavor will be reviewed. First, scholarly work pertaining to culturally fair assessment will be broadly reviewed. Importantly, culture-fair assessment has been of interest to a wide variety of researchers across the social sciences, and therefore this portion of Chapter 2 will pertain to culture-fair assessment generally, regardless of the domain of assessment or construct being tapped. In this section, the motivation for, history of, and statistical definitions of culture-fair assessment will be reviewed and discussed. As with the review of relational reasoning literature, this portion of Chapter 2 will seek to situate the current study within a larger historical context. Specifically, the major trend toward social-justice in educational research, practice, and reform, and how this trend has reverberated through the assessment literature will be discussed.

Second, literature pertaining to the conceptual definition of the construct of relational reasoning will be discussed. It should be noted that, because of the breadth and depth of the literature that either directly or indirectly influences the conceptualization of relational reasoning operating in this study, that literature will be selectively reviewed to provide a meaningful but concise overview. Interestingly, relational reasoning has been an implicit or explicit focus of many programs of psychological research since the formal inception of the field (Dumas et al., 2013), and therefore the portion of this chapter pertaining to the construct itself will have a historical focus. Moreover, the theoretical and empirical distinctions among the specific forms of relational reasoning (i.e., analogy, anomaly, antinomy, antithesis), as well as the accompanying literatures, will be explicated and reviewed.

Third, current and past methods for measuring relational reasoning and its particular forms will be reviewed, with a particular focus on whether and how existing measures of the construct meet criteria for cultural-fairness. Then, a detailed account of the development of the TORR will be presented, including a conceptual explanation of the correspondence between the definition of relational reasoning and its forms and the specific elements of TORR items. Item parameters derived from an initial calibration of the TORR will be presented in this section, and previous empirical work using the TORR will also be discussed in this portion of Chapter 2. Finally, the potential cultural fairness of the TORR will be discussed, including literature-based hypotheses concerning forms of DIF that may or may not indicate bias.

Culture-Fair Assessment

Before describing the extant literature pertaining to the assessment of relational reasoning and the TORR development specifically, it is necessary to overview the history of culture-fair assessment more generally. In this section, three aspects of culture-fair assessment will be overviewed: (a) general motivation for assessments that are culture-fair, (b) a brief history of culture-fair assessment, and (c) available statistical methods for examining cultural-fairness.

Motivation for Culture-Fair Assessment

In any situation in which a measure is being administered to a culturally diverse group of participants, such as is nearly always the case in heterogeneous countries like the United States, valid inferences from test data are impossible if cultural differences unduly affect test scores (Messick, 1980). For example, if a given test is administered to all undergraduate students at a given university, and scores on that test are affected by students' cultural background, than inferences about the difference in ability between the students are not possible, because they are confounded by cultural traits unrelated to the actual purpose of the test. For this reason, culture-fair assessment is a necessity for the formation of valid inferences concerning the ability of participants who vary in terms of cultural background. As such, the 2014 edition of the *Standards for Educational and Psychological Measurement* posit an analysis of the appropriateness of a given measure for a diverse group of participants as an integral stage in the test development process. Because of this explicit industry standard, published measures that contain uninvestigated or non-remedied item bias may be open to legal ramifications if used in the practical setting, and are likely to rapidly become irrelevant in educational practice as well as in the research literature.

Moreover, modern educational researchers and practitioners who believe in the equality of professional, intellectual, or economic opportunity across cultural groups (e.g., Camilli, 2013; Padilla & Borsato, 2008) argue strongly that eliminating cultural bias in assessment is necessary for social justice. Based on this principal, any test-takers, regardless of the cultural group to which they belong, should be able to demonstrate their ability by responding to test items. Because educational and psychological measures are often used to identify students who have high or low ability for various purposes including employment, special education programs, or gifted/honors programs, this social justice aspect is highly relevant. For example, if membership in a particular cultural group biased test scores downwards, students in that group would be systematically underrepresented in whatever program that test was used for selection into. Today, a situation such as this would be considered by many to be highly problematic from an ethical perspective. Interestingly however, this was not always the case. As will be explained, while cultural or ethnic differences in ability have long been a focus of educational and psychological investigation, the actual fairness of measures used by researchers and practitioners has been of interest for a much shorter amount of time.

History of Culture-Fair Assessment

Since the beginning of the formalized measurement of human cognitive abilities, the assessment of those abilities across diverse groups of participants has been a research focus (e.g., Galton, 1869). In the late 19th and early 20th centuries, when initial psychometric methods for measuring cognitive ability were being developed, the interest of many researchers was directed toward the examination of cross-cultural differences in mental ability (Cole & Zieky, 2001). In many of these early investigations, the purpose of assessment was to make explicit comparisons of the cognitive abilities across groups. Because the measures utilized were nearly always developed by researchers of European or European-American descent, and often contained instances of cultural-bias, inferences drawn from cross-cultural comparison studies were often invalid (Poortinga, 1995; Sternberg, 2007; Zurcher, 1998). However, because the results of these studies typically reinforced widely-held gender and racial stereotypes, they were frequently accepted as true by the mainstream psychological community (Gould, 1996).

For example, large-scale cognitive testing in the United States began in World War I, when by 1919 the Army Alpha and Beta tests were administered to approximately two million soldiers (McGuire, 1994). Subsequent analysis of these data supported the stereotype, at that time widely held in the United States, that White Americans had superior cognitive abilities to other ethnic groups (Cole & Zieky, 2001). Despite W.E.B Du Bois's impassioned 1920 essay "Race Intelligence," in which he argued that differential opportunities to learn information that was at the time specific to White American culture, and not genetic predispositions, were the root of these observed group differences, the mainstream view that those abilities were based on genetic predispositions went largely unchallenged.

In another influential study, Klineberg (1935) attempted to rank different cultural and ethnic groups around the world in terms of their intelligence by using traditional IQ tests of the time. Klineberg concluded from his data that those participants of White-European origin had the highest average intelligence, while Australian Aboriginals had the lowest. Although this study did much to fuel the contemporary movement, termed eugenics, that held as one of its principal tenets that cognitive abilities were dependent on an individual's genetic background, it has since been heavily criticized because the measures utilized contained many items that required European cultural knowledge to answer correctly, and as such were probably biased against other groups (Greenfield, 1997; Poortinga, 1995; Sternburg, 2007).

The eugenics movement, fueled in part by psychological and psychometric investigations of human cognitive ability, was widely popular with scholars and laypeople alike in both Europe and North America during the 1920's and 1930's (Bashford & Levine, 2010). In the 1930's and early 1940's, Nazi Germany cited a myriad of published psychological work to support its policies of racist oppression and hatred (Fischer, 2012). However, when World War II broke out, the eugenics movement began to lose popularity in English speaking countries, in part because of its association with Nazism and also because of a growing understanding of cultural and educational effects on the measurement of cognitive ability, such as those pointed out by Du Bois (1920). Moreover, the need to identify cognitively capable individuals to perform military operations during World War II drove the demand for cognitive assessments that were less culturally dependent.

In 1940, Raymond Cattell published his Culture-Free Intelligence Test, in which he attempted to limit the amount of crystallized intelligence required for correctly answering the items. Interestingly, only later would Cattell coin the term "crystallized intelligence" in his factor-analytic research to refer to the dimension of intelligence tests that required language, cultural knowledge, or formalized schooling (Cattell, 1963). Cattell's notion that the concept of cultural-fairness is inherently one of measure dimensionality, and specifically one of limiting the influence of a crystallized dimension, deeply informs modern conceptualizations of culture-fair assessment. Only one year later, in 1941, John Raven published a standardized version of his Progressive Matrices, which featured visuospatial arrays that supposedly carried no specific cultural valence as the primary item stimuli. Interestingly, both of these measures were designed to measure fluid intelligence, a construct that is defined very similarly to relational reasoning. For example, Cattell defined fluid intelligence as the ability "to perceive relations in completely new material" (1987, p. 298). For this reason, both Raven's Progressive Matrices and Cattell's Culture-Free Test have been used as measures of relational reasoning, an issue that will be returned to later in this chapter.

While Cattell and Raven had contributed measures that attempted to be culture-fair during the 1940's, various educational and psychological assessments that remained in regular use during the immediate post-war decade made no such attempt. However, as much greater attention began to be paid to the cultural and educational effects on the measurement of cognitive ability during the 1960's, interest in the cultural-fairness of educational measures increased rapidly. Indeed, the Civil Rights Act of 1964 placed new legal burdens on test designers to empirically demonstrate the cultural fairness of those tests (Cole & Zieky, 2001). Interestingly, in 1969, when Jensen argued that a genetic component was likely the cause of observed differences in educational test scores between White and Black Americans—an argument that was entirely mainstream only 30 years before—he was met with a powerful counterargument. Specifically, in the 6 years after Jensen's argument was published, hundreds of scholarly articles were published arguing that test bias and differences in educational opportunities contributed to the observed test differences among American ethnic groups (Cole & Zieky, 2001). Soon after, in 1971, the Supreme Court ruled in *Griggs vs. Duke Power Company* that employment tests that exhibited mean differences among ethnic or cultural groups could only be used if they had specifically been validated for the job in question (Elliot, 1987). This legal decision reflected a widespread belief held by both psychologists and the general public during that time—that mean differences in scores on a test among cultural groups was an indicator of bias.

However, in the last two decades of the 20th century, the conceptualization of bias as being indicated by mean differences among cultural groups waned, in favor of a conceptualization of bias as being indicated specifically by item parameters, while ability level is held constant (Holland & Wainer, 1993; Swaminathan & Rogers, 1990). In this way, research questions about bias were reformulated into the question: does a test item function similarly across cultural groups? This may have taken place largely because of a decreased emphasis in the literature on the genetic causes of ability differences, and the increased emphasis on learning opportunities as the source of variability. In effect, this specific formulation of bias led to the common practice of discarding items from standardized tests (e.g., SAT, ACT, or IQ tests) that exhibited DIF, under the assumption that they were biased (Penfield & Camilli, 2006). In this way, items were required to be unaffected by group membership, sometimes referred to as "culture-free"—a term that Cattell had coined to describe his 1940 test.

However, it soon became apparent that the creation of an entirely "culture-free" assessment might not be possible. In her classic article, Greenfield (1998) argued that any assessment is inherently an artifact of the culture within which it was created, and assessments simply cannot cross cultural boundaries without their measurement properties being affected.

Sternberg (2007) largely agreed, but added the caveat that the term “culture,” as utilized in the assessment literature, is rather vague and can be used to describe highly divergent populations (e.g., British students and Australian Aboriginals) or populations that are much more similar in terms of their geographic environment, but who have differing histories, socioeconomic status (SES), or access to education on average (e.g., White and African Americans). Therefore, while it is likely not possible to create a culture-free assessment, it may be possible to create one that is culture-fair, especially for a target population with only as certain amount of heterogeneity. For example, the same test may be culture-fair if given to a sample of American college students, with all the differing cultural backgrounds represented within that population, but not be culture-fair if used to compare the same college students to a group of Amazonian tribespeople.

Further, differing modern perspectives on precisely what about testing can be biased has complicated the literature on culture-fair assessment. For some scholars (e.g., Reynolds & Ramsay, 2003), test scores per se are not biased, only inferences drawn from them, if those inferences do not correspond to an appropriate usage of the test. For example, if a test of the English language is given to a sample of middle-school students in China, and the scores on that test are interpreted as a measure of intelligence, than any inference made from test scores is clearly invalid. However, if the same scores are interpreted as a measure of second language proficiency, the inferences made from them may be valid. This example corresponds closely to problems in the early testing literature, in which highly culturally specific assessments were administered to individuals outside of that culture, invalidating inferences drawn from the scores. For example, if Klineberg (1935) had interpreted his testing data not as indicators of intelligence, but as a culturally-specific construct, perhaps termed “westernization exposure,” his inferences about Australian Aboriginals, for example, may not have been invalid. Other modern perspectives on bias (e.g., Penfield & Camilli, 2006) focus on attributes of the test itself, such as an item's content, or the item's parameters calculated across diverse groups of participants. For example, in this conceptualization, even if a test has been created for the assessment of a particular construct within a particular population, and that is indeed how that test is being utilized, the items on that test could still be problematic because they could be affected by cultural factors unknown to the creators of the test.

Finally, in a conceptualization of fairness inherited largely from the factor-analytic work of Cattell (1963), the dimensionality of a particular measure is considered to be of high importance when determining its fairness. Specifically, if a test is measuring an ability that is irrelevant to whatever ability it was designed to measure, than that test could be said to have a construct-irrelevant dimension. If the ability associated with that dimension privileges one cultural group over another within the test's target population, than that test could be regarded as biased. In this way, a modern conceptualization of bias requires the simultaneous consideration of a test's purpose, target population, content, and dimensionality (Cameron, Crawford, Kenneth Lawton, & Reid, 2013; Messick, 1980; Stark, Chernyshenko, & Drasgow, 2006).

Interestingly, the modern conceptualization of cultural fairness in testing corresponds closely to W.E.B. Du Bois's (1920) original argument—that assessments cannot purport to solely measure a particular mental ability (e.g., reasoning or intelligence) if the test items require another ability (e.g., cultural knowledge) in order to be answered correctly. This conceptualization was later formalized in the exploratory factor analytic tradition by Cattell (1963), and then brought into the latent variable measurement model framework in both the CFA and MIRT literatures (e.g., Stark et al., 2006). In the next section, evolving methodologies for detecting DIF and examining cultural fairness will be discussed.

Examining Cultural Fairness

In this section, three methods for examining cultural fairness will be reviewed, methods utilizing a criterion score, the Mantel-Haenszel Test, and latent variable measurement models.

Criterion methods. One way to assess the cultural-fairness of a test is to calculate the predictive validity coefficient for the test separately for different cultural groups (e.g., Cleary, 1968). If the slope of the regression lines between the test scores and the criterion are roughly equivalent across groups, the test may potentially be considered culturally fair, because it may be used to accurately identify individuals who perform best on the criterion regardless of cultural background. However, one major complication to this method is that it requires a meaningful criterion that can be understood as completely unbiased itself. This is because, if a test and a criterion were biased against the same groups in the same amount, the criterion method would hypothetically show no evidence of bias. Unfortunately, a truly unbiased criterion is notoriously hard to find. For instance, Vars and Bowen (1998) used a criterion-referenced method to ascertain the cultural fairness of the SAT among groups of White and Black American test respondents. They found that, contrary to their hypothesis, the SAT actually over-predicted the first year college GPA of Black high-school students. Unfortunately, this finding may not necessarily indicate that the SAT is free of all bias, but that college GPA's are affected by other attributes of students (e.g., SES, educational background) not measured by the SAT. Indeed, from a dimensionality perspective, college GPA may have been a biased criterion, because it was not a pure measure of cognitive ability, but rather it was affected by other factors that were disproportionately distributed among the cultural groups within the target population.

Verney and colleagues (2005) also used a criterion-referenced method to investigate the cultural-fairness of the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 2008).

Specifically, they used an eye-tracker to derive fine-grain measurements of attention to stimulus in WAIS items across groups of White and Mexican American undergraduate students. Despite finding no significant difference in total attention between the two groups, these researchers showed that measures of attention were much better predictors of WAIS scores for White students than they were for Mexican students. This finding sheds doubt on the cultural fairness of the WAIS, but it also relies crucially on the belief that the eye-tracking measurement was a pure assessment of attention for both groups of students without any intervening factors. Because one unanalyzed factor in particular, English language ability, may have mediated the relation between attention and WAIS scores in this study, its generalizability is potentially limited. For reasons such as these, many researchers choose to use other methods to examine cultural fairness that more specifically target the probability of a participant correctly responding to a test item, regardless of their membership in a particular cultural group.

Mantel-Haenszel (MH) Test. The MH test was first proposed by Mantel and Haenszel (1959) as a procedure for studying differences between matched groups in cancer research, and subsequently adapted for use in the detection of DIF by Holland and Thayer (1988). The calculation of the MH statistic requires the creation of multiple 2 × 2 contingency tables, indicating group membership as well as item responses for multiple matched levels of ability. These multiple tables are required because, in a methodology incorporating only one 2 × 2 contingency table, the functioning of the individual item would be confounded by potential group level ability differences that are present in the data. When the MH test is completed, it yields a test statistic that approximates a chi-square distribution with one degree of freedom, making it conceptually straightforward to evaluate its significance. It was perhaps this quality of the MH statistic that made it the DIF detection statistic of choice in large-scale educational testing environments during the 1990's (Dorans & Holland, 1993).

One inherent limitation of the MH test is that it specifically deals with item difficulty, in terms of the likelihood that a particular individual, who is a member of a certain group, will correctly respond to an item. As such, the MH test does not ascertain whether an item discriminates among participants differently depending on their group, nor is it able to find DIF in the guessing parameters of items. In this way, the MH statistic can be conceptualized as a classical test theory approach to DIF, although it also corresponds closely to the Rasch model (Dorans & Holland, 1993), which is an IRT model that constrains all of the items' discrimination and guessing parameters to be equal, and only models item difficulty differences. Moreover, similarly to classical test theory or the Rasch model, the MH test cannot take into account multidimensionality. However, because DIF in guessing and discrimination parameters, as opposed to only difficulty parameters, are of interest in the development of the TORR, another methodology will be required.

Latent variable measurement models. Since the early days of formal psychometric research (e.g., Spearman, 1904), scores on psychological measures have been conceptualized as indicators of participants' actual ability. This true ability is termed "latent" because it underlies or causes the observed scores on a given measure, but cannot be observed directly. Moreover, because test scores are merely indicators of underlying latent ability, they contain measurement error. As such, one of the main goals of psychometric research has been to quantify and reduce that measurement error (Schmidt & Hunter, 1996). During the first two-thirds of the 20th century, this task was undertaken using classical test theory methodology. However, the invention of latent variable measurement models, both in the CFA and IRT traditions, profoundly altered this endeavor. Specifically, latent variable measurement models utilize the covariance and mean structure among measurements to estimate item parameters, participant ability levels, and error.

Differences between CFA and MIRT approaches. Throughout Chapter 1 of this document, and Chapter 2 thus far, latent variable measurement models have been described as a single entity, and terms relating to CFA and MIRT have been used interchangeably. However, despite the conceptual similarities of these approaches, it is important to note that differing theoretical and traditional stances are held by those researchers who commonly use each type of model. What's more, differing terminologies are also employed within each of those latent variable modeling approaches that must be specifically explained here. What follows is a systematic explanation of those terminologies that are commonly utilized across the CFA and IRT traditions.

*Discrimination and loadings***.** Discrimination parameters, in the IRT tradition, are measures of the degree to which an item separates students who have high ability from those who have low. Conceptually, discrimination parameters are also measures of how closely an item relates to the latent ability that it is measuring. In the CFA tradition, the latter aspect of discrimination parameters tends to be emphasized as the principally important purpose of CFA loadings, although the ability of an item to separate students is also indicated by a loading. It should be noted that discrimination parameters are conceptually equivalent, except for scaling differences related to the non-linearity of item response functions, to unstandardized CFA loadings (Embretson & Reise, 2013). Also, whether in the IRT or CFA tradition, the verb "to load on" is used to describe what individual items do to the latent variables they are measuring.

For that reason, using a MIRT model, items on the TORR can be described as loading on various latent ability factors, and as having discrimination parameters.

Difficulty and intercepts. Both IRT and CFA models estimate intercepts for each item that correspond to the expected score on that particular item when the factor is set to zero. In situations in which a factor is standardized, the intercept would therefore represent the mean value for that item. However, because the focus of CFA research is typically the covariance among items, CFA models without a mean structure, and hence intercepts, are often presented in the literature (e.g., Alexander et al., 2015). In the IRT tradition, intercepts are divided by negative discrimination parameters to yield difficulty parameters. Difficulty parameters have the added benefits of representing a specific interpretable quantity for every item: the value of theta that a participant needs to have a greater than 50% (or greater depending on the guessing parameter) chance of getting the item correct.

Guessing. One main difference between the IRT and CFA traditions is the presence of the guessing parameter. Whereas IRT models typically include the guessing parameter (i.e., are 3-parameter models) if the items are selected-response, CFA models do not feature a guessing parameter. As previously mentioned, the IRT guessing parameter corresponds to the probability of an item being answered correctly, given that a participant has no ability.

Non-invariance and DIF. In the CFA tradition, the property of latent measurement models to have the same or very similar item parameters across groups is termed invariance. Conversely, if the item parameters differ significantly across groups, they are termed non-invariant. In the IRT tradition, non-invariance is instead called differential item functioning, or DIF. This difference in terminology likely reflects a greater concern, among IRT researchers, with the functioning of test items, rather than the more model-focused concerns of research typically working within the CFA tradition (Kim & Yoon, 2011).

Latent variables, constructs, factors, dimensions, and abilities. For the purposes of this investigation, these five terms: latent variables, constructs, factors, dimensions, and abilities, are largely used as synonyms. Indeed, all of these terms refer to the particular aspects of a participant's mind that a test is measuring. For example, in this study, relational reasoning is the construct of interest. Therefore, it will be measured as a latent variable, or as a factor, within a measurement model. Moreover, relational reasoning is an ability being measured by the TORR, and therefore it is one of the dimensions of that test. In this way, all of these terms are appropriate to use to describe relational reasoning within the context of this investigation.

Measurement models and cultural-fairness. Further, latent variable measurement models can be fruitfully extended to investigate the cultural-fairness of measures. This can be accomplished through a likelihood-ratio procedure, in which the fit of various measurement models to the data are compared (Stark et al., 2006; Woods, Cai, & Wang, 2013). This likelihood-ratio procedure is iterative and requires multiple steps to complete. In the first step of a likelihood-ratio procedure for examining the cultural fairness of a measure, researchers may fit a multi-group measurement model to the entire dataset, allowing each of the model parameters to be freely estimated across groups. Importantly, this "free-baseline" model also requires one referent item per latent ability to have its parameters set to equality across the groups being compared (Stark et al., 2006)

In the case of the proposed study, this first model would include dimensions corresponding to each scale of the TORR, as well as a general-ability dimension, and would allow the parameters of all items but four (one for each scale of the TORR) to be freely estimated across groups. In the second step, a model that constrains each of the parameters associated with one of the non-referent TORR items to be equal across groups would be run. Then, a nested model comparison between these models would yield a chi-square test value of the significance of the difference between the fit of the free-baseline model and the constrained model. If the chi-square value for a given constrained item is found to be significant at the appropriate degrees of freedom, and generally at a Bonferroni corrected critical *p*-value (Stark et al., 2006), then that item's parameters display DIF.

After iterating this process with the remaining items, DIF can be detected in any of the items on the TORR. Importantly, this iterative likelihood ratio procedure most be repeated for each type of group. In this proposed investigation, DIF will be investigated between gender, ethnic, and language groups. For instance, DIF may exist between males and females, among White, Black, Asian, and Hispanic students, and between students whose first language is English and those whose first language is not English. While the likelihood-ratio test for DIF is computationally and analytically intensive, it yields important information about DIF that can be used to improve the TORR, and make it more suitable for identifying intellectual potential in diverse groups of participants.

Importantly, there are various latent variable measurement model based tests for DIF that have been followed in the literature. These procedures typically differ in terms of the criteria used for determining the significance of DIF or parameter invariance. For example, one previously utilized strategy has been to focus on modification indices, also termed Lagrange-multipliers in a CFA framework, in order to identify parameters that, being constrained across the groups, produce significant misfit. However, the method described, in which the fit of iterative nested models are compared, is considered to be a more modern approach (Mann, Rutstein, & Hancock, 2009; Woods, Cai, & Wang, 2013). This approach requires greater computational resources than the Lagrange-multiplier method, but it allows for a more direct measurement of the change in misfit associated with the freeing or constraining of model parameters across groups. As the processing power of typical personal computers has increased, the iterative likelihood ratio method can be run efficiently in statistical programs specializing latent variable measurement models such as FlexMIRT (Cai, 2013) and Mplus (Muthen & Muthen, 2012).

Moreover, some researchers have followed the likelihood ratio procedure but in the opposite direction from what was described here, meaning they began with a multi-group model in which all the parameters are constrained across groups, and then systematically freed them to be equal across groups, examining the significance of the decrease in misfit each time (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001). While this approach is conceptually similar to the process already described, it holds some statistical pitfalls. For example, in order for the difference in fit between two nested models to follow a chi-square distribution—an important assumption of likelihood ratio tests of nested latent variable measurement models—the baseline model needs to fit the data (Maydeu-Olivares & Cai, 2006). If the baseline model were to have sizeable misfit, perhaps from mis-specification associated with constraining parameters across groups, the likelihood ratio test statistic may not actually follow a chi-square distribution, as it is assumed to.

Interestingly, the latent variable measurement model likelihood-ratio procedure also conceptually combines some of the strengths of both the criterion method and the MH test. In regards to the MH test, which specifically investigates the likelihood of participants of matched ability but different groups getting an item correct, the latent variable measurement model accomplishes the same task, but separates the probability of answering an item correctly in terms of guessing and difficulty parameters. In the proposed study, it may be potentially interesting to observe whether certain groups are more or less adept than other groups at guessing the correct answer to TORR items. This particular information about guessing is not available through the MH test. Further, in terms of the criterion method which focuses on the relation between test scores and a particular criterion, the latent variable measurement model method identifies DIF in discrimination parameters and loadings, which can be conceptually described as measures of the relation between the test item and the underlying ability that it is measuring. In this way, while latent variable measurement models used for testing DIF do not typically include an outside criterion, they do speak to the relations among the test items as well as to their underlying trait, which may accomplish a similar conceptual goal as a criterion method. For example, if significant DIF exists in a particular TORR item's discrimination parameter, that item may be more closely related to relational reasoning ability for one group than another.

For instance, Greiff and colleagues (2013) used a latent variable measurement model likelihood ratio approach to examining the functioning of a complex problem solving measure, which was designed to be culture-fair. They found that the model's parameters adhered to strong factorial invariance among the groups that they included in their analysis, meaning that all item parameters, when constrained to be equal across groups, did not produce significant misfit compared to when those parameters were free to vary. If the findings of the proposed study parallel the findings of Greiff and colleagues (2013), that would indicate that each item on the TORR functions similarly across gender, ethnic, and language groups. Indeed, the construct being measured in Greiff and colleagues's study, complex problem solving, has been conceptualized and defined similarly to relational reasoning, and therefore the operationalization and measurement of the two constructs has been similar as well (Greiff et al., 2015). In this way, there is evidence that higher-order cognitive abilities such as relational reasoning may be assessed in a way so as to produce invariant item parameters across groups.

Relational Reasoning and Its Forms

Relational reasoning has long been regarded as essential to human mental life (e.g., Hofstadter, 2001; James, 1890; Sternberg, 1977). For example, William James (1890) described the ability to uncover relations of “differences and similarity” (p. 346) between mental representations as fundamental to human thinking and learning, as well as to expert cognitive performance in any domain. Later in the 20th century, in *The Abilities of Man* (1927), Charles Spearman described a model of human cognition largely driven by the ability to “bring to mind any relations that essentially hold” (p. 165) between two ideas or concepts. His belief in the importance of this human cognitive function was so strong that he bemoaned what he saw as the neglect of the *cognition of relations* by his contemporaries. Using an idiomatic analogy, Spearman wrote that this could “only be explained by their not seeing the forest for the trees!” (p.166). This strong belief in the importance of reasoning by relations carried forward to Spearman’s students, John Raven and Raymond Cattell, both of whom created tests of intelligence that required individuals to reason with complex relations (Cattell, 1940; Raven, 1941). Even today, the Raven’s Progressive Matrices is a commonly used set of stimuli in empirical research studying relational reasoning (e.g., Baldo, Bunge, Wilson, & Dronkers, 2010).

In the modern educational psychology literature, relational reasoning continues to grow in prominence as more researchers begin to identify its potential for supporting student learning. For example, instructional interventions that specifically incorporate relational mappings have appeared in recent years (e.g., Richland & Begolli, 2015). Moreover, an increasingly diverse array of educational studies incorporating relational reasoning have been published, empirically linking the construct to student outcomes at all levels of schooling from preschool (Collins & Laski, 2015) to advanced studies (Dumas et al., 2014) and across the gamut of domains of learning from reading (Farrington-Flint & Wood, 2007) and writing (Tzuriel, & Flor-Maduel, 2010), to mathematics (Stevenson, Bergwerff, Heiser, & Resing, 2014) and engineering (Chan & Schunn, 2015; Dumas & Schmidt, 2015). This burgeoning of relational reasoning literature is supported by the idea—approaching consensus in the literature—that relational reasoning is not a unidimensional ability, but rather can arise in different forms, depending on the information being reasoned about (Chi, 2013; Sagi, Gentner, & Lovett, 2012). As such, relational reasoning can be conceptualized as a generalized construct arising in multiple manifestations within different learning contexts and domains of study.

Relational reasoning, as it is defined here and elsewhere (e.g., Bassok et al., 2012; Crone et al., 2009; Dumas et al., 2013; Krawzcyk, 2012), requires the discernment of pattern. Importantly, the connecting of multiple relations among pieces of information is required for the formation of a pattern (Simon & Lea, 1974). In this way, patterns are inherently composed of relations among relations. These relations-among-relations have often been characterized as higher order, because they are derived from a series of lower-order relations, or simple associations between individual pieces of information (Chi, 2013; Gentner & Gentner 1983; Goswami, 1992). Others (e.g., Sternberg, 1977) have referred to the multiple lower-order relations required for relational reasoning as inferences, and the higher-order patterns derived from them, as mappings.

Because analogical reasoning, at its core, involves the mapping of higher-order relations (Gentner 1982; Sternberg, 1977), analogy unquestionably fits the definition of relational reasoning. Interestingly, the inherently relational nature of analogical reasoning has led some researchers to use the term *analogy* to describe any process in which a higher-order relation is mapped (e.g., Hofstadter, 2001), while others contend that the broader term *relational reasoning* is more appropriate because it allows for the diversity of mappable relations to be explicitly described (Alexander & the DRLRL, 2012; Chi, 2013; Holyoak, 2012). While the inclusion of analogy as an important form of relational reasoning is uncontested, some researchers recently have suggested that there are further types of mappable higher-order patterns that are relevant to human cognition and education, and that focusing strictly on analogies may restrict what can be learned about human reasoning (Alexander & the DRLRL, 2012; Chi, 2013; Holyoak, 2012; Sagi, Gentner, & Lovett, 2012). Rather, these researchers suggest that relational reasoning can be seen as taking different forms depending on the high-order relation being mapped (e.g., Dunbar, 2013). Because the type of relational pattern that is ultimately mapped can differ depending on the information at hand, it may be essential to investigate different types of patterns in order to understand and potentially support students' ability to think relationally (Chi, 2013). For example, Alexander and colleagues (2012) have suggested that at least four forms of relational reasoning merit further examination: analogy, anomaly, antinomy, and antithesis.

Specifically, in contrast to analogy, which is based on relational similarity, anomaly requires the identification of a pattern based on unusual or unexpected relations between events, occurrences, or objects. In effect, an anomaly is a relational discrepancy or deviation from an established rule or trend (Chinn & Brewer, 1993; Dunbar, 1993; Klahr & Dunbar 1988, Kulkarni & Simon, 1988). As such, the term *anomaly* can be used to refer to both a process and a product predicated on a higher-order relation of discrepancy, just as *analogy* represents both a reasoning process and product based on a higher-order relation of similarity. By comparison, reasoning by antinomy requires the identification of a higher-order relation of incompatibility, and may involve identifying what a pattern *is* by isolating what it *is not*, or recognizing the mutual exclusivity among relationally-defined categories (Chi & Roscoe, 2002; Cole & Wertsch, 1996; Gardner, 1995; Sorensen, 2003). And lastly, an antithesis is predicated on a pattern of directly oppositional relations between ideas, objects, or events (Bianchi, Savardi, & Kubovy, 2011; Kuhn & Udell, 2007; Sinatra & Broughton, 2011).

While each of the forms of relational reasoning considered in this investigation may share overlapping component processes (e.g., encoding, inferring, or mapping; Sternberg, 1977), it is the characterization of the mapped relation that distinguishes one form of relational reasoning from another. Specifically, while each form of relational reasoning requires the mapping of a higher-order relation from multiple lower-order relations (Goswami, 1992; Markman & Gentner, 2000; Sternberg, 1977), the mapped relation could be one of similarity (analogy), discrepancy (anomaly), incompatibility (antinomy), or opposition (antithesis). Importantly, these may not be the only types of higher-order relations that can be mapped, but it has been argued (e.g., Alexander & the DRLRL, 2012; Dumas et al., 2013, 2014) that these four forms are worthy of investigation because of their broad applicability to educational settings in which complex cognitive processes are required.

Assessment of Relational Reasoning

The definition of relational reasoning operating in this proposed study—and which informed the development of the TORR—is the ability to discern meaningful patterns within any stream of information (Alexander et al., 2012; Bassok, Dunbar, & Holyoak, 2012; Goswami, 2013). Importantly, this definition is sufficiently broad as to encompass a wide variety of cognitive tasks. As such, relational reasoning can be thought of as manifesting itself within a great number of psychological assessments and educational activities (Dumas et al., 2013). Indeed, it may be difficult to describe a complex cognitive activity situated within either the educational or professional context, in which relational reasoning or its forms cannot be identified. Therefore, the breadth of possible instances of relational reasoning being measured—even when it is not directly termed “relational reasoning”—available in the educational and psychological literatures is immense, and it is not possible to review in its entirety here. Therefore, a select cross-section of the literature that pertains most directly to the issues of definition, operationalization, and cultural-fairness operating in this investigation will be reviewed and discussed.

Of the possible methods for assessing relational reasoning and its forms, one particular type of assessment has stood out as the most popular: four-term verbal analogies (Alexander, White, Haensly, & Crimmins-Jeanes, 1987; Goswami, 1992; Maguire, McClelland, Donovan, Tillman, & Krawczyk, 2012; Sternberg, 1977). Four-term verbal analogies follow the form A:B::C:D, where the A and B term are linked to the C and D term by a higher-order relation of similarity. For example, the four term verbal analogy *wolf:pack::lion:pride* allows for the identification of the lower-order relation of “is a member of” between *wolf* and *pack* as well as *lion* and *pride*.

Because the characterization of the lower-order relation between the *A* and *B,* and *C* and *D* terms is equivalent, a higher-order relation of similarity links the two lower-order relations, forming an analogy. This description of high-order relations, or relations between relations, is a central theme in the relational reasoning literature, and has even been used to define the construct itself (e.g., Goswami, 2013). Also interestingly, some scholars (e.g., Hofstadter, 2001) use the term *analogy* to describe all instances of higher-order relations, regardless of the direction or type of higher-order relation at work. However, in order to better reflect the diversity of possible higher-order relations, and to facilitate the explanation and instruction of cognitive procedures in the educational context, the more general term *relational reasoning* is utilized by others to describe any instance of higher-order relations being reasoned with (e.g., Chi, 2013; Holyoak, 2012; Krawczyk, 2012).

Four-term verbal analogies have been popular in tests of intelligence and academic aptitude over the past fifty years (Sternberg, 1977; Sternberg & Rifkin, 1979). Indeed, this form of reasoning task was so ubiquitous on tests that, for a time, many schools in the U.S. attempted to formally train students on them, and instructional interventions for improving analogical reasoning were developed by educational psychologists (e.g., Alderman & Powers, 1980; White & Alexander, 1986; White & Caropreso, 1984). However, the issue of cultural fairness has, over the past decade, led to the sharp decrease in instances of four-term verbal analogies on large-scale assessments. For example, analogies were long a staple of the verbal scale of the SAT, but were removed in 2005, in part because of issues of cultural fairness. It has been argued (e.g., Dixon-Román, Everson, & McCardle, 2013) that the particular relations that some analogy items require are overly nested within a particular dominant middle-class White American culture, and as such put other groups of students at risk. For example, this four-term verbal analogy item, formally from the SAT, was found to hold bias privileging White participants:

ANIMAL:HIDE::

(A) Egg:yolk
(B) Deer:hunt
(C) Desk:top
(D) Fugitive:shelter
(E) Fruit:rind*** (O'Neill &McPeek, 1993)

Unfortunately, in the United States, differing access to fresh fruit along income lines meant that knowledge of fruit and rinds was, in a way, culturally specific knowledge. This situation created construct irrelevant variance in this item, where participants who were more familiar with fresh fruit scored higher, regardless of their actual analogical reasoning ability. In the United States, White individuals were more likely than those in other ethnic groups to be familiar with fresh fruit, because of issues of food access and cost, giving White participants an advantage on this analogy item.

Further, verbal analogies sometimes contained bias along gender lines as well. For example, this former SAT analogy exhibited bias favoring women:

TILE:MOSAIC::

(A) Wood:totem
(B) Stitch:sampler***
(C) Ink:scroll
(D) Pedestal: Column
(E) Tapestry: Rug (O'Neill & McPeek, 1993)

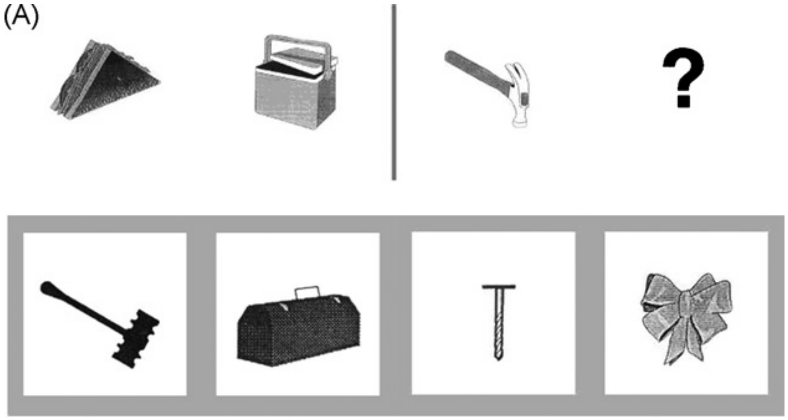
The correct answer in this analogy item pertains to sewing and sewing techniques. In Western culture, women are more likely to receive formal or informal training on sewing and to know terminology associated with that craft. Therefore, participants who were familiar with sewing terms, who were disproportionately female, were more likely to answer this item correct, regardless of their actual analogical reasoning ability.

Because these four-term verbal analogies exhibited significant DIF, and that DIF could be determined to signify bias based on the abilities required to correctly answer the items, they could not be ethically utilized on assessments. As more and more four-term verbal analogies were identified that contained construct-irrelevant variance, the decision was eventually made to cease their use on the SAT all-together. While certain assessments, such as Miller's Analogies

(1960) do still utilize this form of analogy item, far fewer students are currently exposed to them than in the past, in large part because of concerns over their cultural fairness.

For some researchers (e.g., Allalouf, Hambleton, & Sireci, 1999), the root of DIF and bias in four-term verbal analogies was in their heavy reliance on language: a crystallized ability that requires cultural exposure or formalized training to develop. Therefore, some researchers have attempted to expand the A:B::C:D analogical form, creating four-term pictorial analogies (e.g., Krawczyk et al., 2008). These analogies are designed to contain the same four-term relational structure as verbal analogies, but depict images as opposed to utilizing language. For example, the four-term pictorial analogy presented below includes an image of a sandwich, a picnic basket, and a hammer as it's A, B, and C terms, the correct answer is the second choice, which depicts a toolbox.

Figure 1. A four-term pictorial analogy

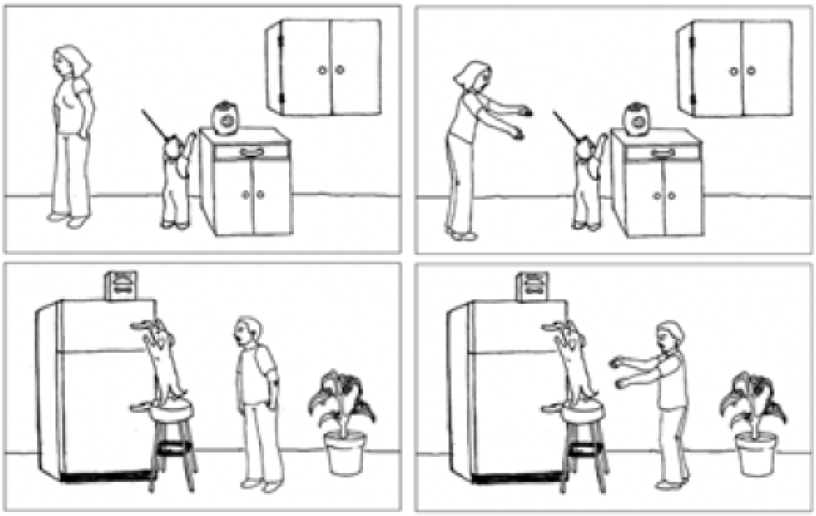


(Krawczyk et al., 2008)

Although it is accurate that the aforementioned problem does not rely directly on reading ability in any particular language, it still remains unclear whether or not it exhibits DIF or bias. For example, the images might not be clear in their meaning to all participants equally, especially if a particular group of participant has more or less experience with any of the objects. Although pictorial analogies may have certain advantages over verbal analogies: they can be hypothetically administered to young children who cannot yet read, and can be administered to different language groups so long as the objects remain recognizable, they do not necessarily solve the problem of cultural-fairness, because the objects depicted in the pictured may yet be culturally specific, or knowledge of their use may be effected by group membership.

As a way to use the four-term format with a potentially richer set of pictorial stimuli, scene analogies ask participants to identify relations among elements in an illustration. For example, this item from Richland, Chan, and Morrison (2010) requires participants to identify an analogy between the child reaching for a snack and the mother attempting to stop them in the first two scenes, and the dog reaching for a snack and the human trying to stop it in the bottom two scenes:

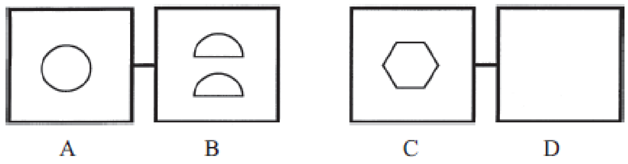
Figure 2. A scene analogy item



Scene analogies such as the example make an attempt to depict scenes that are not overly culturally specific and that will be recognizable to young children, who are the typical participants in studies utilizing scene analogies. However, the depiction of a scene remains open to potential cultural effects, because some participants may have differing schema associated with dogs and cookies, for example, because of their cultural learning opportunities. In some ways, these items may be better described as "culturally-appropriate" for the young middle-class children who are typical participants in developmental research, as opposed to fair across a variety of cultures.

As a way to move the analogy item yet further away from culturally weighted terms, the geometric or figural analogy follows the same four-term A:B::C:D pattern. However, such items use only geometric or figural elements that are manipulated to form an analogical pattern. For example, this geometric analogy created by Tunteler and Resing (2010) features a circle being split into halves, and requires the participant to perform the same operation on a hexagon:

Figure 3. A four-term geometric analogy



Because there is evidence that the operation of halving a whole is likely recognizable across cultural groups (McCrink, Spelke, Dehaene, & Pierre Pica, 2013), it is possible that this item limits possible bias. However, it remains an open question in the literature whether differential knowledge of formal or informal geometry may produce DIF in these types of analogy items. Moreover, the geometric form of these analogy items greatly increases the emphasis in visuospatial rotation and manipulation, an ability that some researchers have argued may privilege male participants (e.g., Kaufman, 2007; Voyer, Voyer, & Bryden, 1995).

Interestingly, the debate continues whether observed male advantages in visuospatial rotation are due to some underlying biological difference (i.e., testosterone levels; Auyeung et al., 2011; Vuoksimaa, Kaprio, Eriksson, & Rose, 2012) or differences in learning opportunities caused by cultural beliefs (Miller & Halpern, 2014; Pontius, 1997). However, from the perspective of cultural fairness, the source of differences between males and females on visuospatial ability is of less importance. Of more importance is the decision of whether or not visuospatial ability is relevant to the construct being assessed by the items. For example, in a visuospatial assessment of relational reasoning, some visuospatial rotation ability, or at least visuospatial working memory, may be necessary by not sufficient to complete the items correctly (Grossnickle, Dumas, & Alexander, 2015) making it potentially construct-relevant.

One expansion of the geometric analogy, which breaks with the A:B::C:D format in favor of a more complex set of stimuli is the matrix analogy. The matrix analogy is associated with Raven (1941) for his Progressive Matrices—still one of the most often utilized measures of relational reasoning. Since Raven's initial use of the form, many other researchers have created their own matrix analogies for use in the research or clinical setting (Krawczyk et al., 2011; Naglieri & Insko, 1986). In fact, the matrix form of analogy is utilized on the TORR, and therefore this form of analogy assessment will be returned to in more detail later in this chapter.

Today, the matrix format remains the gold standard for visuospatial analogical reasoning assessment, and is also used for the assessment of fluid intelligence, which, as discussed, is a closely related construct. DIF has been found on Raven's items between male and female participants in the past (Abad, Colom, Rebollo, & Escorial, 2004), and it is likely caused by the items' emphasis on visuospatial processing. However, whether or not matrix analogies can be biased remains an open question. In effect, the question is one of dimensionality and relevancy: is the visuospatial component of matrix analogies relevant to the construct of relational reasoning or not? This question is central to the proposed study and will be returned to later in this chapter and also in Chapter 3.

In terms of the assessment of reasoning with anomalies, verbal semantic anomalies have been administered regularly to participants since Binet and colleagues (1913) used a child's ability to identify *absurd phrases* as a test of intelligence. Verbal semantic anomalies are short passages or sentences that contain unusual information. For example, this short passage taken from Filik and Leuthold (2008) can be presented as nonanomalous:

Terry was very annoyed at the traffic jam on his way to work.

He glared at the [truck] and carried on down the road.

Or as anomalous:

Terry was very annoyed at the traffic jam on his way to work.

He picked up the [truck] and carried on down the road.

It can also be presented in a fictional context, negating the anomaly:

The Incredible Hulk was annoyed that there was a lot of traffic in his way.

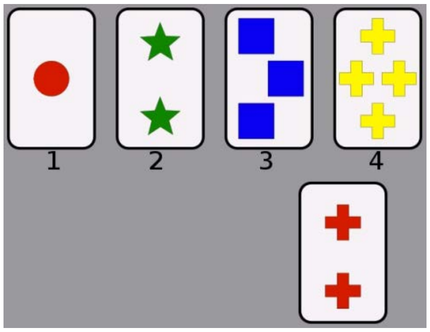
He picked up the [truck] and carried on down the road.

Verbal semantic anomalies that describe something unusual, unexpected, or in this case, physically impossible, are far and away the most widely used measure in the literature pertaining to anomaly (Faustmann, Murdoch, Finnigan, & Copland, 2007; Filik, 2008; Weber & Lavric, 2008). Their popularity may be largely due to the use of Event Related Potentials (ERP) and the predictable effect, termed N400, which is elicited by verbal semantic anomalies in ERP studies. The N400 effect refers to a potential in the negative direction, observed approximated 400 milliseconds after the stimulus onset, that is closely associated with the processing of anomalies. This clear denotation of an anomaly by the N400 effect has enabled researchers to closely study how people reason with verbal semantic anomalies and how altering the context of the phrase can negate the anomaly, and the N400 effect that goes with it.

While verbal semantic anomalies are the most popular measures of anomaly, other measures were also utilized. These include mathematical problems solved incorrectly (Chen et al., 2007), and inconsistent information presented in narrative texts (Stewart, Kidd, & Haigh, 2009). Importantly, many of these measures rely on language, and verbal semantic anomalies in particular often rely on particular culture-specific stories or fictional characters (e.g., the Incredible Hulk) and as such are susceptible to the effect of culture. For that reason, geometric anomalies, inspired by the format and visuospatial emphasis of matrix analogies were created for the TORR.

Antinomous reasoning, which requires the conceptualization of mutually exclusive categories, has been assessed through the use various categorization or sorting tasks such as the Wisconsin card sorting task (Grant & Berg, 1948). In this task, participants must select matching cards by correctly inferring rules both about what card go together and what cards do *not* go together. A screenshot of the online version of this task appears below:

Figure 4. Screenshot from the online version of the Wisconsin card sort task



Although this task has been used for decades to assess the flexibility and speed with which participants can conceptualize mutual exclusivity, it can be conceptualized as a measure of antinomy, because it requires participants to reason simultaneously with what a given card does not represent. Importantly, because the Wisconsin card sort task is open ended, it is not straightforward to administer or score, and before electronic versions with automatic scoring functions were available, administration and scoring of the task was time-consuming. In order to create an assessment of antinomous reasoning that was more readily scored and interpreted by practitioners, closed-ended selected-response type items that required similar higher-order categorization were conceptualized and created for the TORR.

The assessment of antithetical reasoning, similarly to analogical reasoning, has been very popular on large-scale high-stakes assessments in the United States in the form of verbal antonyms. Verbal antonyms present a single word, typically a relatively difficult vocabulary word, and require participants to select the answer choice that is most directly its opposite. Verbal antonyms, because of their heavy reliance on vocabulary, also occasionally exhibited DIF among language groups. Whether or not this DIF constitutes bias is debatable, because English language vocabulary is an ability, which these antonym items are often designed to measure (O’Neil & McPeek, 1993). However, as could be expected, the particular direction of DIF differed depending on the specific language background of participants. Namely, because many English words have their root in Latin and other romance languages, Spanish speakers can be privileged by certain antonym items such as:

TURBULENT:
(A) Aerial
(B) Compact
(C) Pacific***
(D) Chilled
(E) Sanitary (O’Neil & McPeek, 2012)

In American English, the word “pacific” is principally utilized to refer to the Pacific Ocean, although a seldom used alternate meaning is “tranquil.” However, in Spanish and other romance languages, that alternate usage is much more common. Therefore, students with knowledge of Spanish were more likely to get this item correct, regardless of their actual antithetical reasoning ability. As would be expected, the DIF of this item was in the opposite direction for participants whose first language was not English or a romance language, because lower exposure to English or romance language vocabulary meant a lower likelihood of answering this item correctly.

Another way that relational reasoning has been operationalized in the research literature without referencing any particular form of the construct is through relational-deductive tasks, or, as Acredolo and Horobin (1987) defined the task, “deductive reasoning about the relations between objects” (p. 1). One example of this type of assessment of relational reasoning, taken from Van der Henst and Schaeken (2005) is:

A is to the left of B
B is to the left of C
D is in front of A
E is in front of C

What is the relation between D and E?

Although this type of task has been popular in the cognitive and neuropsychological research literatures (Acredolo & Horobin, 1987; Van der Henst & Schaeken, 2005), it has been much less common in larger scale assessments. Perhaps for this reason, no investigation of which I am aware has examined the cultural-fairness of relational deductive tasks.

Another way that relational reasoning and its forms have been measured is through naturalistic observations and discourse analysis (Dunbar, 1995; Trickett et al, 2009). This method of studying reasoning is termed *in vivo* methodology, because it seeks to capture the relational reasoning of individuals as it unfolds naturally in human interactions (Dunbar, 1995). For instance, Trickett and colleagues (2009) observed meteorologists as they reasoned with unexplained weather patterns, and used those meteorologists' verbalized reasoning as a data source. Moreover, Dumas and colleagues (2014) used the recorded discourse of a team of clinical neurologists as they reasoned about patient symptoms and treatment in order to ascertain how the forms of relational reasoning operate in concert with one another to achieve a goal. Importantly, *in vivo* studies of relational reasoning are always situated within a particular context in which the reasoning is taking place (e.g., meteorology, clinical neurology), and for that reason cultural-fairness is rarely relevant in the same way as it is to traditional test-based assessment. Instead, a rich description of the culture in which a study took place allows researchers to make reasoned inferences about the generalizability of *in vivo* findings across cultures and contexts.

Today, relational reasoning is a burgeoning area of inquiry with much promise for insight into the mind and education (Sinatra, 2015). However, while the large variety of measures and assessments utilized in the field point to the depth and richness of this research, there is a need to consolidate many of the extant measurement paradigms into a single test that can be used with confidence by a variety of scholars across research contexts. For example, many of the most often utilized measures are susceptible to the effects of culture, and cannot be said to be culture-fair. Moreover, while measures exist for each of the identified forms of relational reasoning, no current measure brings together all four forms into one assessment structure. In this way, a multidimensional, culture-fair assessment of relational reasoning is needed in the field. For that reason, development of the TORR was undertaken.

TORR Development

The process of measure development for the TORR had five phases: (a) theoretical conception; (b) item creation; (c) retrospective interviews, and (d) calibration, and; (e) development of a scoring method. Please see Table 1 for a presentation of the stages of TORR development, with citations. Each of these stages will now be further explicated.

Table 1

Steps of TORR Development with Citations

Step	Citation
Definition of construct	Alexander et al., 2012
Review of literature	Dumas, Alexander, & Grossnickle, 2013
Item development	Alexander et al., 2015
Expert validation	
Pilot testing	
CTT item parameters	
Reliability	
Convergent/discriminant validity	
MIRT Calibration	Dumas & Alexander, 2015
RRQ norming	
Predictive validity	Dumas, Alexander, & Schmidt, 2015
Cultural fairness	This dissertation

Theoretical Conception

In the later 1980's, Alexander and colleagues (Alexander, White, Haensly, & Crimmins-Jeanes, 1987; Alexander, Willson, White, & Fuqua, 1987; Alexander et al., 1989; White & Alexander, 1986) published a series of studies investigating the analogical reasoning ability of young children. In order to measure analogical reasoning in that population, they created the Test of Analogical Reasoning in Children (TARC; Alexander et al., 1987b). This measure utilized visuospatial four-term analogies based on attribute blocks that varied on their size, shape, and color. Using this measure, Alexander and colleagues (Alexander, White, Haensly, & Crimmins-Jeanes, 1987; Alexander, Willson, White, & Fuqua, 1987) were able to demonstrate that even very young children (i.e., 4 and 5 year olds) could manifest analogical reasoning abilities. This research informed later investigations of the analogical reasoning ability of young children (e.g., Goswami, 1992; Richland et al., 2006), and formed the basis for the later theoretical conceptualization of the TORR.

Later, Alexander (2012), being influenced generally by set theory in mathematics, and specifically by the writing of Bertrand Russell (1973), conceptualized analogical reasoning as one form of a broader construct. This construct, termed *relational reasoning*, could take multiple forms, depending on the nature of the higher-order relations being formed. Although the set of possible higher-order relations to be identified is likely extensive, four particular types of higher-order relations, were seen as particularly salient. These higher-order relations (i.e., similarity, discrepancy, incompatibility, and opposition) were each conceptualized as an iteration of the broad construct of relational reasoning.

In order to conceptually organize these functions, and provide a lens through which to study the process of reasoning relationally, specific forms of relational reasoning were posited, each defined by the presence of one of the theoretically identified higher-order relations. As already discussed, these forms of relational reasoning were termed *analogy* (defined by a higher-order relation of similarity), *anomaly* (discrepancy), *antinomy* (incompatibility), and *antithesis* (opposition). Potential problem sets that would represent each of the four forms were than developed within the DRLRL. From the outset, it was determined that those problem sets would be entirely figural to parallel earlier work by Alexander and colleagues on analogical reasoning

(Alexander, White, Haensly, & Crimmins-Jeanes, 1987; Alexander, Willson, White, & Fuqua, 1987). It was felt that novelty and nonlinguistic elements were critical to capture the underlying relational processes without the complications that arise from more crystallized knowledge elements (e.g., language). In addition, a systematic literature review was undertaken to empirically explore the viability of the conceptualized forms and further consider item parameters (Dumas et al., 2013).

Item Creation

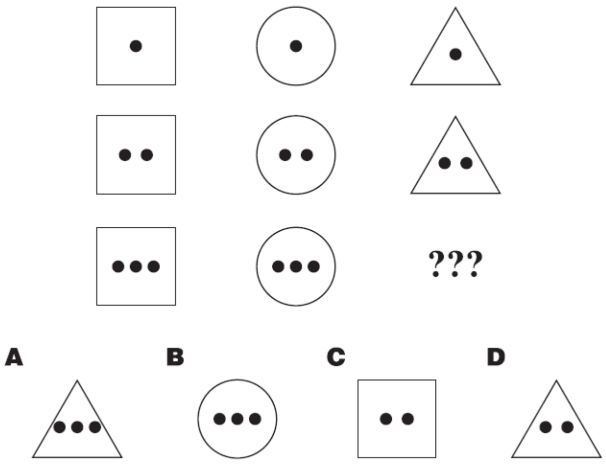
Each of the items on the TORR was originally created collaboratively, through a group brainstorming process. The correspondence between the items and the theoretical description of the forms of relational reasoning they were being designed to tap was a principal focus during the item creation stage. Each item was designed so that a correct answer would demonstrate participants' ability to utilize the corresponding relational process (e.g., similarity or discrepancy) and was structured to ensure that those underlying processes would differ across scales. Further, to avoid the need for participants to inhibit reasoning processes used on a previous scale, the patterns used in any one of the scales were not repeated in the other scales. What follows is a more detailed description of each scale with a sample item to illustrate the corresponding reasoning.

Analogy. The items on the analogy scale of the TORR were designed using a matrix format, which is a common item configuration within the analogical reasoning literature, as previously discussed. As they were in matrix format, the analogy items on the TORR consisted of a three by three matrix of figures, with the figure in the lower right unspecified. Participants are asked to discern the pattern underlying the given problem, and then decide which of the possible answer choices would complete that pattern. This underlying pattern could be identified by processing item components either vertically or horizontally. In effect, to solve matrix analogies like those on the TORR, a participant must identify the answer choice that makes one row or one column of elements relationally similar to another row or column of elements. It is this basic requirement to establish the similarity across elements of the problem in order to reach the solution that defines these items as analogical in nature (Carpenter, Just, & Shell, 1990), because analogical reasoning is generally defined by the presence of a fundamental relation of similarity (e.g., Holyoak, 1985; Novick, 1988),

The analogy scale directions and a sample item from that scale are provided here. It is important to note that these sample items on the TORR are used to familiarize participants with the structure of the test, and as such are intended to be relatively easy, compared to most of the items within the scale.

Figure 5. A sample matrix analogy item from the TORR

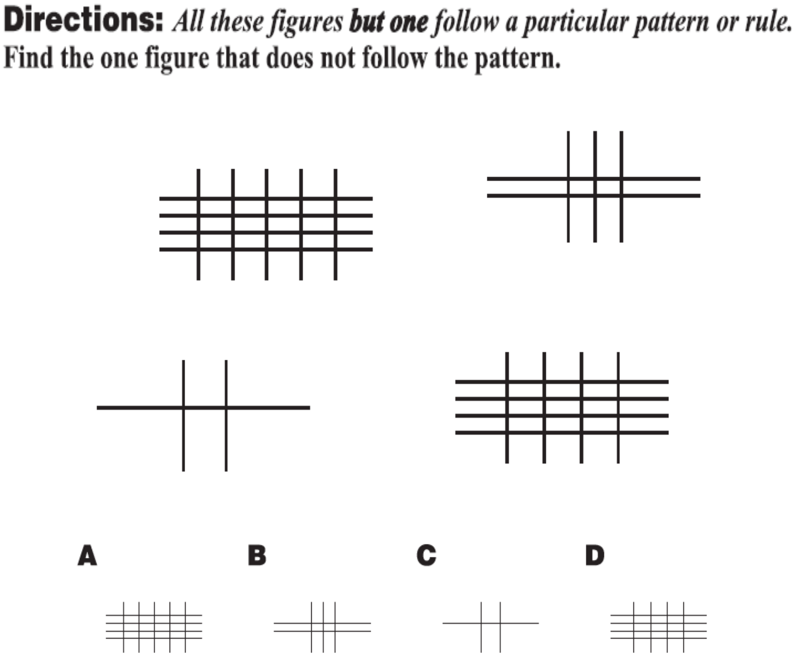
Directions: ***Below is a pattern that is not yet complete.***
Select the figure from those shown below that completes the pattern.



In this sample item, the correct answer is A, because that answer choice completes the pattern of changing shape (square, circle, triangle) across the rows and increasing number (1,2,3) of dots down the columns of the matrix.

Anomaly. Unlike the analogy scale, which was designed in the widely used matrix format, there were no the measures of anomaly currently utilized by the field that were graphical in nature. So, a format for the items on this scale of the TORR was formulated based upon the conceptual nature of an anomaly. Because an anomaly is defined as an unexpected deviation from a pattern or rule (Chinn & Brewer, 1993; Klahr & Dunbar, 1988), the items on the anomaly scale necessarily had to depict a discernable pattern that could subsequently be broken by the anomaly itself. So, an "odd-one-out" format was used. Four figures were presented in a non-linear array: three of the figures follow the same pattern, but one of them— despite being visually similar to the other three—does not. In order to answer correctly, a participant must attend to the features of each of four figures, recognize the pattern governing the array, and then select the figure that deviates from the pattern. A sample item used to familiarize participants with this scale of the TORR is:

Figure 6. A sample anomaly item from the TORR

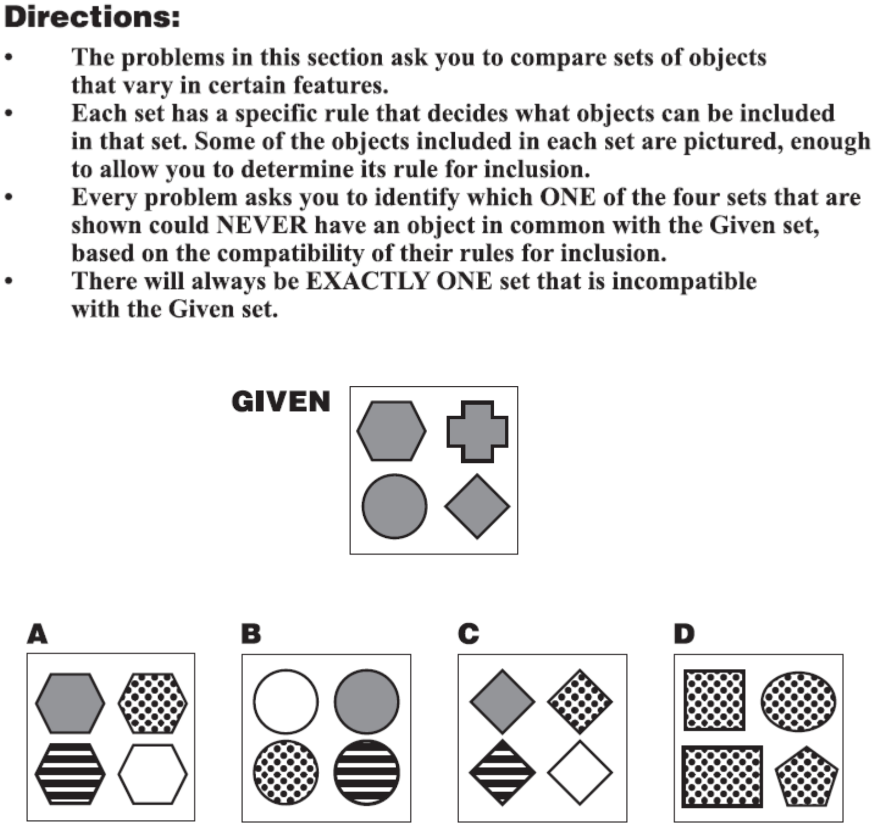


Directions: *All these figures* **but one** *follow a particular pattern or rule.*
Find the one figure that does not follow the pattern.

In this sample item, the correct answer choice is D because it is the only figure in the array that does not follow the pattern of having one less horizontal line than vertical lines. Although figures A, B, and C each follow this pattern, D breaks the rule, and thus can be identified as the anomaly.

Antinomy. An antinomy occurs when mutually exclusive sets or categories of objects or ideas are brought together (Chi & Slotta, 1993; Russell & Lackey, 1973). To reflect this theoretical definition of antinomy, the items on the antinomy scale of the TORR were designed with a given set, governed by a rule for inclusion, and four answer choice sets, also governed by their own rules for inclusion. It is the participant’s task to select the answer choice set that has a rule for inclusion that is antinomous, or incompatible, with the rule for inclusion in the given set. In other words, the correct answer choice is the set that could never have an item in common with the given set. A sample item from the antinomy scale of the TORR, along with its directions, follows.

Figure 7. A sample antinomy item from the TORR



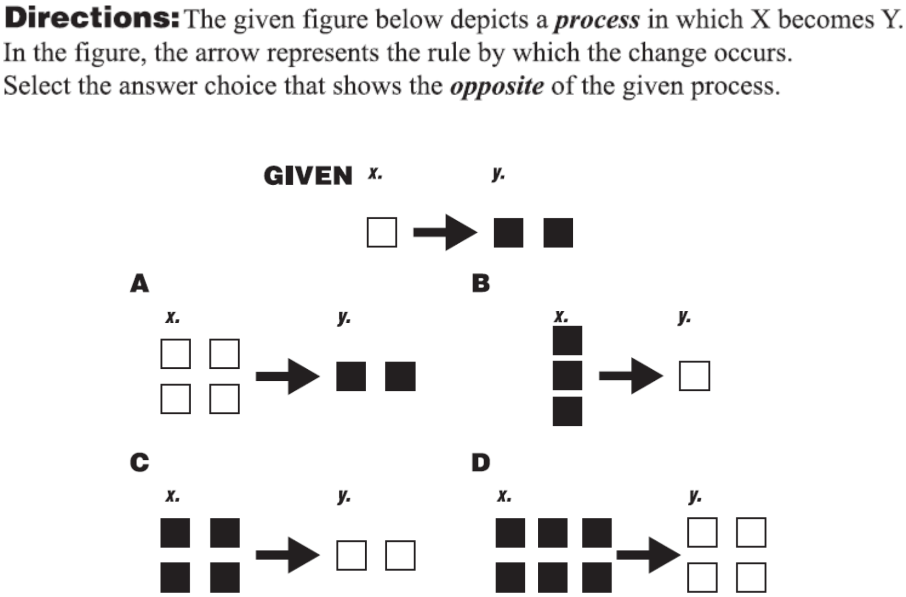
Directions:

- **The problems in this section ask you to compare sets of objects that vary in certain features.**
- **Each set has a specific rule that decides what objects can be included in that set. Some of the objects included in each set are pictured, enough to allow you to determine its rule for inclusion.**
- **Every problem asks you to identify which ONE of the four sets that are shown could NEVER have an object in common with the Given set, based on the compatibility of their rules for inclusion.**
- **There will always be EXACTLY ONE set that is incompatible with the Given set.**

For this sample problem, the rule governing the given set allows differing shapes to be included, as long as they are of the same, designated color (i.e., gray). Options A, B, and C are comprised of one shape each (hexagon, circles, and diamonds, respectively), but those shapes are of varying fills. In this sample item, each of these options includes a gray shape corresponding to one in the given set. Thus, Options A, B, and C, while different from the given set, can have a member in common with the given and are, thus, not incompatible. Only Option D has a rule for inclusion that is incompatible with the given set. Because Option D can only include dotted shapes, it could never have a member in common with the given set, marking it as the correct choice. In effect, membership in set D precludes membership in the given set.

Antithesis. An antithesis is theoretically described as directly oppositional concepts (e.g., Bianchi, Savardi, & Kubovy, 2011). Therefore, the items on the antithesis scale of the TORR were designed so that the correct choice would have a relation of opposition to the given figural array. To achieve this, the given array was created to depict a process where one figure (labeled "X") is changed into another figure (labeled "Y"). To correctly solve the item, a participant must ascertain what process has taken place in the given, and then select the option that depicts the opposite of that given process. A sample problem and its directions from the antithesis scale of the TORR follows:

Figure 8. A sample antithesis item from the TORR



Directions: The given figure below depicts a ***process*** in which X becomes Y. In the figure, the arrow represents the rule by which the change occurs. Select the answer choice that shows the ***opposite*** of the given process.

In this sample item, the process being depicted in the given array is a doubling of the number of squares and a changing of the color of the squares from white to black. Answer choice C depicts the antithesis of the given process because, in process C, the number of squares is being halved and the color is being changed from black to white. In order to select the correct answer to an antithesis item, a participant is required to reverse the process depicted in the given figure mentally, and then select the answer choice that depicts that oppositional process.

Retrospective Interviews

After the initial item creation process, feedback about the TORR items from experienced reasoners was required in order to judge whether or not the items were eliciting the appropriate mental processes in participants. Therefore, in-depth retrospective interviews were utilized. A small select sample (N=5) of skilled thinkers was recruited that included advanced graduate students in science education (n=2), mathematics education (n=2), and a professor in human development (n=1) from the University of Maryland. These individuals were invited to participate because of their demonstrated capacity to think and reason, as well as their ability to explain their reasoning processes in educational and psychological terms.

Before being interviewed, each participant completed the TORR silently. Then, they verbally explained to a researcher how they had come to select their answer choice for each item. Where a participant's answer choice differed from what the researcher had expected, the researcher and participant discussed the discrepancy. Moreover, if the participant experienced any difficulties in comprehending the scale directions, or if the format of the items or the whole test was confusing to them, that difficulty was noted and discussed. Then, after discussion had ended, cognitive lab participants responded to a series of questions concerning their perceptions of the TORR and its scales. For example, participants were asked to generate a possible title for each of the scales as well as for the test as a whole. Questions such as, "if you were to give this test to a student, what do you think it would tell you about them?" and "how similar or different do you think these scales are?" were also posed to participants.

Data gathered during this stage of test development was used to identify features of the draft TORR items that did not function in the way they were intended to function, ultimately leading to item revision and deletion. For instance, in some cases, participants would point out that two of the available answer choices for an item were potentially correct. In these situations, the answer choices were revised to eliminate the ambiguity. If a problem identified in a draft item was deemed too large to be effectively responded to through revision, items in the initial pool were also deleted and replaced before being again brought to the cognitive lab participant for feedback. In this way, item revision associated with the retrospective interviews continued, and the calibration phase of the investigation was not initiated, until there was agreement among the cognitive lab participants that only one of the answer choices on the TORR was logically keyable.

Moreover, participants' answers to the questions posed during the interviews were also reviewed to determine whether the TORR and its scales appeared to measure the constructs of interest. Overall, participants described the TORR as being comprised of four distinct but related scales, each requiring participants to reason differently about graphical figures and their associations. For example, one participant gave the title, "test of complex patterns" to the TORR as a whole, while another named it the "fluid reasoning test." Given the guiding postulations and definition of constructs that guided the creation of the TORR, this feedback was interpreted as evidence that the TORR was beginning to accomplish its purpose. In terms of the individual scales of the TORR, the analogy scale was generally described as focusing on relations of similarity and included such titles as "find the match" and "complete the pattern." For the anomaly scale, titles like "one of these things is not like the other" and "odd-one-out," captured the relation key to its formulation. Titles like “mutually exclusive sets” and “shapes that do not fit” drew on the relation of incompatibility for the antinomy scale.

Finally, labels like “opposites” and “reverse the rule” were offered for the antithesis scale. In general, this feedback was taken to show that the scales were discernably different to participants, and that their differences were based on the relations being reasoned with. Indeed, cognitive lab participants described the scales of the TORR as being discrete but complementary. One participant described the scales as “different pieces of the puzzle” of the entire test, and went on to point out that a single scale could provide an educator with information about “a student’s ability to reason with a particular pattern,” but the entire test would “see if they were able to reason with a variety of patterns.” It should be noted that only when we were satisfied that the scales and the items were functioning well for our select sample of cognitive lab participants did we move forward with the calibration of the TORR.

Also based on the feedback of the retrospective interview participants, the particular length, in terms of numbers of items, of the scales of the TORR was decided upon. In general, the number of items should be high enough to meaningfully sample from the possible levels of item difficulty that are of interest. Moreover, because classical test-theory reliability coefficients (i.e., Cronbach’s alpha) are still commonly used by substantive researchers who may make use of the TORR, the TORR scales should have enough items so that it may achieve adequate CTT reliability. However, these length requirements must be balanced with considerations of participant cognitive fatigue, and challenges related to administering lengthy measures to a large numbers of participants. Therefore, in order to balance these various pressures, the decision was made to include eight items per scale of the TORR, for a total of 32 items on the entire test. Moreover, because of the high novelty of the stimuli on the TORR, and the complexity of the cognitive processes required for relational reasoning, two low-difficulty sample items were constructed for each scale, in order to familiarize participants with the structure of the items. As a way to improve the likelihood of them being able to meaningfully reason with the scored items, participants are given the correct response to the sample items immediately after selecting an answer choice. In this way, specific choices about the construction of the TORR were made in

order to maximize the likelihood that participants could provide their best responses to TORR items.

Calibration

In order to ascertain the internal structure of the TORR, as well as the parameters of each of the TORR items, a CTT and MIRT calibration was undertaken. The calibration was completed with a large (N = 1379) representative sample of the University of Maryland. Full demographic information on this sample, as well as chi-square tests for representativeness, are available in Table 2.

Table 2

Demographics and Representativeness of Calibration Sample

Variable	Group	N	Percentage of Sample	Percentage of University Population	χ^2 test
Gender	Male	700	50.76	53.49	χ^2 =.14, df = 1, *p* =.70
	Female	679	49.23	46.51	
Ethnicity	White	712	51.71	52.22	χ^2 =6.86, df = 6, *p* =.33
	African American/Black	256	18.56	12.82	
	Hispanic	173	12.54	9.21	
	Asian	190	13.77	15.94	
	Native American/ Islander	0	0	0.12	

	Other	31	2.24	9.69	
	Prefer not to respond	17	1.23	n/a	
First Language	English	1204	87.31	86.21	χ^2 =.05, df = 1, *p* =.82
	Not English	175	12.69	13.79	
Major Domain	Arts/Humanities	126	9.13	11.23	χ^2=1.60, df = 5, *p* =.91
	Business	116	8.41	10.44	
	Social Sciences	450	32.63	25.63	
	Natural Sciences/Mathematics	290	21.02	21.61	
	Engineering	214	15.52	14.82	
	Undecided undergraduate studies	183	13.27	16.27	
Level	Freshmen	122	8.84	12.59	χ^2=1.59, df = 4, *p* =.81
	Sophomore	265	19.21	21.19	
	Junior	355	25.74	25.86	
	Senior	398	28.86	27.95	
	More than 4 years	239	17.33	12.41	

A variety of statistics and parameters related to the functioning of TORR items were calculated from the data collected from this sample. For example, the correlations among the observed scale scores of the TORR are available in Table 3.

Table 3

Correlations Among the Scales of the TORR

	TORR	Analogy Scale	Anomaly Scale	Antinomy Scale	Antithesis Scale
TORR	1.00				
Analogy Scale	.79**	1.00			
Anomaly Scale	.78**	.52**	1.00		

Antinomy Scale	.65**	.32**	.33**	1.00	
Antithesis Scale	.76**	.48**	.51**	.30**	1.00

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

As can be seen, each scale correlates in the strong-positive direction with the TORR total score, and the strengths of the correlations among the scales themselves are positive, but more moderate in strength. As another indicator of these strong correlations among the elements of the TORR, the classical reliability statistic Cronbach's alpha was calculated to be $\alpha = .84$, which is especially high for a figurally presented reasoning test, which tend to have lower reliabilities than verbal measures (Atkins et al., 2014; Senturk, Yeniceri, Alp, & Altan-Atalay, 2014).

Classical test theory. Item analysis based on classical test theory was also undertaken at this stage of TORR development. Classical test statistics for each item on the TORR, including difficulty and discrimination values, can be found in Table 4.

Table 4

Classical TORR Item Statistics

			Full-Instrument	Scale-Specific
Scale	Item	Difficulty	Discrimination	Discrimination
		p	*Item-Total*	*Item-Scale*
			Correlation	*Correlation*
Analogy	1	.51	.44	.52
	2	.36	.47	.63
	3	.67	.43	.53
	4	.35	.52	.60
	5	.64	.39	.56
	6	.40	.38	.53
	7	.58	.40	.55
	8	.36	.52	.61
Anomaly	1	.57	.31	.45
	2	.76	.32	.41
	3	.43	.44	.56
	4	.46	.51	.61

	5	.63	.47	.58
	6	.55	.48	.57
	7	.43	.44	.53
	8	.37	.36	.47
Antinomy	1	.76	.25	.57
	2	.62	.51	.33
	3	.47	.41	.70
	4	.60	.40	.67
	5	.55	.34	.56
	6	.43	.18	.54
	7	.39	.31	.38
	8	.54	.36	.48
Antithesis	1	.32	.38	.48
	2	.46	.31	.41
	3	.79	.37	.51
	4	.55	.29	.50
	5	.37	.53	.43
	6	.67	.53	.65
	7	.60	.47	.53
	8	.68	.45	.56

In terms of classical difficulty, all of the items on the TORR fell between .3 and .8, with 15 out of the 32 items being within a tenth of .50. This range of scores is widely considered to be ideal from a classical test theory perspective, because it maximizes the variability in participant scores (Wainer & Thissen, 2001), and therefore maximizes the covariance the TORR can potentially have with other related measures. As a case in point, Anomaly 7 exhibited a classical difficulty of .43, just 7 hundredths less than an ideal value of .50. In terms of a classical discrimination index, nearly all of the TORR items displayed strong-moderate item-total correlations, with only a small minority of items displaying moderate or weak-moderate item-total correlations. Based on existing measurement literature within a classical test theory tradition (e.g., DeVellis, 2006) these item-total correlations indicate that individual TORR items are generally associated with relational reasoning ability, as indicated by the TORR total score.

MIRT. Then, because classical test theory item statistics have a number of shortcomings, including the unavailability of a guessing parameter and an inability to accommodate multidimensionality, a MIRT calibration was undertaken. Before generating item parameters, however, a MIRT model that appropriately described the multidimensionality of the TORR needed to be selected. To do this, the fit of three theoretically plausible models to the data were compared. Importantly, each of these MIRT models was three-parameter logistic (3PL). Therefore, discrimination, difficulty, and guessing parameters were estimated for each of the items. Because the TORR is a selected-response test in which participants can potentially guess answers, each of these parameters is potentially informative about the functioning of a given item.

Model comparison. The first model that was compared was a unidimensional model in which a general relational reasoning factor loaded on all 32 items. This model is the same as traditional 3PL IRT models that do not account for multidimensionality. If this model fit best, then the TORR could be described as a unidimensional test in which all 32 items tap a single latent relational reasoning ability. The second model that was fit to these data were a correlated factor model, in which the items from each scale of the TORR loaded on their own specific latent ability, and these factors were free to covary. If this model were to fit best, then the TORR could best be described as tapping four distinct yet related abilities. As such, scoring would only validly take place at the scale level, and students would be evaluated based on their profile of analogical, anomalous, antinomous and antithetical reasoning abilities, but not on their general relational reasoning ability. The mathematical definition of the item parameters estimated by the unidimensional model and the correlated factor model can be thought of as largely equivalent to equation one, because both of these models define the cause of variability in each of the individual TORR items to be a single latent ability. That is to say that, despite the positing of four distinct abilities by the correlated factor model, each item only loads on one ability. Therefore, although the correlated factor model does describe the TORR itself as multidimensional, each item within that model is only loading on one of those dimensions.

In contrast, the third model that was fit to these data were a bi-factor model, in which a general relational reasoning factor loaded directly on all 32 of the items and four specific factors (i.e., analogy, anomaly, antinomy, and antithesis) loaded on their corresponding scales. Therefore, each item loads on two latent abilities—general relational reasoning ability, and the specific form (e.g., analogy) that is associated with that item. Therefore, a greater number of item parameters are estimated in this model than in either the unidimensional model or the correlated factor model. Specifically, while the former two models estimate three parameters—difficulty, discrimination, and guessing— for each item, the bi-factor model estimates five: guessing, general discrimination, specific discrimination, general difficulty, and specific difficulty.   These parameters are defined within the following logistic equation:

$$P(x_{ij} = 1 | \theta_i, \xi_i) = c_j + \frac{1 - c_j}{1 + e^{-a_j(\theta_i + \xi_i - b_j)}} \qquad (2)$$

As can be seen above, the probability of a participant responding correctly to a given item (i.e., having it coded as “1”) depends on that participant’s general (θ) and specific (ξ) abilities. It should be noted that the Greek character ξ is used here to represent the specific ability on which an item loads. In this way, the bi-factor model has the advantage of simultaneously modeling general relational reasoning ability as well as accounting for the dependencies among scale-items with specific factors. As such, the bi-factor model allows students to be assessed in terms of their general relational reasoning ability, while also supplying information on their profile of analogical, anomalous, antinomous, and antithetical reasoning abilities.

Although the conditional independence of item responses is a hallmark assumption of many psychometric models, including unidimensional IRT, the particular formulation of the bi-factor model allows for a relaxation of that assumption. Specifically, the conditional independence assumption of the bi-factor model pertains to all latent variables within the model simultaneously, meaning that the covariance among the items is assumed to be due entirely to those latent variables that are posited in the model (Cai, Yang, & Hansen, 2011). However, one strong assumption that does hold for the bi-factor model is the normal distribution of latent traits being measured. Further, an assumption of orthogonality of the specific ability factors in the bi-factor model ensures model identification (Rijmen, 2009). Another important assumption of all single-group MIRT models is that item parameters are invariant across sub-groups within the same population (i.e., no DIF). Of course, it is this very assumption of the bi-factor model that is the subject of this dissertation.

Relevant fit statistics corresponding to each of our tested models are available in Table 5.

Table 5

Limited Information Fit Statistics for Model Comparison

Model	-2Loglikelihood	AIC	BIC	Limited information RMSEA	M_2*	df
Unidimensional	53054.73	53246.73	53748.72	0.065	2323.65	432
Correlated Factor	52411.30	52615.30	53171.97	0.058	1798.16	426
Bi-Factor	52244.09	52516.09	53227.25	0.052	1664.04	392

This discussion of fit is greatly abridged, and for a full discussion, see Dumas and Alexander (2015) Among the available fit statistics, -2loglikelihood, Akaike information criterion (AIC), Bayesian information criterion (BIC), as well as the limited information RMSEA and M_2^* are available in this table. Because each of these statistics is a measure of misfit, smaller values indicate better fit throughout Table 5. As can be seen, the unidimensional model exhibited the worst fit in terms of each of the available fit statistics. Although the correlated factor model fit significantly better than the unidimensional model, it was not the best fitting model of the ones being compared. The bi-factor model exhibited lower values for the -2loglikelihood, AIC, and M_2. However, the correlated factor model did fit better than the bi-factor in terms of BIC, probably because this fit statistic placed much weight on parsimony, especially at large sample sizes, and the correlated factor model retained 34 more degrees of freedom then the bi-factor model did. However, the limited information M_2^* statistic, which was specifically formulated for the MIRT context (Cai & Hansen, 2013), showed a strong preference for the bi-factor model.

Therefore, based on the available fit statistics, the bi-factor model should be retained for calibration. It should also be noted that the bi-factor model corresponds more closely to the theoretical structure of the construct of relational reasoning, because it allows for the measurement of the general dimension that is a major gap in the relational reasoning literature (Dumas et al., 2013). Practically speaking as well, a general relational reasoning score may be the most meaningful for interpretation in the educational context, because it allows practitioners a general understanding of a students' relational reasoning ability, before getting into the student's profile of abilities with specific forms of relational reasoning. For these theoretical reasons, as well as the statistical evidence already examined, the bi-factor model was retained for calibration of the TORR, and used to produce item parameters.

Another important difference between the models fit during this stage of the investigation is the meaning and interpretation of the specific ability factors (i.e., analogy, anomaly, antinomy, and antithesis). For example, the correlated factor model posits four specific abilities causing variation in the TORR items, and no general abilities. Therefore, in that model, the specific abilities, and the covariance between them, would be the focus of a calibration and scoring procedure. However, in the bi-factor model, the specific ability factors represent the cause of variation in a given TORR item that are *not* relational reasoning. Therefore, determining what precisely the specific ability factors represent in the bi-factor model is a complex theoretical exercise. For example, such factors may measure the ability of a participant to mentally manipulate the specific visual stimuli of a given scale, or represent a participant's familiarity with logical exercises similar to that of a particular scale of the TORR. Therefore, the bi-factor model is inherently concerned with the reliable measurement of relational reasoning ability in general, and not with the associated specific abilities. In previous calibration work (Dumas & Alexander, 2015), the recommendation was forwarded that, should a specific form of relational reasoning (e.g., analogy) be of interest in research, that a single-factor scale-specific calibration and scoring procedure be followed. Such a procedure inherently changes the focus of measurement from general relational reasoning to a specific sub-form of the construct. Indeed, the observation that the bi-factor model relegates the specific ability factors, compared to the general factor, to a subordinate role, is an important one to communicate to researchers interested in using the TORR. However, potentially equally important is that the bi-factor model produces a highly reliable general ability factor, free from the conditional dependence associated with the scale-specific TORR stimuli.

Relatedly, the bi-factor model requires that the specific ability factors must be orthogonal to the general ability factor. This means that the specific abilities, as measured by the bi-factor model, are unrelated to the general ability being measured. So, in the case of the bi-factor model, the specific abilities that are labeled analogy, anomaly, antinomy, and antithesis should not be interpreted as instantiations of the general construct of relational reasoning, but rather as separate abilities—not relational reasoning—that support students' likelihood of getting a particular item correct. Also, in the bi-factor model utilized in this investigation, the specific ability factors are also kept orthogonal from one another. Although this choice was not mathematically required for the bi-factor model to be identified, it follows common practice in the applied psychometric literature, in which specific abilities measured by the bi-factor model typically do not covary with one another (e.g., Patrick, Hicks, Nichol, & Krueger, 2007). That being said, there may be reason to believe that, theoretically, the specific abilities associated with each of the scales of the TORR do covary. For example, the same visuo-spatial, test-taking, or reading strategies used for the interacting with the anomaly items and directions may also be used for other types of items as well. For that reason, allowing the specific ability factors of the TORR bi-factor model to covary may be an interesting next step in this line of research. However, in order to follow commonly accepted practice in the literature, the specific ability factors were not allowed to covary in this study. As such, interpretation of the relations among the specific residualized factors will not be possible, because the covariances among those factors were not estimated. In the future, uncovering such relations among the specific factors—although their interpretation will be necessarily limited because of the residualized nature of the factors themselves—may be motivation for including covariance among the specific abilities in the bi-factor model.

Item parameters. In Table 6, general dimension difficulty and discrimination, specific dimension difficulty and discrimination, as well as the guessing parameter that is shared between the general and specific dimensions, are available.

Table 6

MIRT Parameters for TORR items based on the bi-factor model

Scale	Item	General Dimension Parameters *(SE)*			Specific Dimension Parameters *(SE)*	
		Guessing	Discrimination	Difficulty	Discrimination	Difficulty
Analogy	1	.23 (.04)	1.53 (.28)	0.16 (.02)	0.45 (.09)	0.54 (.08)
	2	.07 (.02)	1.65 (.15)	0.36 (.04)	1.03 (.21)	0.57 (.11)
	3	.28 (.04)	1.87 (.34)	-0.52 (.06)	0.35 (.07)	-2.77 (.36)
	4	.09 (.02)	1.66 (.26)	0.48 (.07)	0.84 (.11)	0.94 (.17)
	5	.20 (.05)	1.30 (.30)	-0.58 (.07)	1.24 (.15)	-0.61 (.12)
	6	.22 (.03)	1.58 (.33)	0.75 (.09)	0.81 (.09)	1.46 (.19)
	7	.28 (.04)	1.40 (.37)	-0.07 (.01)	1.89 (.23)	-0.05 (.04)
	8	.09 (.02)	2.19 (.60)	0.37 (.04)	1.08 (.14)	0.75 (.10)
Anomaly	1	.43 (.02)	2.14 (.34)	0.71 (.08)	0.20 (.03)	.75 (.20)
	2	.20 (.08)	0.96 (.16)	-1.19 (.18)	0.64 (.10)	-1.78 (.22)
	3	.17 (.03)	2.00 (.30)	0.43 (.06)	0.96 (.13)	0.89 (.15)
	4	.12 (.02)	2.23 (.34)	0.11 (.01)	0.20 (.04)	1.22 (.18)
	5	.11 (.03)	1.53 (.18)	-0.65 (.09)	0.80 (.11)	-1.24 (.19)
	6	.11 (.04)	3.01 (.60)	-0.26 (.03)	2.88 (.40)	-0.27 (.08)
	7	.17 (.03)	1.58 (.31)	0.52 (.06)	0.24 (.04)	3.42 (.46)
	8	.20 (.03)	1.30 (.28)	1.10 (.13)	0.20 (.05)	.71 (.11)
Antinomy	1	.21 (.05)	1.35 (.23)	-1.47 (.17)	1.54 (.19)	-1.28 (.16)
	2	.55 (.02)	4.51 (.58)	1.54 (.19)	1.75 (.23)	3.16 (.39)
	3	.13 (.01)	3.07 (.52)	0.11 (.02)	3.39 (.45)	0.09 (.02)
	4	.24 (.02)	1.88 (.48)	-0.22 (.02)	3.12 (.38)	-0.13 (.04)
	5	.17 (.04)	1.17 (.29)	-0.08 (.01)	0.94 (.13)	-0.09 (.03)
	6	.13 (.03)	0.84 (.12)	0.81 (.10)	0.89 (.09)	0.76 (.12)
	7	.24 (.04)	0.44 (.25)	3.03 (.32)	0.68 (.05)	1.91 (.23)
	8	.24 (.05)	1.11 (.26)	0.48 (.06)	0.93 (.09)	0.57 (.08)
Antithesis	1	.13 (.03)	1.32 (.30)	1.23 (.15)	0.90 (.12)	1.80 (.22)
	2	.22 (.04)	2.13 (.37)	0.65 (.08)	0.47 (.04)	2.94 (.36)

3	.41(.06)	1.18 (.20)	-0.81 (.10)	1.46 (.18)	-0.65 (.09)
4	.20 (.05)	1.51 (.21)	0.04 (.01)	1.41 (.13)	0.04 (.03)
5	.31 (.01)	6.95 (.78)	1.87 (.16)	0.79 (.11)	.16 (.08)
6	.17 (.03)	3.48 (.49)	-0.55 (.07)	2.37 (.30)	-0.81 (.10)
7	.25 (.04)	2.31 (.35)	-0.13 (.02)	0.73 (.08)	-0.41 (.07)
8	.18 (.05)	1.43 (.40)	-0.70 (.08)	0.87 (.12)	-1.15 (.14)

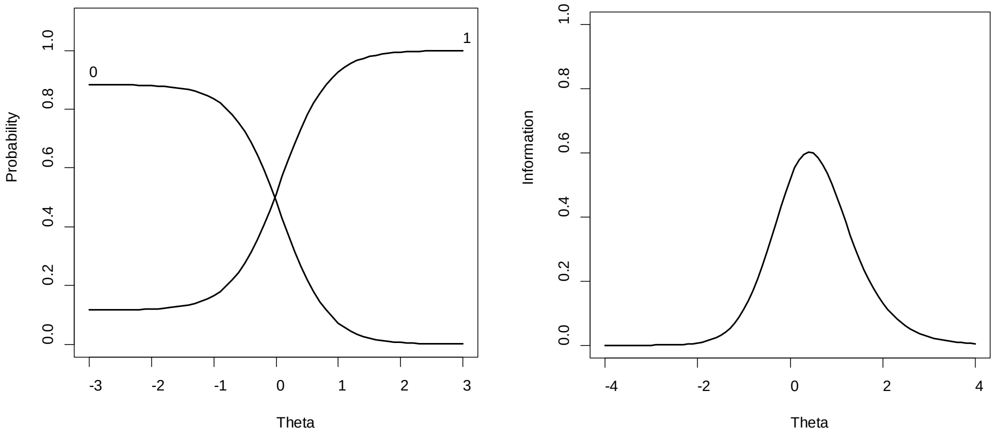
In terms of these guessing parameters, because each TORR item has four answer choices, an expected guessing parameter for each item would logically be .25. However, most items (25 out of 32) exhibited empirically derived guessing parameters that were lower than .25, suggesting that TORR items did not generally contain easily-guessed answers, and that it was not often possible for low-ability participants to narrow down their answer choices past the initially available four.

In terms of discrimination parameters, the TORR items tended to have strong discrimination parameters in terms of the general relational reasoning factor. Indeed, 29 of the 32 items had discrimination parameters stronger than one, 11 had discrimination parameters stronger than two, and five items had discrimination parameters stronger than three. These findings suggest that the TORR items are strongly related to general relational reasoning ability, and that the items are effective at separating those participants with low relational reasoning ability and those with higher ability. Interestingly, for the majority (i.e., 25 out of 32) of TORR items, the discrimination parameter associated with the general dimension was stronger than the parameter associated with the specific dimension. This is by design, because the bi-factor model aims to account for the covariance among the items first with the general factor, and the specific factors are mainly intended to account for residual covariance and dependency among the items on each specific scale

Difficulty parameters of the TORR items on the general relational reasoning dimension exhibited a healthy spread with a minimum of -1.47 (Antinomy 1), and a maximum of 3.03 (Antinomy 7). This finding suggests that TORR items are variable in their difficulty, and as such may produce a large amount of variability among participants. Moreover, a large majority (i.e., 25 out of 32) of the general dimension difficulty parameters fell between -1 and 1, suggesting that a large proportion of TORR items require between 1 standard deviation below and 1 standard deviation above average general relational reasoning ability in order to be correctly answered. Interestingly, for 25 out of 32 TORR items, specific dimension difficulty parameters were more extreme than general dimension difficulty parameters. That is to say, if an item was relatively difficult in terms of general relational reasoning ability, it was typically even more difficult in terms of specific ability, and vice versa. This finding logically follows from the discussed finding that most items had stronger discriminations on the general dimension, because discrimination parameters form the denominator of the formula for difficulty parameters.

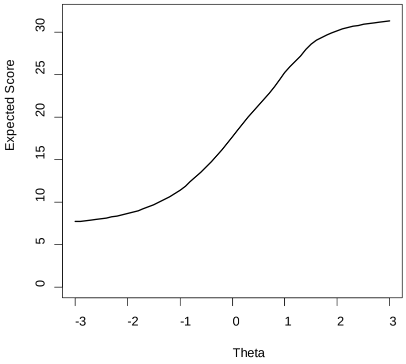
Characteristic curves. It should be noted that because the calibration model used in this investigation is multidimensional, including each of the dimensions in the model would yield multidimensional item-characteristic surfaces. However, because the general relational reasoning factor is of principal interest here, and to ease the visual interpretation of the data, unidimensional curves that solely focus on the general relational reasoning dimension are presented. An example ICC from a proto-typical item (i.e., Anomaly 7) is available in Figure 9.

Figure 9. Item characteristic curve and item information function for Anomaly 7 pertaining to the general relational reasoning dimension of the TORR.



Here, the curve labeled “1” represents the probability of a participant getting the item correct, and the curve labeled “0” represents the probability of getting the item incorrect. Focusing on the curve labeled “1”: the guessing parameter (.17) appears as the lower asymptote and the inflection point is the probability of 58% for getting the item correct. The difficulty parameter (.52) is the location of the inflection point on the theta axis, and the discrimination (1.58) is the slope or first derivative of the ICC at the inflection point. Taking into account the parameters of all 32 items on the TORR, a test-characteristic curve (TCC) was also constructed to describe expected TORR score of a participant, given their general relational reasoning ability level. The TCC for the general dimension of the TORR appears in Figure 10.

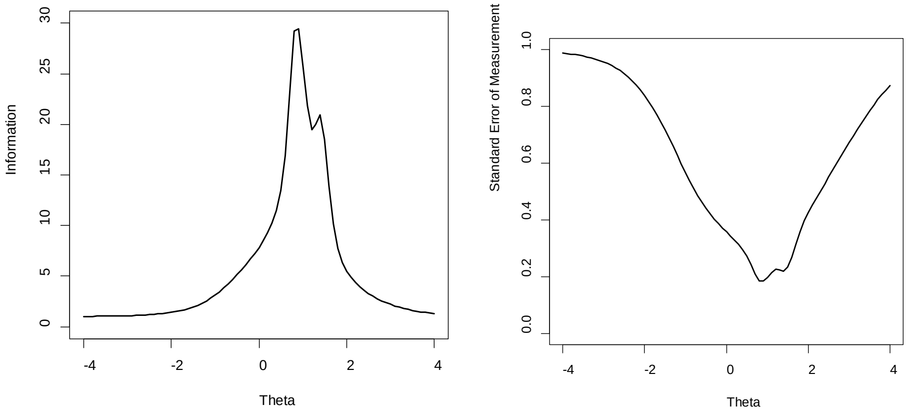
Figure 10. Test characteristic curve pertaining to the general relational reasoning dimension of the TORR.



Interestingly, the TCC rises steeply after a theta level of zero, implying the TORR as a full measure is effective at discriminating between participants who have average relational reasoning ability, and those who have above-average ability.

Information. The item information function (IIF) pertaining to the general relational reasoning dimension is available for Anomaly 7, in Figure 9. As can be seen, the item provides the greatest amount of information about a participant whose ability level matches the item's difficulty parameter, and the amount of information is related to the discrimination of the item. Moreover, the IIF's of each of the TORR items can be summed to produce the test information function (TIF) available in Figure 11.

Figure 11. Test information and conditional standard error of measurement curves pertaining to the general relational reasoning dimension of the TORR.



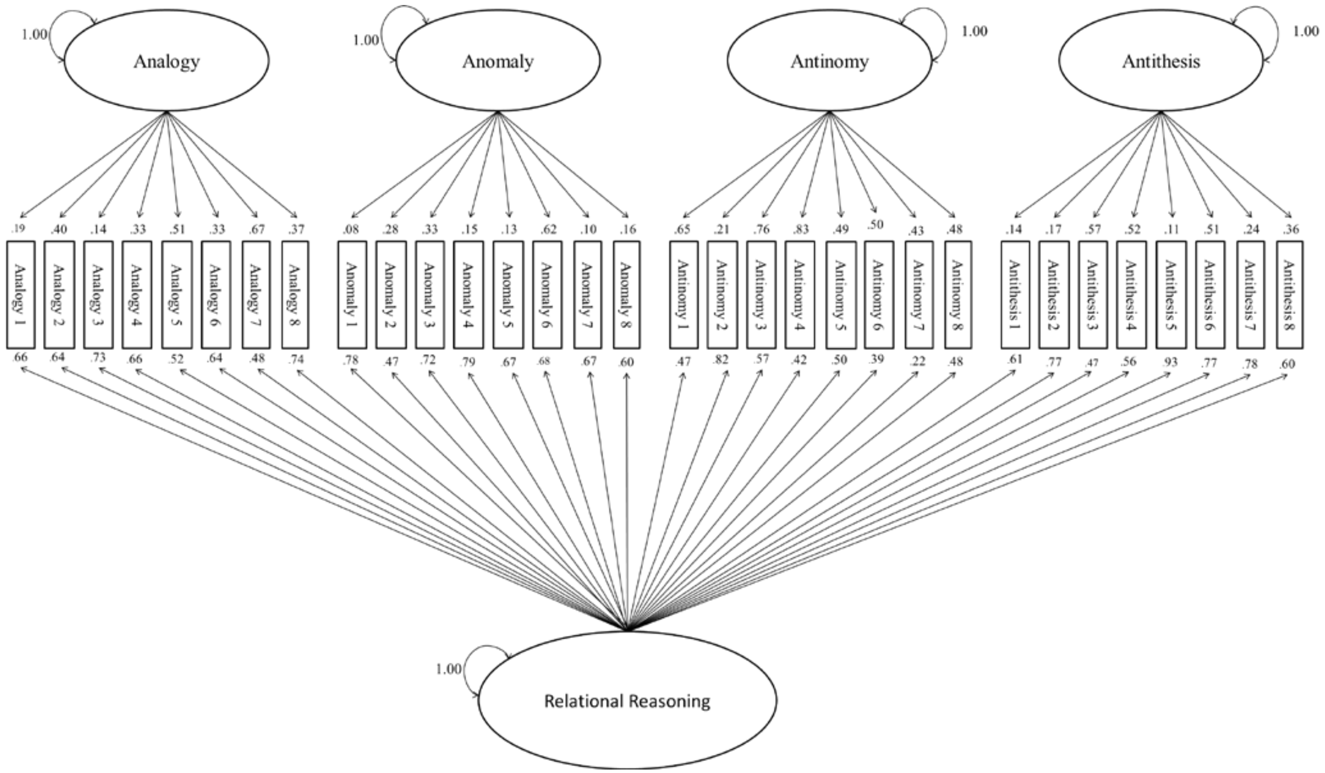
The TIF for the TORR shows that highest attained information (29.20) is found at a theta level of approximately .8. This finding suggests that measurement of general relational reasoning ability was most informative, and therefore contained the least error, when participants were approximately .8 of a standard deviation above the mean ability level. Also available in Figure 1 is the standard error of measurement curve, which is simply the inverse of the information function. The maximum information value attained by the TORR (i.e., 29.20) is high for a visuospatial test (Vock & Holling, 2008).

Construct reliability. One principal method for quantifying the stability or reproducibility of a latent variable is through its construct reliability. Construct reliability can be calculated via coefficient *H* (Hancock & Mueller, 2001). Coefficient *H* may be conceptualized as the proportion of variance in a latent variable that its indicators can account for if that latent variable were regressed on each of its indicators, similar to an R^2 value (Hancock & Mueller, 2001). Coefficient *H* can be expressed for a latent factor with *k* indicators and *a* standardized loadings as follows in equation 3:

$$H = \frac{\sum_{i=1}^{k} \frac{a_i^2}{(1-a_i^2)}}{1+\sum_{i=1}^{k} \frac{a_i^2}{(1-a_i^2)}} \tag{3}$$

The standardized bi-factor model, with all standardized loadings, appears in Figure 12.

Figure 12. Bi-factor model with standardized loadings.



Given these estimated standardized loadings, the general relational reasoning dimension exhibited a construct reliability of H = .96. This very strong level of construct reliability implies that latent relational reasoning ability is highly reproducible. The strength of the item loadings on the general dimension, and the fact that all 32 TORR items load on the general dimension contributed to this high reliability. Because of the nature of the bi-factor model, the construct reliability of the specific factors is expected to be much lower than that of the general factor. This is because, in the bi-factor model, the general ability factor loads on each of the items first, accounting for as much variance in those items as possible. Then, the specific factors load on their designated items, and account for the residual dependency and variance among those items. Because of this procedure, the standardized loadings of each item on the general factor of the bi-factor model are typically stronger than they are on the specific factors. This is by-design, in order to produce as reliable and reproducible a general factor as possible Therefore, in this case, the analogy specific factor had a construct reliability of H = .64, the anomaly specific factor H = .47, antinomy H =.84, and antithesis H =.59.

Scoring. One of the principal goals of this MIRT calibration of the TORR was to produce readily interpretable TORR scores for practical use in the educational context. Importantly, because the TORR is calibrated and scored in this study based on a large representative sample of our university's population, scores produced here coupled with an easy-to-use conversion between simple summed-scores and MIRT scores can allow TORR scores to be highly meaningful for educators and students even in single-subject scenarios. Toward that end, expected a posteriori (EAP) scores were produced for the general relational reasoning factor for every participant. It is important to note here that, in the MIRT context, calibration and scoring are not interchangeable terms, and denote separate steps of a larger process. More specifically, the calibration phase involves fitting a measurement model (e.g., the bi-factor model) to test data, interpreting data-model fit, and if the fit is good, producing item parameters and ICC’s for each item that describe the relation between that item and the latent ability being measured. After that process is complete, those item parameters are used to calculate scores, or estimates of how much latent ability a particular participant had during that testing occasion. In the IRT or MIRT contexts, every possible response pattern on a given test (about 4.3 billion possible response patterns exist on a 32 item test like the TORR) have a different score depending on the item parameters of the particular items constituting that pattern. A number of different methods are available for the calculation of response pattern scores in the IRT context, but when tests are multi-dimensional, Bayesian scoring methods, such as the EAP method used here, are the most commonly used (Brown, & Croudace, 2015)

As previously mentioned, unique EAP scores exist for every discrete response pattern (i.e., 4.3 billion on the TORR). However, in the educational context, it is not feasible to expect practitioners to account for each of these 4.3 billion possible patterns by scoring the TORR using the bi-factor MIRT model. Therefore, in order to support the interpretability of TORR scores, these EAP scores were averaged over each response pattern that yielded the same summed-score. For example, the EAP scores corresponding to all the response patterns that yield a total score of 17 were averaged, and the same was done for each summed-score possible on the TORR (from one to 32). Because MIRT models estimate abilities as continuous latent variables, EAP score estimates are also available for scores lower than any present in our dataset (i.e., below five).

However, because many practitioners are not familiar with the standardized EAP metric, EAP scores were linearly transformed allowing for placement on a more typically utilized scale. Specifically, EAP scores were multiplied by 15 and then added to 100 in order to produce scores with a mean of 100 and standard deviation of 15: a basic IQ-type metric. TORR EAP scores placed on this metric have been termed relational reasoning quotient (RRQ) scores. Table 7 holds a full listing of TORR summed scores, EAP scores, and scaled RRQ scores.

Table 7

Summed Scores, EAP scores, and RRQ Scores for the General Relational Reasoning Dimension of the TORR

Summed Score	EAP Score	RRQ Score
1	-2.05	69
2	-1.97	70
3	-1.90	71
4	-1.82	73
5	-1.73	76
6	-1.63	77
7	-1.52	79
8	-1.40	81
9	-1.27	83
10	-1.14	85
11	-0.99	87
12	-0.85	90
13	-0.69	91
14	-0.55	94
15	-0.41	96
16	-0.26	98
17	0.01	100
18	0.14	102
19	0.27	104
20	0.39	106
21	0.52	108
22	0.65	110
23	0.77	112
24	0.91	114
25	1.05	116
26	1.19	118
27	1.35	120
28	1.52	123
29	1.71	126
30	1.92	129
31	2.15	132
32	2.40	135

It should be noted that in order to facilitate interpretation, RRQ scores are rounded to the nearest unit. Interestingly, these scores range from 69 to 135, illustrating that the TORR can meaningfully measure the relational reasoning ability of students two standard deviations above or below the mean of the university population consisting of older adolescents and young adults. With the TORR effectively calibrated, its validity has also begun to be established through a variety of studies that utilize it as a predictive measure.

Previous Research Using the TORR

To date, the TORR has been fruitfully utilized to predict and understand a variety of cognitive and academic outcomes, establishing the validity and usefulness of the TORR in studying relational reasoning and associated educational and psychological variables. For example, in Alexander and colleagues' (2015) recent investigation, TORR scores were found to significantly predict scores on SAT released items, both for the verbal section [$F(1, 28) = 16.13$, $p<.001$; $\beta=0.36$, $t=4.02$, $p<0.001$; $R^2 = 0.37$] and for the mathematics section [$F(1, 28) = 4.34$, $p<.05$; $\beta=0.2$, $t=2.08$, $p<.05$; $R^2 = 0.13$]. Interestingly, while the TORR items are entirely visuospatial in nature, TORR scores predicted verbal SAT scores better than math SAT scores in this study. In the same investigation, the TORR also significantly correlated with the Raven Progressive Matrices at $r=.49$ ($p<.001$), which is to be expected given the similarity in the construction of the two measures, although the correlation is not so strong as to imply the measures to not account for unique variance as well. Moreover, the TORR correlated with the ShapeBuilder measure of visuospatial working memory (Sprenger et al., 2013) at $r=.31$ ($p=.02$), indicating that while working memory capacity may play a significant role in an individual's ability to correctly respond to TORR items, that construct does not account for an undue proportion of variance in scores.

Also in regards to working memory capacity, recent work (Grossnickle, Dumas, & Alexander, in revision) has sought to determine at exactly what point in the process of reasoning with a TORR item working memory capacity makes the greatest impact. Interestingly, the conditional probability of reaching a given stage in the reasoning process can correlate with working memory capacity as weakly as *r* =.1, and as strong as *r* =.57, indicating that limitations to working memory capacity produces "bottle-necks" in the relational reasoning process at certain points. This work is an example of how the TORR has been used to improve the field's understanding of the reasoning process and its cognitive requirements.

In a series of studies in the domain of engineering design (Dumas & Schmidt, 2015; Dumas, Schmidt, & Alexander, under review), the TORR was also used as a way to gauge the extent to which relational reasoning ability influenced design students' ability to produce innovative designs to solve an engineering problem. In these investigations, the TORR was a significant predictor of engineering design innovation ($\beta = .84$, $p = .01$) and a significant correlate of a commonly used measure of creative thinking ability, the Uses of Objects Task ($r = .37$, $p = .042$). These findings indicate that the TORR may be used as a predictive measure both for other domain-general cognitive abilities (e.g. working memory, creativity) as well across a wide variety of domain-specific academic variables (e.g. engineering design innovation). Indeed, other research groups within the field of educational psychology have begun to use the TORR across diverse lines of inquiry with promising results, and it is expected that the TORR will continue to gain popularity as a measure of relational reasoning (e.g., Kendeou & O'Brien, 2015).

Cultural-Fairness of the TORR

All of the previous work with the TORR, from its initial development and calibration, to the extant studies it has been utilized in, point to a reliable, valid, and useful measure for assessing relational reasoning in the university population. Indeed, the TORR has already begun to serve its intended purpose, and has been effective in supporting the field's ongoing effort to understand relational reasoning, and its role in education. At this stage, the TORR is nearly ready for widespread use in the educational setting. In fact, given the reliable, valid, and normative nature of TORR scores, it is conceivable that the TORR could currently be widely utilized to identify fluid relational reasoning ability in students. However, one critical phase of test development has yet to be completed for the TORR. Despite being developed with culture-fair assessment in mind, the TORR has yet to undergo a rigorous statistical examination of its actual cultural-fairness. Therefore, the goal of this proposed research is to investigate bias on the TORR among multiple gender, ethnic, and language groups. To do this, DIF must be tested for, and if found, determined to indicate or not to indicate bias. Based on existing literature, some predications can be made about what type of DIF, if any, may be found on the TORR, and under what conditions that DIF may indicate bias. These predictions are organized into three different types of groups for which it is proposed that DIF will be tested: gender, language, and ethnic groups.

Importantly, these groups are not the only types of groups that may be salient for DIF analysis. Also especially interesting may be socio-economic status (SES) groups, age groups, or groups of students who do or do not have particular identified disabilities. The three types of groups (i.e., gender, language, and ethnic groups) analyzed here were chosen as a reasonable sub-set of the universe of potentially meaningful groups that also have a strong basis in the extant literature for the analysis of DIF on measures of relational reasoning. What follows are particular literature-based hypotheses related to DIF within the groups chosen for analysis.

Gender and Sex

Although gender and sex have often been collected as a single variable in psychological research (Nye, 2010), they differ conceptually in important ways. Specifically, sex refers to the biological categories of male and female, typically defined by genitalia or hormone levels (Diamond, 2004). In contrast, gender refers to the socially defined categories man and woman, and includes the various culturally related differences in behavior associated with these categories (Kenneavy, 2013). This distinction is relevant in research related to group differences on visuospatial measures among males and females and men and women, because many researchers have observed group level differences and measurement invariance among these groups, but debates continue about the source of those differences that are potentially confounded with the variables that were measured. Specifically, on visuospatial reasoning measures, males/men have historically outperformed females/women (Nazareth, Herrera, & Pruden, 2013). Some researchers (Auyeung, et al., 2011; Vuoksimaa, et al., 2012) have argued that biological differences, including testosterone levels, drive the discrepancy between the groups, while others (Miller & Halpern, 2014; Pontius, 1997) contend that differential opportunities to develop visuospatial abilities early in life create the differences.

Because of these open questions, participants who identify as male/women or female/man, known as transgender individuals, complicate the research endeavor because it is often unclear, without collecting much more detailed biological or qualitative data, what an individual’s gender related socialization experiences or hormone levels are. Such biological investigations are important future directions in the field of visuospatial cognitive processing.

However, because transgender individuals do appear to be statistically rare in most populations (Xavier & Bobbin, 2005), investigations concerned with more generalizable differences in visuospatial ability among more common participants typically do not pose explicit research questions concerning them. Unfortunately, because sex/gender are often collected as a single variable, it is usually not possible to identify transgender individuals in a dataset. Therefore, more recent research collects each variable separately (Nye, 2010). In this proposed study, all transgender individuals in the data, if any, may be removed from the more general DIF analysis and saved for potential follow-up investigations. For this reason, the terms males/men and females/women may be used relatively interchangeably in the discussion.

In terms of the more general analysis of DIF on TORR items among men and women, it is necessary that DIF be interpreted in terms of the existing literature on visuospatial reasoning differences. For example, DIF in difficulty parameters indicating that certain items are less difficult for men than they are for women would fit with the current literature, although DIF in discrimination or guessing parameters may represent new information for the field. For example, if males are better at guessing on TORR items than females are, that may explain some observed mean differences in the literature. Moreover, if the discrimination parameters contain DIF, that may indicate meaningful differences in the way relational reasoning ability is distributed through each population, or differences in the way those abilities are organized in the minds of men and women.

Perhaps an even more important question is whether or not DIF on TORR items in regards to gender, if any is found, constitutes bias. Based on the large literature concerning sex differences in visuospatial rotation, this possible question may hinge on whether or not the TORR should be described as a measure of “relational reasoning” or “visuospatial relational reasoning." If, for example, the more general terminology of "relational reasoning" is utilized, than variance in scores due to visuospatial ability may be considered construct irrelevant, and therefore be seen as an indicator of bias. However, if the more specific term "visuospatial relational reasoning" is used, then variance in scores attributable to differences in visuospatial ability—regardless of whether they are biologically or socially produced—would not be an indicator of bias. Indeed, if the TORR is designated as a measure of "visuospatial relational reasoning" it could be fruitfully utilized as a measure within the large and growing literature on sex differences in visuospatial ability, and might even be used as a training tool in intervention studies designed to improve the visuospatial processing of females, and possibly demonstrate that formerly observed differences were culturally determined and not biological.

Moreover, because the TORR has been calibrated using the bi-factor model, it could be that the specific ability factors may be functioning as dimensions related to the particular visual aspects of each of the items. Therefore, we may expect DIF to be more likely in the specific ability factors than in the general relational reasoning factor. Because score estimates on the general relational reasoning dimension are currently the only scores utilized in calculating the RRQ, this property of the bi-factor model may help to prevent DIF or bias from entering RRQ scores, by using the specific factors to account for abilities that may be differentially distributed across groups. For this reason, DIF on specific ability factors may be less likely to be considered bias, because one major purpose of specific ability factors in the bi-factor model is to account for problematic properties of items such as residual dependency, and DIF could be conceptualized in the same way.

Language

Although TORR items do not contain a heavy proportion of language use, the language background of a participant may still affect scores and item parameters. One major way in which language background can affect scores on visuospatial tests is through the understanding of the directions (Richards, 2007), which on the TORR, differ by scale. Moreover, language background can deeply influence a student's ability to learn in US primary and secondary schools, thus altering their general cognitive abilities when they arrive at university (Freeman, 2012). Interestingly however, recent examinations of differences in scores on cognitive measures among language groups that utilized university samples have typically found that those participants whose first language was not English scored higher (e.g., Jiang et al., 2015; Ratiu & Azuma, 2015). While this was at first a surprising result, it should be noted that in university samples, those students whose language background is not English are much more select in terms of their cognitive abilities than in the general American population. Unfortunately, having a language background that is not English may put a student at risk for not going to college (Burke & Sunal, 2010), and therefore, those who do attend the university are a select sample, with potentially higher-than-average cognitive ability (Kroll & Fricke, 2014). In contrast, those university students whose first language is English are much more variable in terms of their ability, and may score lower on average. Further, it has been shown that those students who take assessments in their second language typically attend to them more closely, and are more likely to carefully read directions before proceeding to the test itself (Bialystok & Martin, 2004; Koo, Becker, & Kim, 2014).

Therefore, DIF that generally favors students whose second language is not English may be less surprising, and potentially less problematic than DIF that generally favors students whose first language is English. This is because, for any test of cognitive ability, constructs such as attention or willingness to read directions must necessarily be considered relevant, because they are required of all students in order to complete the measure with any degree of accuracy (Pashler, 1998). Potentially more problematic on the TORR would be DIF in difficulty parameters that generally favor native English speaking students. This is because the TORR was constructed from the beginning to limit the amount of crystallized ability, including reading that would be necessary to correctly respond to the items. However, for example, if native English participants were more adept at guessing the items because of test-taking strategies, that could be a potentially interesting finding for the field. DIF favoring native English speaking students would be more likely than other DIF to be considered bias, because reading and language ability has never been considered a relevant construct to the TORR. For this reason, if such DIF is observed, revision of item directions may be necessary to make them easier to read for second language students. However, based on the previous literature, a more likely finding is that those students whose first language is not English will attend more carefully to the directions, thus boosting their performance.

Race/Ethnicity

Test bias in terms of ethnicity is by far the most publicized type of bias in the history of psychometrics, especially regarding differences among White and Black American students (Cole & Zieky, 2001; Freedle & Kostin, 1997; Vars & Bowen, 1998). This widespread interest in ethnic DIF is based on the long history of oppression in the United States, beginning with slavery, of Black individuals by White individuals. Unfortunately, these cultural, economic, and ethical issues of oppression are still highly relevant today, with the observation that being a person of color is a risk factor for low educational attainment, and not pursuing university education (Steele, 2003; Stulberg, 2015). Therefore, just as with language groups, American university samples are quite a bit less diverse than the general American population, and those individuals of color who attend university may represent a select sample (Iverson, 2007). In this way, because the population of interest in this proposed study is university students, DIF or mean differences in performance between these groups may be much less likely than it would be in the general population. However, DIF among ethnic groups in American university samples has been found with frequency in the past, potentially due to cultural differences in cognitive organization, or differing access to education (Rosselli & Ardila, 2003).

Unfortunately, if such DIF is found in the proposed study, deciding whether or not that DIF represents bias is particularly convoluted. In contrast to sex differences, explanations of differences between ethnic groups are almost never attributed to genetic differences, and instead are explained in terms of differential access to economic and educational resources (Poortinga, 1995; Sternberg, 2007). However, because those educational differences can cause major variance in the development of cognitive abilities (Weatherholt, Harris, Burns, & Clement, 2006), they may be producing genuine differences in terms of relational reasoning. For this reason, it is important to conceptually separate mean differences in performance among ethnic groups (sometimes termed *impact*) with DIF and bias. Mean differences in TORR performance among ethnic groups may not represent construct irrelevant variance, but instead be capturing true differences that are caused by ongoing oppressive economic and social systems. In contrast, DIF or measurement invariance, especially among discrimination parameters or loadings, may represent a more problematic situation, in which the comparison between the ethnicities is not direct. Indeed, if the measurement of relational reasoning across ethnicities is highly variant, with much significant DIF, then the old saying “comparing apples to oranges” may be relevant, because the scores produced for each ethnicity cannot be necessarily said to represent the same mental construct.

Another potentially relevant finding from the DIF literature concerning ethnicity has centered around timed vs. untimed tests. In general, timed tests have been much more likely than untimed tests to exhibit significant DIF (Rosselli & Ardila, 2003; Wicherts, Dolan, & Hessen, 2005). Moreover, a number of potentially construct irrelevant abilities have been posited to explain this phenomenon, such as increased test anxiety for certain groups in a timed setting (Wicherts et al., 2005), differential usage of time management strategies (Chen, 2005), and differing cultural beliefs about the relation between ability and speed of processing (Rosselli & Ardila, 2003). Specifically, certain students, especially those from a Hispanic background, may hold a cultural belief that a slower pace of reasoning reflects greater attention and deep processing, while the belief that more able thinkers reason more quickly is a typical belief of White students (Rosselli & Ardila, 2003). It should be noted here that the TORR is an untimed test, an attribute that may reduce the likelihood of significant DIF or bias.

CHAPTER 3:

METHODOLOGY

This chapter describes the participants, measures, and procedures that were the basis for this study of the cultural fairness of the Test of Relational Reasoning (TORR). But first, the analysis plan that was pursued when conducting this study is overviewed. The analysis plan is relevant here, because it formed a conceptual basis for the specific methodological steps taken in this study.

Overview of Analysis

The purpose of this study was to: (a) test for DIF on the TORR items among gender, language, and ethnic groups using a multi-group latent variable measurement model approach; (b) theoretically or empirically determine if that DIF is an indicator of construct irrelevant variance, or bias, on the TORR; and, (c) if bias is identified, respond to it in an appropriate and effective way. Therefore, this investigation unfolded in three conditional phases. First, the presence of potential DIF in TORR items across gender, ethnicity, and language groups will be tested using a multi-group latent variable measurement model procedure. Secondly, if DIF is detected, it would be further scrutinized theoretically and empirically in order to determine whether it constitutes item bias or not. Finally, if bias is determined to be present in one or more TORR items, systematic responses to that bias, would be undertaken. It should be noted here that, if any of the phases of this investigation reveal null results (e.g., no significant DIF is uncovered) than the subsequent stages would not be necessary.

Model Comparison Procedure

In order to evaluate DIF using a latent-variable measurement model paradigm, iterative multi-group MIRT models were fit to the TORR data. The procedure followed methodological recommendations formalized through simulation work in the MIRT context by Stark and colleagues (2006), but that have been meaningfully employed in the unidimensional IRT context for some time (e.g., Reise et al., 1993). This procedure began by fitting a two-group bi-factor model across the groups being compared (e.g., males and females). Of these two groups, the one with the larger N (e.g., males) was designated as the *reference* group, and the latent means and variances of that group's model were set to zero and one, respectively. In contrast, the latent means and variances of the other, or *focal* group were free to vary in relation to the latent means and variances of the reference group.

Next, *referent* items, whose parameters are constrained to be equal across groups, were specified. In the case of the bi-factor model, one referent item must be specified for each specific latent ability being measured, and the loading of each of those items on the general ability factor was also set to equality. Because the scale of the latent abilities across the groups is determined by these referent items, choice of the referent items is methodologically important. In general, referent items should be strongly related to the abilities they measure, and should be moderate in difficulty (Stark et al, 2006). If possible, referent items should also be chosen that are theoretically less likely to display DIF, or that have empirically shown no DIF in previous invariance testing work. In this investigation, referent items were chosen based on previous calibration work, and are further discussed in Chapter 4.

Then, the free-baseline model, in which only the parameters associated with the referent items are constrained across groups, was run, and its chi-square fit statistic was recorded. Next, a model that constrained the parameters associated with another of the items, in addition to the parameters associated with the referent items, was fit and its chi-square fit statistic recorded. The increase in the model chi-square value associated with that more-constrained model was tested for significance at 4 degrees of freedom, which is the difference in degrees of freedom between the free-baseline and constrained models. Whether or not this chi-square increase reached significance allowed for inferences about whether or not the constrained item displayed significant DIF across the groups being compared. This procedure was repeated for each of the non-referent items on the TORR. Then, in order to test the referent items, the TORR item from each scale that, when constrained, displayed the lowest chi-square increase from the free-baseline model (and therefore, the least DIF) was chosen as a new referent item. With these four empirically chosen referents, the procedure was repeated in order to test the referent items for DIF.

It should be noted that all MIRT analysis in this investigation was conducted using flexMIRT (Cai, 2013), software utilizing the expectation-maximization (EM) algorithm and priors of 2.0 for the estimation of item parameters (Bock, Gibbons, & Muraki, 1988). Also, the supplemented EM algorithm (Cai, 2008) was utilized for the calculation of standard errors.

Moreover, this procedure required 32 consecutive model comparisons (one for each TORR item) on the same set of group data. Importantly, 32 chi-square tests conducted on the same data would greatly inflate the family-wise Type-I error rate of each full likelihood-ratio procedure (Klockars & Hancock, 1994; Rice, 1988). Inflated family-wise type-I error rate is particularly problematic in cross-cultural or demographic investigations such as this one, where the identification of differences between groups may be politically delicate. Therefore, researchers examining DIF among demographic groups such as those in this study often choose to correct for inflated family-wise Type-I error rate, in order to be conservative about the instances of DIF that they identify (Elosua, 2011).

Although there are a variety of methods for correcting for inflated type-I error rate, the most widely used, and most conservative, is the Bonferroni correction. The conservatism at the root of this correction stems from the fact that it ignores the correlated structure of a series of tests (in this case DIF tests) and corrects based on their assumed orthogonality. Because DIF tests of items from the same measure are likely correlated, meaning that if one item displays DIF, another item is more likely to do so, this correction may be stronger than is absolutely necessary to retain a family wise error rate of .05. However, previous simulation work (e.g., Elosua, 2011; Stark et al., 2006) has found that, when DIF is large, the Bonferroni correction is capable of effectively eliminating type-I error, while preserving power. Although such simulation work does demonstrate that the Bonferroni correction limits power to detect small levels of DIF on a given item, larger levels of DIF that may be practically significant are typically the focus of investigations of DIF in the substantive research literatures. In this way, the Bonferroni correction allows for a criterion for significance that largely guarantees that an item flagged for DIF is actually problematic across groups. In the test-development process, time spent attempting to remedy DIF that is not practically significant may be time wasted, and for that reason the Bonferroni correction is commonly used.

Therefore, the Bonferroni correction was applied to the Type-I error rate, in order to maintain a family-wise error rate of .05, as is recommended in the IRT measurement invariance literature (Stark et al., 2006). Specifically, the appropriate critical *p*-value to utilize in order to limit the Type-I error rate to .05 is calculated via the Bonferroni correction through the formula $p_{crit} = \frac{\alpha}{k}$, where *k* is the number of tests conducted on the same data, in this case *k* = 32. Using this correction formula, the critical *p*-value that will be utilized in the likelihood ratio tests in this proposed investigation is:

$$p_{crit} = \frac{\alpha}{k} = \frac{.05}{32} = .001$$

Therefore, chi-square values that are significant at the *p* = .001 level will be taken as indicators that statistically significant DIF is present. This *p*-value corresponds to a critical chi-square value of 18.50, based on conversion tables (e.g., Field, 2013). Finally, it should also be noted that this procedure was repeated for each of the types of groups (i.e., gender, language, and race) included in this study. In the case of race/ethnicity groups, one particular group that was the most populous (i.e., White) was used as the focal group for comparisons to each of the other groups included in the study. In this way, all likelihood-ratio procedures were undertaken with two-group models.

Power Analysis

Before conducting this study, an a priori power analysis (i.e., sample size determination) was conducted for testing data-model fit, as outlined by Hancock and French (2013). As a first step in this process, the desired level of power (π = .90) and type-I error rate (α = .05) must be specified. Then, the degrees of freedom associated with the model being fit to the data must be calculated. In the case of MIRT models and their associated limited-information fit statistics, such as are being interpreted in this investigation, the number of degrees of freedom are written as $\mathrm{df} = \kappa - \nu$ where κ is the number of reduced first- and second- order marginal residuals, and ν is the number of parameters being estimated (Cai & Hansen, 2013). Additionally, for a multi-group model with two groups, κ is calculated as:

$$\kappa = \frac{(n_1)(n_1+1)}{2} + \frac{(n_2)(n_2+1)}{2} \tag{4}$$

Where n is the number of items on the test. In this case, because the TORR has 32 items:

$$\kappa = \frac{(32)(32+1)}{2} + \frac{(32)(32+1)}{2} = 1056$$

Then, in the case of the free-baseline model, 250 parameters are being estimated. This value is reached from 240 item parameters being estimated across both groups (taking into account that 4 items have each of their 4 parameters constrained across groups). Additionally, 5 latent means and 5 latent variances are being estimated for the comparison group model. So, in order to calculate the degrees of freedom associated with the free-baseline model:

$$\text{df} = \kappa - \nu = 1056 - 250 = 806$$

It should be noted here that the models being compared to the free-baseline models will have 4 additional degrees of freedom, for a total of 810. This is because the four model parameters associated with each item (i.e., one intercept, two discriminations, and one guessing for each item) will be constrained across groups. When the item parameters are constrained across groups, they need not be estimated for both groups, therefore saving one df per parameter. Importantly, both of the discrimination parameters for each item—pertaining to both the general and specific factors that the item loads on—are being constrained in this procedure. Therefore, the DIF being tested for in this study pertains to both abilities (general and specific) that an item is measuring.

With the degrees of freedom calculated, a theoretically appropriate increase in misfit associated with the constraining of an items' parameters must be posited. In this case, in order to be conservative, I have chosen to hypothesize an increase of .01 units of RMSEA when an item's parameters are constrained. In order to calculate the required sample size for achieving these specified parameters, a web application, created by Preacher and Coffman (2006) will be utilized. This application operates by generating R code based on a user's entered parameters. After running that generated code in R, a minimum per model sample size is produced. The statistical operations accomplished by this code are based on work from Hancock and Freeman (2001), MacCallum, Brown, and Cai (2006), and Preacher, Cai, and MacCallum (2007).

For this investigation, the required sample size per group is N = 113. It should be noted that this required sample size is feasible for all of the gender, language, and ethnic groups relevant to this investigation except for Native Americans. Unfortunately, the University of Maryland has such a low proportion of Native American students that collecting the required sample may be untenable, and as such, the functioning of the TORR with Native or Tribal populations must remain a future direction. A demographic breakdown of the University of Maryland population is displayed in Table 2.

Participants

Participants were 1,379 undergraduate students enrolled at a large public research university in the mid-Atlantic region of the United States. The sample was representative of the full university population in terms of the gender, ethnicity, language background, major, and year in school. The resulting demographic information is displayed in Table 2 for the sample, as well as the university population, along with corresponding chi-square tests for representativeness. Additionally, students ranged in age from 18 to 26, with a mean age of 21.34 (*SD* = 1.96). A one-sample t-test was used to confirm that this mean was not significantly different than the university reported mean age of 21.0 [*t* (1,378) = 1.06, *p* =.28]. Moreover, students reported GPAs ranging from 1.5 to 4, on a 4-point scale, with a mean of 2.81 (*SD* = .24).

Measures

Test of Relational Reasoning

As described in Chapter 2, the TORR ($\alpha = .84$) is a 32-item reasoning test, designed to limit the need for participant prior knowledge, and language through the use of graphical, non-linguistic items. The TORR is comprised of four scales representing each of the four forms of relational reasoning previously described (i.e., analogy, anomaly, antinomy, and antithesis). Each scale consisted of two practice items followed by eight test items. A full review of the development of the TORR, as well as accompanying psychometric information concerning the measure's reliability and validity, is presented within Chapters 1 and 2.

Demographic Questionnaire

A standard demographic questionnaire, focusing on the variables relevant to this investigation, was included. This demographic questionnaire is included in Appendix A. As can be seen, the variables gender and sex were collected separately in accordance with findings from the extant literature. However, no participant in this sample reported having a gender that did not directly correspond to their sex (e.g., no males were also women). Participant race, and whether or not that participant was a native English speaker was also collected. Further, the variables age, year in school, majors, minors, and overall GPA were also collected. While DIF was not examined across groups defined by these other variables, those variables were used to check for sample representativeness, and saved for potential future investigations.

Procedures

The sample was collected through direct communication with instructors across the university, who received information about the study as well as a link to the online version of the TORR to disseminate to their students via email. In exchange for their students' participation in this study, instructors agreed to offer extra course credit. The online version of the TORR was powered by Qualtrics (2014) survey software, and was programed to present the scales of the TORR (i.e., analogy, anomaly, antinomy, and antithesis) in a randomized, counterbalanced order across participants. Consistent with previous research utilizing the TORR (e.g., Alexander et al., 2015), students could participate in this study from any computer connected to the Internet, but could not participate on a smartphone or tablet. Additionally, students were permitted to take as much time as they needed to complete the TORR, with the average time being 29.61 minutes (*SD* = 7.32). No student took more than 50 minutes to complete the measure. After students had completed the TORR, they provided demographic information, and logged out of the study website. To avoid redundant data, students enrolled in multiple courses that were participating in this study were not permitted to re-take the TORR, but were offered an alternate extra credit assignment in that course. Moreover, to avoid issues related to missing data, the online version of the TORR is programmed not-to-allow participants to skip items without supplying an answer. This strategy is effective at eliminating systematic item-skipping across participants related to item difficulty, but it further highlights the need for the guessing parameter of each item to be estimated, because it likely results in increased guessing, especially in highly-difficult items. Moreover, participants in this investigation were not permitted to revisit items after submitting a response. Finally, because 34 faculty members around the university were contacted for recruitment, each with more than one-hundred undergraduate students in their classes during the semester in which data were collected; participant non-response is likely an important issue to consider when interpreting results from this investigation. As such, non-response will be discussed in-depth in the limitations section of Chapter 5.

CHAPTER 4:

RESULTS

Dependence Among Grouping Variables

Any iterative process for testing DIF among multiple types of groups (e.g., gender and language groups) has as an assumption the independence of those groups. For example, if female students are significantly more likely than male students to be non-native English speakers, and DIF exists that, in reality, is attributable to language background, than the subsequent tests will conflate these two potential causes of DIF and possibly indicate DIF among gender groups as well. So, before proceeding with an iterative multi-group model comparison process, the independence of each of the relevant grouping variables was tested using χ^2 tests of independence. See Table 8 for summary data and associated χ^2 tests.

Table 8

Independence of Grouping Variables for DIF Analysis

	Male	Female	Chi-Square
English	614	590	$\chi^2(1) = .21$, *p* = .65
Non-English	86	89	
White	369	343	$\chi^2(3) = 1.74$, *p* = .63
Black/African-American	130	126	
Hispanic/Latino	80	93	
Asian	96	94	

	English	Non-English	Chi-Square
White	702	10	$\chi^2(3) = 247.9$, *p* <.01
Black/African-American	248	8	
Hispanic/Latino	127	46	
Asian	127	63	

As may be expected the distribution of participants into gender groups was not dependent on the distribution within language or ethnic groups. However, there was evidence of significant dependence among language and ethnic groups. Specifically, Hispanic and Asian participants were more likely than White and Black participants to report English not being their first language. This finding makes intuitive sense, because Hispanic and Asian individuals tend to be more recent immigrants to the United States and the Maryland region than are White and Black individuals, making them more likely to not speak English as a first language.

Based on this finding, DIF on the TORR affecting Hispanic or Asian participants should be interpreted as potentially being related to the observation that those participants are more likely to have first spoken a language other than English—the language in which the TORR instructions are written. However, it should be noted that, despite the significance of the chi-square test associated with these grouping variables in Table 8, the proportion of non-native English speakers within the Hispanic and Asian groups in this sample is not as large as may potentially be expected. For example, only 26% of Hispanic participants in this study reported being a non-Native English speaker. Further, 33% of Asian participants in this study reported being a non-native English speaker. These proportions may pertain to the particular make-up of the undergraduate population of the University of Maryland, and will be discussed further in Chapter 5.

Differential Item Functioning

After examining the dependence among the demographic grouping variables of interest in this study, the presence of DIF on TORR items among those groups is able to be examined. As overviewed earlier, the process of detecting DIF requires comparing the fit of MIRT models that do or do not allow for an item's parameters to vary across groups. In order to account for participant membership in particular groups the bi-factor model can be modified as follows:

$$P\ (x_{ijk} = 1 \mid \theta_{ik}, \xi_{ik}) = c_{jk} + \frac{1 - c_{jk}}{1 + e^{-a_{jk}(\theta_{ik} + \xi_{ik} - b_{jk})}} \tag{5}$$

Here, the subscript *k* indicates that the parameters being estimated by the model pertain not only to the participant (*i*) and to the item (*j*), but also to the group from which that participant is drawn (Kim & Yoon, 2011). This equation can be compared to the bi-factor model utilized in earlier work with the TORR (see Equation 2), in which group membership was not accounted for. As explained in Chapter 2, during the earlier TORR calibration (Dumas & Alexander, 2016) the invariance of model parameters across sub-groups (*k*) of the population of undergraduate students was an important but unexamined assumption of the bi-factor model. By relating Equation 5 to Equation 2, the null hypothesis of this investigation, that any given TORR item has no DIF, can be formally posited as:

$$H_0 : P(x_{ij} = 1 \mid \theta_i, \xi_i) = P(x_{ijk} = 1 \mid \theta_{ik}, \xi_{ik}) \tag{6}$$

In the above equation, the null hypothesis refers to a situation where the probability of participants selecting the correct response on a given item is conditional on their general relational reasoning ability (θ) and their scale-specific ability (ξ), but that the participants membership in a given demographic group (*k*) does not significantly affect the estimation of that probability. In contrast, the alternative hypothesis of this investigation, that significant DIF does exist on one or more of the TORR items, can be posited as:

$$H_1 : P(x_{ij} = 1 \mid \theta_i, \xi_i) \neq P(x_{ijk} = 1 \mid \theta_{ik}, \xi_{ik}) \tag{7}$$

The above equation represents a situation in which significant DIF is present, because it shows that the probability of a participant selecting the correct response on a TORR item is conditional on their demographic grouping. Going forward, this hypothesis is tested through the model comparison procedure overviewed in Chapter 3.

Choice of Referent Items

As mentioned in Chapter 3, in order to utilize the free-baseline approach to model comparison and DIF detection, referent items, which will have their parameters constrained across groups, need to be identified. In the DIF detection context, referent items are particularly important, because their constraint across the groups allows for the latent variables being measured to have comparable scale across the groups. As such, the magnitude of DIF detected in non-referent items is conceptually on the scale of the referent items themselves. For these reasons, testing for DIF on the TORR requires the identification of four referent items: one for each scale of the TORR. Going forward, these items' parameters will be constrained across groups, giving scale to each of the specific factors, as well as the general dimension.

When choosing referent items, it is ideal to draw information about the invariance of parameters of particular items from previously published invariance work (Stark et al., 2006). However, because this study is the first to examine the invariance of TORR item parameters, no such published work exists. Therefore, educated choices of referent items must be made based on other available information. In the measurement of more crystallized abilities (e.g., verbal SAT items; O'Neill &McPeek, 1993), in the absence of prior invariance work, the choice of referent items may be driven largely by the specific content of the item and reasonable assumptions that particular content may or may not differentially affect different demographic groups. However, in the measurement of fluid visuo-spatial abilities such as the TORR is designed to tap, it is not straight-forward to separate the items based on content, because it is not clear which of the stimuli within the items may function differentially. Therefore, the most valid way to select referent items in this context is to utilize item parameters estimated from previous calibration work.

Fortunately, simulation research (e.g., Rivas, Stark, & Chernyshenko, 2009) has been conducted to determine what types of items, in terms of their estimated parameters, are most suitable to be referent items in a DIF detection procedure. Specifically, Rivas and colleagues (2009) recommend the use of items that are highly, but not too highly, discriminating (ideally $1.0 < a < 2.0$), and that have a difficulty parameter close to the level of theta at which the test is most informative (on the TORR $b \approx .8$). Moreover, the guessing parameters of referent items should be relatively close to the expected guessing parameter given the number of answer choices (with four answer choices on the TORR $c \approx .25$). This is done so that the effect of guessing on the referent items is representative of the guessing effect for the entire test. In general, these guidelines depict a well-functioning but proto-typical item as the appropriate choice for a referent item.

For this investigation, referent items were chosen based on these criteria, using item parameters estimated during the initial calibration of the TORR with the bi-factor model (Table 6). Importantly, these criteria were produced via simulation research only in the unidimensional IRT context, so no guidelines exist on choosing referent items based on general- and specific-ability parameters. However, the bi-factor model places greater emphasis, in terms of discrimination parameters and loadings, on the general ability factor. Moreover, the RRQ norms that were previously derived (Dumas & Alexander, 2016) were based solely on the general ability factor. Therefore, referent items were chosen solely for their parameters associated with the general ability factor. Specifically, the items that were chosen as referent items in this investigation were: Analogy 1, Anomaly 7, Antinomy 8, and Antithesis 4. For model comparisons related to each of the demographic grouping variables of interest in this investigation, the parameters associated with these four items were constrained across groups in order to test the invariance of each of the other 28 TORR items. Then, in order to fully ascertain whether these referent items were appropriate choices (i.e., they did not display DIF), they were also tested for invariance. Specifically, after the other 28 items had been tested, the item from each scale that displayed the lowest chi-square increase associated with the constrained model was chosen as a new referent item. Then, these four new-referents were constrained, and the previously chosen referents were tested one at a time. In this way, the referent items utilized in this investigation were chosen using the best-available simulation-based criteria, and empirically verified as appropriate referent items at each stage of analysis. Therefore, in general, inferences concerning DIF across demographic groups are based on the most best available methodological choices.

Gender

Results from each of the likelihood-ratio tests for DIF between males and females are displayed in Table 9.

Table 9

Likelihood-ratio Tests for DIF between Males and Females

Scale	Constrained Item	Model df	Model Chi-Square	Chi-square increase from baseline $\chi^2_{crit} = 18.50$
Baseline	Only referents	806	52158.75	--
Analogy	1	--	--	--
	2	810	52159.19	0.44
	3	810	52168.79	10.04
	4	810	52158.84	0.09
	5	810	52158.95	0.20
	6	810	52160.27	1.52

	7	810	52161.36	2.61
	8	810	52158.90	0.15
Anomaly	1	810	52164.96	6.21
	2	810	52164.33	5.58
	3	810	52161.26	2.51
	4	810	52159.91	1.16
	5	810	52161.51	2.76
	6	810	52176.19	17.44
	7	--	--	--
	8	810	52161.55	2.80
Antinomy	1	810	52161.95	3.20
	2	810	52168.53	9.78
	3	810	52158.87	0.12
	4	810	52159.21	0.46
	5	810	52161.37	2.62
	6	810	52165.54	6.79
	7	810	52158.95	0.20
	8	--	--	--
Antithesis	1	810	52166.84	8.09
	2	810	52164.26	5.51
	3	810	52163.47	4.72
	4	--	--	--
	5	810	52162.08	3.33
	6	810	52160.19	1.44
	7	810	52160.90	2.15
	8	810	52164.32	5.57

Note: Referents for this analysis were Analogy 1, Anomaly 7, Antinomy 8, Antithesis 4

As can be seen, the free-baseline model, which constrained the parameters of only the referent items across groups had 806 degrees of freedom and a model chi-square value of 52158.75. Each of the models that follow constrained four additional parameters, associated with the particular item being tested for DIF, resulting in 810 degrees of freedom. Moreover, each of these models had a chi-square value higher than the free-baseline model. The increase in chi-square value between a constrained model and the free-baseline model was tested for significance on a chi-square distribution with four degrees of freedom. As already mentioned, the Bonferroni correction was applied to these tests in order to hold the type-I error rate at .05 across each group comparison. Therefore, the critical chi-square value, which would indicate that significant DIF existed on an item of the TORR, was 18.50. As can be seen in Table 9, none of the items on the TORR displayed significant DIF between gender groups based on that criterion. In fact, only three items on the TORR displayed a chi-square increase from the free-baseline model that was at least half of the magnitude of the critical value. One item, Anomaly 6, came within two chi-square units of the critical value, but did not reach significance.

As previously mentioned, the items on each scale that displayed the lowest chi-square increase were used as new-referents when testing the previously selected referents for DIF. As can be seen in Table 9, the new free-baseline model, with only the newly chosen referent items' parameters constrained across groups, displayed a chi-square value of 52133.62 and 806 degrees of freedom. Then, just as constrained models were previously compared to the baseline model in Table 9, chi-square increase values are presented in Table 10.

Table 10

Likelihood-ratio Tests for DIF between Males and Females: Testing Referents

Scale	Constrained Item	Model df	Model Chi-Square	Chi-square increase from baseline $\chi^2_{crit} = 18.50$
Baseline	Only new referents	806	52133.62	--
Analogy	1	810	52137.75	4.13
Anomaly	7	810	52134.11	0.49
Antinomy	8	810	52133.83	0.21
Antithesis	4	810	52134.78	1.16

Note: New referents for this analysis were Analogy 4, Anomaly 4, Antinomy 3, Antithesis 6

As can be seen, the referent items did not display DIF, and all showed a chi-square increase less than a third of the critical value. This finding implies that the statistical criteria used to select referent items was effective at determining items that were appropriate referent items. Therefore, Tables 9 and 10 converge on the finding that no significant DIF exists on TORR items across gender groups.

Language Background

Analogously to Tables 9 and 10, which pertain to gender groups, Tables 11 and 12 present the results of likelihood-ratio tests for DIF across language groups.

Table 11

Likelihood-ratio Tests for DIF between Language Groups

Scale	Constrained Item	Model df	Model Chi-Square	Chi-square increase from baseline $\chi^2_{crit} = 18.50$
Baseline	Only referents	806	52184.34	--
Analogy	1	--	--	--
	2	810	52190.33	5.99
	3	810	52190.35	6.01
	4	810	52185.65	1.31
	5	810	52184.78	0.44
	6	810	52185.24	0.90
	7	810	52184.92	0.58
	8	810	52185.91	1.57
Anomaly	1	810	52189.97	5.63
	2	810	52187.88	3.54
	3	810	52192.47	8.13
	4	810	52184.90	0.56
	5	810	52184.51	0.17
	6	810	52185.67	1.33
	7	--	--	--
	8	810	52185.60	1.26
Antinomy	1	810	52185.51	1.17
	2	810	52188.54	4.20
	3	810	52186.53	2.19
	4	810	52184.50	0.16
	5	810	52191.80	7.46
	6	810	52187.25	2.91

	7	810	52187.14	2.80
	8	--	--	--
Antithesis	1	810	52187.41	3.07
	2	810	52187.40	3.06
	3	810	52189.75	5.41
	4	--	--	--
	5	810	52188.08	3.74
	6	810	52186.88	2.54
	7	810	52184.82	0.48
	8	810	52186.90	2.56

Note: Referents for this analysis were Analogy 1, Anomaly 7, Antinomy 8, Antithesis 4

As can be seen in Table 11, the free-baseline model that constrained only the originally selected referent items (i.e., Analogy 1, Anomaly 7, Antinomy 8, and Antithesis 4) displayed a chi-square value of 52184.34. Chi-square increases from that value associated with each of the constrained models give information about the significance of DIF on TORR items across language groups. None of the items tested in Table 11 showed chi-square increases greater than the Bonferroni corrected critical value of 18.50. In fact, none of the items showed chi-square increases that were half the magnitude of that critical value.

The items that showed the smallest chi-square increase per scale were Analogy 7, Anomaly 5, Antinomy 4, and Antithesis 7. Therefore, these items were used as referent items to test for DIF in the original referent items, none of which displayed significant DIF (See Table 12).

Table 12

Likelihood-ratio Tests for DIF between Language Groups: Testing Referents

Scale	Constrained Item	Model df	Model Chi-Square	Chi-square increase from baseline $\chi^2_{crit} = 18.50$
Baseline	Only new referents	806	52177.62	--
Analogy	1	810	52179.86	2.24
Anomaly	7	810	52178.30	0.68
Antinomy	8	810	52180.04	2.42
Antithesis	4	810	52181.83	4.21

Note: New referents for this analysis were Analogy 7, Anomaly 5, Antinomy 4, Antithesis 7

Perhaps interestingly, none of the items that displayed the lowest chi-square increases per scale when testing for DIF among language groups were the same as those that displayed the lowest chi-square increase when testing across gender groups. This finding may imply that, despite the fact that no significant DIF was uncovered among language groups in this study, the underlying mechanisms that drive differences among gender and language groups differ in important ways.

Race/Ethnicity

In this investigation, tests for DIF among race/ethnicity groups were conducted by fitting multi-group MIRT models to one focal group and one reference group at a time. In this way, despite there being five race/ethnicity groups included in this analysis (i.e., White, Black, Hispanic, and Asian), each of the likelihood-ratio procedures featured two-group models. Specifically, White participants were chosen as the reference group for each model comparison procedure. This choice was made because White students were the most populous group in the sample, and White students remain the majority in the population of interest (i.e., undergraduate students). Therefore, separate likelihood-hood ratio procedures were conducted to detect DIF between White and Black participants, White and Hispanic participants, and White and Asian participants. The results of each of these procedures are detailed below.

Black/African American. Tables 13 and 14 contain information related to the likelihood-ratio tests for DIF between White and Black participants.

Table 13

Likelihood-ratio Tests for DIF between White and Black Participants

Scale	Constrained Item	Model df	Model Chi-Square	Chi-square increase from baseline $\chi^2_{crit} = 18.50$
Baseline	Only referents	806	36545.82	--
Analogy	1	--	--	--
	2	810	36545.98	0.16
	3	810	36552.37	6.55
	4	810	36548.22	2.40
	5	810	36546.75	0.93
	6	810	36546.85	1.03
	7	810	36548.41	2.59
	8	810	36546.86	1.04
Anomaly	1	810	36547.16	1.34
	2	810	36554.19	8.37
	3	810	36546.11	0.29
	4	810	36546.56	0.74
	5	810	36546.89	1.07
	6	810	36546.46	0.64
	7	--	--	--
	8	810	36548.52	2.7
Antinomy	1	810	36563.2	17.38
	2	810	36548.49	2.67
	3	810	36552.77	6.95
	4	810	36554.15	8.33
	5	810	36559.51	13.69
	6	810	36551.7	5.88
	7	810	36552.79	6.97
	8	--	--	--

Antithesis	1	810	36548.69	2.87
	2	810	36548.52	2.70
	3	810	36547.02	1.20
	4	--	--	--
	5	810	36549.61	3.79
	6	810	36545.87	0.05
	7	810	36551.87	6.05
	8	810	36546.55	0.73

Note: Referents for this analysis were Analogy 1, Anomaly 7, Antinomy 8, Antithesis 4

The free-baseline model described in Table 13, which constrained the item parameters of only the referent items, had 806 degrees of freedom and a model chi-square value of 36545.82. When the parameters associated with each of the other TORR items were constrained across groups, no item displayed a chi-square increase from the baseline that was greater than the critical value of 18.50. Two items (i.e., Antinomy 1 and 5) did display chi-square increases that were greater than half of that critical value, but did not reach significance. The items that displayed the lowest chi-square increases per scale were Analogy 2, Anomaly 4, Antinomy 2, and Antithesis 6. These items were used as new-referent items to test the original referent items for DIF, none of which displayed a significant chi-square increase (see table 14).

Table 14

Likelihood-ratio Tests for DIF between White and Black Participants: Testing Referents

Scale	Constrained Item	Model df	Model Chi-Square	Chi-square increase from baseline $\chi^2_{crit} = 18.50$
Baseline	Only new referents	806	36517.22	--
Analogy	1	810	36517.62	0.4
Anomaly	7	810	36518.44	1.22
Antinomy	8	810	36521.22	4.00
Antithesis	4	810	36520.45	3.23

Note: New referents for this analysis were Analogy 2, Anomaly 4, Antinomy 2, Antithesis 6

Interestingly, Anomaly 4 was the item with the lowest chi-square increase on the Anomaly scale across both male and female and White and Black participants. This finding may imply that item is particularly suited for measuring relational reasoning across demographic groups.

Hispanic/Latino. As with the likelihood ratio tests for DIF between White and Black participants, no item on the TORR displayed significant DIF between White and Hispanic students. As can be seen in Table 15, no item, when constrained across groups, produced a chi-square increase from the baseline model greater than the critical value.

Table 15

Likelihood-ratio Tests for DIF between White and Hispanic Participants

Scale	Constrained Item	Model df	Model Chi-Square	Chi-square increase from baseline $\chi^2_{crit} = 18.50$
Baseline	Only referents	806	33518.39	--
Analogy	1	--	--	--
	2	810	33518.55	0.16
	3	810	33519.39	1.00
	4	810	33519.39	1.00
	5	810	33519.86	1.47
	6	810	33519.31	0.92
	7	810	33518.79	0.40
	8	810	33519.18	0.79
Anomaly	1	810	33521.55	3.16
	2	810	33527.34	8.95
	3	810	33519.41	1.02
	4	810	33518.5	0.11
	5	810	33529.36	10.97
	6	810	33522.88	4.49
	7	--	--	--
	8	810	33519.94	1.55
Antinomy	1	810	33521.51	3.12
	2	810	33519.56	1.17

	3	810	33518.84	0.45
	4	810	33518.67	0.28
	5	810	33530.12	11.73
	6	810	33519.64	1.25
	7	810	33520.84	2.45
	8	--	--	--
Antithesis	1	810	33520.68	2.29
	2	810	33522.66	4.27
	3	810	33524.53	6.14
	4	--	--	--
	5	810	33519.1	0.71
	6	810	33521.77	3.38
	7	810	33519.73	1.34
	8	810	33524.66	6.27

Note: Referents for this analysis were Analogy 1, Anomaly 7, Antinomy 8, Antithesis 4

The items from each scale that displayed the smallest chi-square increases from the baseline were: Analogy 7, Anomaly 4, Antinomy 4, and Antithesis 5. It should be noted that Anomaly 4 displayed the least chi-square increase on the Anomaly scale in the likelihood-ratio tests pertaining to gender, Black participants, and Hispanic participants. This convergent finding suggests that this particular item is highly invariant across groups. However, none of the other items that displayed the smallest chi-square increase between groups have repeated across analyses, suggesting that the underlying causes of invariance are different depending on the kinds of groups being analyzed. As before, these new-referent items were used to test the original referent items for DIF, which none displayed significantly (see Table 16 for details).

Table 16

Likelihood-ratio Tests for DIF between White and Hispanic Participants: Testing Referents

Scale	Constrained Item	Model df	Model Chi-Square	Chi-square increase from baseline $\chi^2_{crit} = 18.50$
Baseline	Only new referents	806	33510.08	--
Analogy	1	810	33511.24	1.16
Anomaly	7	810	33511.99	1.91
Antinomy	8	810	33512.33	2.25
Antithesis	4	810	33511.57	1.49

Note: New referents for this analysis were Analogy 7, Anomaly 4, Antinomy 4, Antithesis 5

Asian. At this point in the analysis, no significant DIF had been uncovered on any TORR items across any of the demographic groups being tested. The likelihood ratio tests for DIF between White and Asian participants were no exception to this pattern, with no item, when constrained across groups, producing a significant chi-square increase. Please see Table 17 for full information on these likelihood-ratio tests.

Table 17

Likelihood-ratio Tests for DIF between White and Asian Participants

Scale	Constrained Item	Model df	Model Chi-Square	Chi-square increase from baseline $\chi^2_{crit} = 18.50$
Baseline	Only referents	806	34206.24	--
Analogy	1	--	--	--
	2	810	34207.15	0.91
	3	810	34216.15	9.91
	4	810	34209.9	3.66
	5	810	34208.13	1.89
	6	810	34207.13	0.89
	7	810	34210.1	3.86
	8	810	34210.95	4.71

Anomaly	1	810	34209.75	3.51
	2	810	34206.25	0.01
	3	810	34206.34	0.1
	4	810	34215.28	9.04
	5	810	34206.86	0.62
	6	810	34206.45	0.21
	7	--	--	--
	8	810	34211.26	5.02
Antinomy	1	810	34210.11	3.87
	2	810	34206.58	0.34
	3	810	34206.99	0.75
	4	810	34210.11	3.87
	5	810	34210.6	4.36
	6	810	34207.67	1.43
	7	810	34212.77	6.53
	8	--	--	--
Antithesis	1	810	34210.73	4.49
	2	810	34207.06	0.82
	3	810	34209.15	2.91
	4	--	--	--
	5	810	34211.33	5.09
	6	810	34210.71	4.47
	7	810	34209.08	2.84
	8	810	34211.61	5.37

Note: Referents for this analysis were Analogy 1, Anomaly 7, Antinomy 8, Antithesis 4

As can be seen, the items on each scale that produced the smallest chi-square increase were: Analogy 6, Anomaly 2, Antinomy 2, and Antithesis 2. As in the previous analyses, these items were used as new-referents to confirm that the original referents did not display significant DIF (see Table 18).

Table 18

Likelihood-ratio Tests for DIF between White and Asian Participants: Testing Referents

Scale	Constrained Item	Model df	Model Chi-Square	Chi-square increase from baseline $\chi^2_{crit} = 18.50$
Baseline	Only new referents	806	34196.93	
Analogy	1	810	34201.97	5.04
Anomaly	7	810	34197.78	0.85
Antinomy	8	810	34197.42	0.49
Antithesis	4	810	34196.94	0.01

Note: New referents for this analysis were Analogy 6, Anomaly 2, Antinomy 2, Antithesis 2

Summary of Analysis

This investigation put forward three main goals, each conditional on the one that preceded it: (a) test for DIF on the TORR items among gender, language, and ethnic groups using a multi-group latent variable measurement model approach; (b) theoretically or empirically determine if that DIF is an indicator of construct irrelevant variance, or bias, on the TORR; and, (c) if bias is identified, respond to it in an appropriate and effective way. Because of the conditional aspect of these analysis goals, should any of them produce wholly null results (i.e., the null hypothesis is retained across all tests conducted), the subsequent goals are not necessary. This is because, if no significant DIF is uncovered, no items are flagged for further investigation into the existence of cultural of bias.

As can be seen in Tables 9 to 18, no TORR item displayed significant DIF across any of the demographic groups of interest in this study. Given these findings, there are no items that, given the goals of this investigation, should be further investigated for the existence of bias. As such, the analysis related to this investigation can finish with an empirical answer to the first research question: no significant DIF appears to exist on the TORR among gender, language, or race/ethnicity groups within the undergraduate population. And a logical answer to the second two research questions: Given no significant DIF, there is no empirical evidence from this study that suggests cultural bias may exist on the TORR, and therefore there is no need to statistically correct such bias. Consequently, empirical analysis related to this investigation may stop, and further discussion of these findings commence.

CHAPTER 5:

DISCUSSION

This study was designed to assess the cultural-fairness of the Test of Relational Reasoning (TORR). The likelihood-ratio procedure undertaken in this investigation as an empirical means to detect DIF or non-invariance on the TORR, exposed no significant DIF on any of the TORR items across any of the demographic groups of interest. Therefore, the TORR items appear to function similarly across these demographic groups, and the TORR may be tentatively described as culture-fair. However, as described in Chapters 1 and 2, the issue of cultural fairness in measurement is highly complex, and generally no total claim of universal cultural fairness is possible for any measure (Poortinga, 1995; Sternberg, 2007; Zurcher, 1998). With this issue in mind, the null-finding of no significant DIF on the TORR from this investigation is discussed, delimitations and limitations of the current study presented, and future directions for TORR research are explained.

The TORR as Culture-Fair

One important assumption of the bi-factor model, and most other psychometric models used to measure cognitive abilities or psychological traits, is that the parameters of the measurement model used to estimate participants' ability are invariant across those participants, regardless of the demographic group from which they come (Cai, Yang, & Hansen, 2011). Unfortunately, in the psychological assessment literature, whether or not a given measure and its accompanying measurement model meet this assumption is not always empirically tested (Sternberg, 2008). After this investigation, empirical evidence exists to bolster the argument that the TORR can be meaningfully calibrated, normed, and scored in the undergraduate population, without explicitly accounting for the demographic group membership of participants from that population. Interestingly, this assumption of latent-variable measurement models (i.e., invariance) has been found to be untenable for a number of cognitive assessments (Poortinga, 1995; Rosselli & Ardila, 2003). Moreover, it is a goal of psychometrics in general (Verney et al., 2005) to produce measures that are culturally fair, or that at least meet the assumption of invariance across demographic groups. Therefore, it is interesting to discuss what aspects of the TORR or TORR administration may have contributed to no significant DIF being detected in the present study, so that these aspects may be used in future psychometric work as strategies for creating culturally fair measures. Going forward, these aspects will be organized in three broad categories: (a) administration procedures, (b) fluidity of stimuli, and (c) avoiding stereotype threat.

Administration Procedures

During the development of the TORR, a number of administrative and procedural choices were made that may have improved the likelihood of the TORR items being invariant across demographic groups. For example, the TORR was designed to be an untimed measure, on which students may spend as much time as they deem necessary. Second, participants who take the TORR online, as they did in this study, are not permitted to skip items, leave items blank, or revisit items after they have selected an answer. Finally, the scales of the TORR (i.e., analogy, anomaly, antinomy, and antithesis) were presented in a random order.

Timing. Many psychometric measures, including those administered in the school setting, are timed (Ardila, 2005). This methodological choice is predicated on a fundamental assumption, present even in the earliest days of psychometric research (Spearman, 1927), that those students who are more adept at a given cognitive skill can perform that skill more quickly. However, increasing evidence has suggested that this belief is in fact culturally specific to

Western society, from which most early psychometric research arose. In contrast, some cultures, namely Asian and Latin-American cultures, my hold the inverse belief—that greater time spent on a given task reflects a greater depth of processing, and thus better performance (Chen, 2005; Roselli & Ardila, 2003). Interestingly, the actual relation between speed-of-processing and cognitive ability, as well as how that relation comes to be, has been hotly debated for decades (e.g., Sternberg, 1986; Vernon, 1986), with accepted scientific consensus not yet reached.

In the meantime, untimed psychometric measures have become the norm for students with any identified cognitive impairment, including impairments arising from social or emotional sources, such as anxiety (Ashcroft & Moore, 2009). Although the extant body of work explicitly on the question of the relation between timing of a test and cultural-fairness is limited, there is some empirical evidence that untimed tests are more capable at identifying high-ability students from non-dominant cultures (Shaunessy, Karnes, & Cobb, 2004). For this reason, it may be that untimed tests are more likely to be culturally-fair than timed tests (Roselli & Ardila, 2003). Therefore, while this proposition remains empirically un-verified, the untimed nature of the TORR may have contributed to its invariance across demographic groups. Going forward, the relation between speed-of-processing and cognitive ability may remain a topic of interest to educational psychologists, and systematic investigations of timed and untimed versions of the TORR may inform that discussion.

Skipping and revisiting items. On the TORR as it was administered in this investigation, participants were not permitted to skip an item without selecting an answer choice, nor were they permitted to return to an item after making a selection. This strategy is utilized not only to prevent missing data, but also to preserve the novelty of the stimuli of each item to participants. Indeed, this strategy is used in the assessment of other fluid cognitive abilities besides relational reasoning (e.g., Abad et al., 2004). Interestingly, because there is evidence that students with differing educational or demographic backgrounds are differentially trained on test taking strategies such as skipping and returning to items. For example, Rindler (1980) found that test administration procedures that permitted students to skip and return to items increased variability among participants by advantaging those students who had higher ability and disadvantaging those students with lower ability. Drawing from this finding, it may be that tests, such as the TORR, that do not allow skipping and revisiting items limit the observed differences among students based on test-taking strategies. In this way, not allowing skipping and revisiting of items may contribute to the cultural fairness of a measure.

Ordering of scales. Relatedly, the online administration of the TORR utilized in this study was programed to randomize the order of the scales of the TORR (i.e., analogy, anomaly, antinomy, and antithesis). This strategy is utilized to combat a potential order-effect, in which certain scales are systematically affected by participant fatigue or testing effects. However, because it may be that the educational and demographic background of students is related to their ability to learn strategies for reasoning with particular stimuli while they are actually engaged with a measure, the randomization of the scales may have acted to control DIF as well. In effect, the randomization of the scales of the items of the TORR ensured that different participants reasoned with the items of the TORR in an essentially different order, negating the possibility that the order of presentation of items may benefit particular students depending on their background.

Fluidity

From the first attempts to create a culturally-fair measure of cognitive ability (Cattell, 1940), the fluidity or novelty of the stimuli on the measure has been considered to be highly important. This is because participants who differ on demographic background variables probably systematically differ on their exposure to particular topics within the educational system, and as such have developed differential crystallized abilities. However, if the stimuli on a measure are such that none of the participants have had any prior experience with them, then the effect of such differential exposure is lessened. It should be noted here that absolute novelty, meaning that no member of the target population could possibly have been prepared better for particular stimuli, is likely impossible. Indeed, in the case of the TORR, a number of particular cognitive strategies for mentally manipulating visual stimuli, as well as test-taking strategies for eliminating incorrect answer choices, likely played a role in participants' scores. In this way, it is probably not possible for any stimuli to elicit purely fluid cognitive processing, because some previously developed declarative or procedural knowledge (i.e., strategies) will likely always be involved in any reasoning process. However, the relative novelty of stimuli can be maximized, therefore placing the greatest possible emphasis on fluid intellectual processing and the least emphasis on crystallized abilities developed through education. As detailed in Chapters 1 and 2, the novelty of TORR stimuli was an explicit focus in item development, potentially leading to the finding of no significant DIF across TORR items.

Interestingly, there are those within the scholarly community who criticize fluid cognitive measures, in part because the novelty and perceived simplicity of their stimuli may encourage students to think of them as irrelevant (Rosselli & Ardila, 2003). This line of argument has its genesis in W.E.B Dubois's 1920 essay in which he pointed out that many rural Black soldiers during World War I may not have perceived the relevance of the psychometric assessments they were given, and as such were probably not motivated to do their best. However, given the findings of the current investigation, it appears that, at least at the undergraduate level, fluid reasoning on the TORR was effective at being measured invariantly across demographic groups. Clearly, the social and cultural landscape has changed since DuBois's time, and these changes may have contributed to a greater diversity of students perceiving relevance in fluid cognitive tasks.

Avoiding Stereotype Threat

Since the mid-1990's, psychologists have been aware of the phenomenon of stereotype threat—in which individuals who are aware of negative stereotypes about a group that they belong to "live down" to those stereotypes when they are active in their consciousness (Steele & Aronson, 1995). Importantly, this phenomenon is exacerbated when participants are asked to supply their demographic information before taking a test, or when students take a test in a setting, such as a high-stakes testing environment or psychological laboratory, that activates anxiety related to those stereotypes (Palumbo & Steele-Johnson, 2014; Spencer, Steele, & Quinn, 1999). With the TORR, two test administration choices were made that may have helped to diminish the effect of stereotype threat.

First, students were asked to supply demographic information only after they had completed the TORR, and secondly, students were able to participate in this study from any computer connected to the Internet. This second choice meant that students who may have otherwise felt anxiety upon entering a psychological laboratory and completing a cognitive measure were able to complete the TORR in their own private space, without necessarily feeling the anxiety a laboratory may produce. Of course, the completion of the TORR outside of the laboratory came with its own set of possible problems including a lack of control over whether or not the TORR was completed individually, and despite directing participants to focus solely on the TORR while taking it, whether or not participants were distracted with other stimuli (e.g., TV, websites) while completing the TORR. However, one major benefit of Internet-based test administration, besides making data collection more rapid and cost-effective, may be to limit the stereotype threat related anxiety students feel when taking a cognitive measure.

Limitations

As with any scientific endeavor, this investigation had certain limitations, which may be better addressed in future studies. For example, despite the collection of a relatively large (N=1379) sample of undergraduate students, who were representative of the population of the University of Maryland, participant non-response is necessary to consider. Also, while this study has shown that no significant DIF exists on the TORR within that particular population, the cultural fairness of the measure across an even more diverse population remains an open question. Finally, the general limitation of psychometric assessment in accounting for differences among demographic groups, while not over-emphasizing those differences, is discussed as it pertains to the TORR.

Non-Response

As outlined in Chapter 3, the sample utilized in this study was collected through direct contact with instructors across the University of Maryland campus. These instructors were situated within a variety of disciplines, in order to ensure the diversity, in terms of academic major, of the sample. 34 instructors around campus were contacted. Each of these instructors was chosen to be contacted because they were listed on the University of Maryland course scheduling website (Testudo) as teaching greater than 100 undergraduate students during the semester in which data-collection occurred. This means that, at the lowest, the potential sample-size, after each of those faculty members were contacted, was 3,400. With an actual sample size of 1,379 students, a maximum response rate of 41% was observed. This response rate, although greater than many documented response rates in higher-education samples (e.g., Sax, Gilmartin, & Bryant, 2003) is still low enough to potentially introduce non-response bias into the study. This form of bias is caused when those students who choose not to respond to a solicitation for research participation systematically differ on variables related to the study. For example, if students with generally high relational reasoning ability, or low test-anxiety, had systematically chosen to participate in this study, the results may have been different then they would have been without that systematic non-response.

Importantly, because the goals of this investigation particularly pertained to DIF among demographic groups, the hypotheses being tested would be most affected by non-response if that non-response was more or less systematic across demographic groups. For example, if male students responded or did not respond for reasons unrelated to the study, but female students were more likely to respond if they had high relational reasoning ability, or vice versa. Fortunately, because the collected sample was representative of the university population in terms of the demographic variables of interest, the likelihood of this differential non-response phenomenon having occurred appears to be low.

Moreover, there is no evidence of which I am aware that undergraduate students from various demographic groups are more likely to respond to calls to participate in research, especially when those calls are available across university majors. Therefore, while non-response was substantial in the data collection for this study, it seems unlikely that non-response significantly altered the findings of the investigation. However, going forward, future research with the TORR should conceptualize ways to improve the participant response rate, whether they be greater participant compensation or more tightly controlled data collection procedures.

Population Diversity

For a variety of economic, cultural, and social reasons, the population of undergraduate students in the United States is substantially less diverse—especially on economic variables related to educational attainment— than the population of the nation as a whole (U.S. Census Bureau, 2010). For that reason, although the findings of this investigation speak directly to the cultural fairness of the TORR within the undergraduate population, inferences cannot necessarily be drawn to the population of older adolescents and adults across the country. For example, those non-native English speakers who are actively engaged in undergraduate education are likely systematically different on a number of relevant variables from non-native English speakers of the same age who are not enrolled in college. Of course, the same could be said for nearly any demographic variable analyzed here or elsewhere.

For that reason, DIF may yet exist on the TORR across sub-groups of the young adult population, but was not detectable here, because the undergraduates sampled constituted a select subset of that population. For that reason, the population for which the TORR is considered culturally fair should be the undergraduate student population specifically, and not the older adolescent or young adult population generally. Moreover, whether or not the TORR is invariant across samples of undergraduate students from different countries or continents remains an open question. For example, after translating the directions, can the TORR be a reliable measure of relational reasoning in Japan, or Israel, for example? Although evidence now exists for the invariance of TORR items across sub-groups of the American undergraduate population, empirical investigations of its invariance across a broader, more diverse population, are still in the future.

Although the generalizability of the findings in this study to other populations outside of the undergraduate student population is not known, the procedure undertaken in this study does allow for generalizability from the specific sample collected at the University of Maryland to other groups of undergraduate students, at least within the United States and other similar countries (e.g., Canada). For one, the University of Maryland is a reasonably diverse public institution that allowed for the collection of data from students of varying demographic backgrounds, and secondly, the particular latent variable model utilized in this study supported the generalizability of findings by not relying solely on observed item-responses, but by positing latent abilities that caused variation in those item-responses. Because the same latent structure of abilities theoretically underlies the TORR regardless of the university from which a sample is drawn, and the differences among gender, race, and language groups are also likely constant, the findings of this investigation may be generalizable across the undergraduate population. One possible notable exception may be institutions with a high proportion of Native American or indigenous students, which were not available in large numbers at the University of Maryland. Moreover, the TORR may potentially also be culturally-fair within samples of senior-level high-school students, or beginning graduate, law, or medical students, because samples of such students are likely to have similar mean values on the distribution of latent relational reasoning ability, and potentially include similar levels of diversity as undergraduate students. Of course, senior-level high school students and beginning law or medical students are often only a few months removed from undergraduate status themselves. In institutions at which this is not the case because a much greater proportion of older and continuing students are enrolled as undergraduates, the cultural-fairness of the TORR may also need to be re-examined, to ensure the validity of its use in such a population. In these ways, the findings from the present investigation are generalizable, within reason, to populations that share characteristics with the undergraduate population of the University of Maryland.

Possibility for Multi-Group Calibration

In the likelihood-ratio procedure for DIF using latent variable measurement models, the fit of models that are free to estimate item parameters across groups are compared to the fit of models with the parameters of at least one other item constrained. Importantly, the more constrained models always fit worse, although in the case of this investigation, not significantly worse. This general observation, which is true for all latent variable measurement models and not only for the one used to calibrate the TORR, seems to beg the question: if a totally free model fits best, why not always calibrate every test with multi-group models? Indeed, if maximizing model-data-fit is a principal goal of psychometricians, than this strategy seems reasonable.

However, there are a number of practical, political, and social reasons why a default multi-group calibration strategy may be untenable. First, using multi-group models to calibrate any given test would necessitate the use of separate scoring procedures and norms for each group. For example, scoring manuals of a test fully calibrated using multi-group models would feature scoring conversions and normative information for each demographic group within the target population. Such differential scoring procedures or norms may create political or social tension surrounding the test, because they may be perceived by students, parents, or teachers as unfairly benefiting one demographic group over another. Further, such multi-group calibration and scoring would require definitively sorting every student who takes a given test into a particular demographic group.

This need for definitive sorting may lead to the use of heuristics such as the “one-drop” rule used for decades in the Southern States to define whether or not an individual was Black, and not accurately reflect the actual cultural background of many students. One can imagine difficult situations being created in the clinical setting, in which a school psychologist does not know, given a particular student’s demographic background, which scoring and normative tables to use. At least in my view, it is clear that situation is less than ideal for students and practitioners within the educational setting. Therefore, ascertaining, as was done in the present study, whether or not it is reasonable to calibrate and score a test using the same measurement model for each sub-group within a target model should remain a goal of test development.

Delimitations

As with any empirical investigation, the methods used in this dissertation deeply affect the inferences that are able to be drawn from the findings. In this context, delimitations are an explicit formulation of the effect of particular methodological choices made in this investigation on the interpretations of the findings. Three specific delimitations will be discussed here: (a) the interpretation of the bi-factor model, (b) the methodological choice to correct for Type-I error rate using the Bonferroni correction, and (c) the particular demographic groups that were utilized in this study.

Measurement Model Interpretation

In this study, the TORR data are modeled using a particular measurement model: the bi-factor model. As detailed in Chapter 2, this model was chosen both because of its empirical fit to the data, and because of its correspondence to one main purpose of the TORR: to model a highly reliable relational reasoning factor while also taking into account variance in items due to specific abilities. However, it is important to note here that, should a different measurement model have been used, the results of the DIF procedure may not have been the same. For example, the bi-factor model prioritizes the measurement of general relational reasoning ability, and does not model scale-specific abilities that bear a relation to that general construct. Therefore, it is still currently not known, if another model such as the correlated-factor model or a higher-order factor model (often termed a testlet model in IRT) would have shown converging results. For instance, the correlated factor model would have necessitated chi-square difference tests among the free-baseline and constrained models at only three degrees of freedom, as opposed to the four used here. This is because each item would only load on one particular factor, and therefore items would only have three parameters each to constrain. Such a methodological change may have altered the results, and their interpretation.

Moreover, it should be noted here that the specific ability factors in the bi-factor model do not correspond to forms of relational reasoning theoretically. Instead, they correspond to other abilities, specific to each particular scale of the TORR, that are not relational reasoning, but nonetheless contribute to variance on a given item. Because of this aspect of the bi-factor model, the interpretation of participant ability on those residualized factors (ξ) can be convoluted. In future research, if meaningful scores that pertain directly, for example, to participants' antithetical reasoning ability, are desired, then an alternate measurement model, such as the correlated factor model, or individual scale uni-dimensional measurement models, should be utilized. Importantly, the actual measurements produced by any test can differ depending on the measurement model utilized to calibrate and score that test. So, the appropriateness of the measurement model for the particular purpose of the test—in any measurement context, not only the present study—needs to be carefully considered.

Bonferroni Correction

As described in Chapter 3, the Bonferroni correction was utilized in this study to correct for Type-I error rate across the 32 likelihood-ratio tests associated with each demographic group comparison (e.g., TORR item functioning across males and females). The choice to utilize the Bonferroni correction was based on common practice in the applied psychometric literature, as well as the recommendation of methodological simulation work (e.g., Elosua, 2011; Stark et al., 2006). Moreover, the use of the Bonferroni correction aligned with the goal of this investigation to flag items for significant DIF only if that DIF was practically significant, and required further revision or consideration. However, it must be noted that, should no correction for family-wise Type-I error have been used, or if a different correction had been used other than the Bonferroni correction, it is likely that some significant DIF would have been found. For example, when constrained across gender groups, Anomaly 6 produced 17.44 chi-square units of mis-fit. At four degrees of freedom and with the Bonferroni correction applied, the critical chi-square value associated with that test was 18.50. Therefore, the chi-square increase of 17.44 did not reach significance. However, it is easy to see that this value is far above the uncorrected critical chi-square value for an alpha of .05 with four degrees of freedom: 9.48. So, one crucial delimitation related to the methodology of this investigation is that the finding of no DIF closely depends on the choice to use the Bonferroni correction.

Demographic Groups

Another critical delimitation of the current investigation is that the principal finding—that the TORR may be capable of measuring relational reasoning in a culturally-fair way—is dependent on the particular cultural groups included in this study. Following common practice in the culture-fair assessment literature; gender, race, and language groups were utilized in this study. However, while less commonly investigated, SES groups, which vary systematically by the amount of social or economic resources available to them, may also be interesting to study. Other possible types of groups for DIF analysis, reflecting the diversity of students from which psychological data is collected, may be: identified gifted students, students with intellectual disability or autism, or students with other genetic disorders that affect cognition such as Klinefelter's syndrome. It should be noted here that the demographic form utilized in this study did not prompt students to supply information related to their membership in any of these categories, and as such the invariance of TORR item parameters across such groups must necessarily remain a future direction.

Besides these possible groups, DIF among more specific groups, such as the interactions between the grouping variables used in the present study (e.g., Asian females and Asian males) , may be interesting to investigate. It may be that the increased homogeneity of such specific groups would increase the power of a likelihood ratio procedure to detect practically significant DIF, and could potentially elucidate important measurement non-invariance among individuals. In contrast, larger, more heterogeneous groups (e.g., all white and all non-white students) may also be used. While this is not common practice in the culture-fair assessment literature, such a general grouping may be an effective strategy for balancing the goals of invariance testing with sample size constraints within certain populations.

Future Psychometric Research on the TORR

Although many critical steps of measure development have now been completed with the TORR (see Table 1), some potentially interesting psychometric investigations into the functioning of the TORR are still to come. For example, the ability of the TORR to measure the relational reasoning ability of students at different points in their development is currently unknown. Also, further discriminant and convergent validity research may be important for the TORR's usefulness within the educational psychology literature. Finally, criterion methods for detecting possible differential predictive relations of the TORR to other measures across demographic groups—an important aspect of cultural fairness research—has not been undertaken.

Longitudinal Invariance

In most lines of research inquiry within the educational psychology literature, the particular developmental trajectories of academically related variables are of interest, and the relational reasoning literature is no exception. However, as far as I am aware, no measure of relational reasoning currently exists that is able to tap the construct in an invariant way across student of different ages. Therefore, in the future, investigating the longitudinal invariance of the TORR may be an interesting and important next step. Of course, no measure can be longitudinally invariance across the entire lifespan, but for what age ranges is the TORR reliable and invariant?

Moreover, outside of this identified age-range, alternative measures of relational reasoning may be created to tap the construct within much younger individuals, for example. Then, the TORR and those alternative measures may, if appropriate, be vertically scaled to create measurement of relational reasoning ability that is on the same scale across developmental stages. Although it would take substantial psychometric measure-development effort, this line of work, if completed, could allow for a rigorous quantitative understanding of the way relational reasoning develops through childhood and adolescence. Therefore, it may be an interesting future direction for research with the TORR.

Further Validity Research

While some discriminant, convergent, and predictive validity work has been completed with the TORR (Alexander et al., 2015; Dumas & Schmidt, 2015), important sources of evidence of the validity of the measure have yet to be investigated. Perhaps chief among these sources of validity information is a multi-trait, multi-method study. In a multi-trait, multi-method investigation, the target construct (i.e., relational reasoning) is measured using multiple methods (e.g., visuo-spatial and verbal items). Concomitantly, measures that use each of those methods, but measure different constructs (e.g., mental rotation and reading comprehension) are also given. Then, a latent variable measurement model that includes factors for both the abilities being tested and the methods being utilized is fit to the data. After the model is fit, the correlations among the latent factors are interpreted as sources of information concerning the validity of the measures.

Indeed, a verbal measure of relational reasoning is currently under development (Alexander, Singer, Jablansky, & Hattan, in press), so a multi-trait, multi-method investigation of the validity of both measures may be a logical next step relatively soon. That being said, perhaps the best evidence of validity for any psychoeducational measure comes from strong predictive relations to academically important outcome variables frequently being reported in the empirical literature. For this to happen, the TORR must be used widely in educational psychology research, something that is now beginning to occur (e.g., Kendeou & O'Brien, 2016).

Analysis of Error Patterns

Traditionally in psychometric research on the functioning of selected-response tests and their constituent items, analysis is focused on the correct item-response, and the probability that students will select that response (e.g., Embretson & Reise, 2013). However, it is also possible, and potentially interesting, to conduct analyses that are focused not on the correct answer choice, but on which incorrect answer choice students selected if they got the item wrong. If the cognitive process required to select the correct answer choice is well understood, incorrect choices may be created based on common deviations from that correct process, in order to ascertain which particular mistakes a given participant is prone to making. Indeed, advanced models within both the psychometric and cognitive-science literatures have been developed that could be used to accommodate such a testing strategy, including polytomous IRT models (e.g., Samejima, 1998) and cognitive diagnostic models (Brown & Burton, 1978).

A possible way to extend these methods to research questions related to cultural-fairness may be to fit such polytomous or diagnostic models across multiple demographic groups, and examine any systematic differences in error patterns among the groups. If such systematic differences in error patterns exist, that may be evidence that meaningful differences in the process of reasoning also exist among demographic groups. Such a finding may be explained by differential education or training relevant to cognitive assessment, varying cultural beliefs about the appropriate way to approach a cognitive task, or both. While deeply understanding cultural differences in the process of reasoning may require further follow-up studies using think-aloud or eye-tracking methods, a psychometric analysis of error patterns is a logical place to start in this potentially rich line of research inquiry.

Criterion Methods for Examining Fairness

As reviewed in Chapter 2, latent variable measurement model comparisons for DIF are not the only way to empirically investigate the cultural fairness of any given measure. Another important tool for cultural-fairness research is criterion methods, in which the possibility that a measure differentially predicts an outcome variable across demographic groups is tested. For example, if the TORR were to predict achievement in a particular academic domain strongly for females, but not for males, that may be evidence of differential validity of the TORR. Conversely, not significantly different predictive relations between the TORR and outcome variables across demographic groups would be evidence for the TORR's cultural fairness.

In the future, multi-group structural equation models may be constructed to formally test whether or not the predictive relation between latent relational reasoning ability and a particular outcome variable is the same across groups. Investigations of differential predictive validity are particularly important if the TORR is ever to be used as a selection tool in the academic or professional settings, because such selection decisions are based on the belief that the predictive validity of a measure is equal across all participants, regardless of their demographic background. For these reasons, it may be important for future psychometric research on TORR validity and fairness to proceed in conjunction with applications of the TORR for answering substantive research questions.

Future Applications of the TORR

As interest in the construct of relational reasoning grows among educational and psychological researchers, so too does the evidence of the importance of relational reasoning to educational outcomes (e.g., DeWolf, Bassok, & Holyoak, 2015). However, a variety of salient research questions concerning relational reasoning remain to be asked and investigated within the educational psychology literature. Some of these substantive research questions that may be addressed using the TORR are discussed here. For example, the relation between creativity and relational reasoning remains an open question in the literature, and the TORR may be a useful tool to uncover that relation. Second, the TORR may be used to investigate the role of relational reasoning in a number of domain-specific learning outcomes. Third, the specific cognitive or neurological processes associated with successful relational reasoning may be studied using items from the TORR. Finally, the TORR may be used as a measure within intervention research projects. Each of these future directions will now be further discussed.

Relation to Creativity

An increasing amount of empirical evidence (e.g., Green, Kraemer, Fugelsang, Gray & Dunbar, 2010, 2012) has shown that relational reasoning and creative thinking may be closely related constructs. For example, Dumas and Schmidt (2015) found that master's level engineering design students' scores on the TORR strongly predicted the originality of their design ideas. However, the particular relation between the two constructs is still relatively unknown. For example, it has been hypothesized that creative thinking requires the mental manipulation of higher-order relations, and therefore variation in relational reasoning ability within the population causes variation in creative thinking (Green et al., 2012). In contrast, others have argued that observed correlations between the two constructs only exist because of their indirect link through another, more basic attribute, such as speed-of-mental-processing (Rindermann & Neubauer, 2004).

One observation that complicates this question is that cognitive abilities, such as relational reasoning, appear to be correlated with creativity only at lower levels of relational reasoning ability (Jauk, Benedek, Dunst, & Neubauer, 2013). This "threshold theory" posits a particular level of reasoning ability at which creativity and reasoning are no longer correlated. However, what that specific level of reasoning ability may be is still unknown in the literature. Going forward, the TORR may be a useful measure to use in studies that seek to test the threshold theory, and to locate the particular threshold itself on the theta distribution. Further, measurement procedures and models that tap creative thinking as a multidimensional construct (e.g., Dumas & Dunbar, 2014) may be interesting to apply to such an investigation as well. In this way, such a studies would provide meaningful information about the nature of relational reasoning, creativity, and the relation between them.

Domain Specific Outcomes

As reviewed in Chapters 1 and 2, relational reasoning ability has been empirically linked to academic outcomes in a variety of domains of learning including reading (Ehri, Satlow, & Gaskins, 2009), chemistry (Bellochie & Ritchie, 2011; Trey & Khan, 2008), mathematics (DeWolf et al., 2015), engineering (Dumas & Schmidt, 2015), and medicine (Dumas et al., 2014). Given the findings from these investigations, domain-specific studies in other, yet-to-be studied domains may also find success in linking the construct to academic outcomes. Indeed, theoretically, it is difficult to determine if there are any domains of learning in which relational reasoning does not play a role. For that reason, explicating the particular mechanisms by which relational reasoning supports achievement within various academic domains may be an interesting endeavor for those researchers within the relational reasoning literature. For example, while STEM domains have typically been the focus of researchers in the field to-date, it is not yet known what role relational reasoning may play in the arts, including music composition, visual art, or literary writing. It is likely that students engaged in learning within these domains are also engaged in much higher-order cognitive processing, including relational reasoning. For that reason, investigations of the construct within those domains may be highly fruitful.

Modeling the Process of Reasoning

Within the cognitive science literature, a large body of work has sought to understand the specific cognitive processes by which individuals reason with analogies (see Holyoak, 2012 for a review). However, efforts to understand the cognitive processes underlying the other forms of relational reasoning—only recently theoretically identified—have been limited. One recent study (i.e., Grossnickle et al., 2016) found that the same componential processes identified as contributing to analogical reasoning were also identifiable within the other forms of relational reasoning, and that the same Bayesian network was able to describe the conditional probabilities associated with the successful completion of each of these componential processes across forms of relational reasoning. Therefore, the application of other various methodologies (e.g., computational modelling, neurological imaging) that have been previously used to understand analogy may be fruitfully applied to anomaly, antinomy, and antithesis.

Currently, it is not yet known whether the same types of computational systems can solve analogies and antinomies, for example, or whether the same neurological regions form the cortical substrate of anomaly and antithesis. For these reasons, expanding the already rich cognitive science literature on analogy to include the full multi-dimensional construct of relational reasoning may be especially interesting going forward. Such investigations may also serve to inform current questions about the role of particular cognitive processes (e.g., inductive, deductive, and abductive processes) in relational reasoning and its forms. In general, such an in-depth understanding of the cognitive and neurological mechanisms underlying successful relational reasoning can form a foundation for other lines of work, which more explicitly seek to support and improve the relational reasoning ability of students.

Intervention Research

While there is some evidence that relational reasoning is a malleable ability that can be improved in students through either direct (e.g., Alexander et al., 1987) or indirect (e.g., Murphy et al., 2016) intervention, no intervention research of which I am aware has sought to explicitly train students on each of the four forms of relational reasoning, and then measure gains in that construct or related academic variables. With appropriate experimental design, such an intervention may be an important test of the causal relation between relational reasoning ability and high-value academic outcome variables. Based on existing research (e.g., Begolli & Richland, 2015), it may be meaningful to incorporate instruction on strategies for reasoning with relations into wide-spread educational practice.

However, a causal link between gains on the construct and gains on academic outcome variables is necessary to demonstrate before such potential alterations to educational practice can be meaningfully posed. Importantly, all causal claims from future intervention work would be predicated on the reliability and validity of the measures utilized to assess relational reasoning and other abilities. For that reason, the development of the TORR is relevant to potential intervention work, and to the possibility of informing educational practice through research on relational reasoning.

Based on the findings of the present investigation, the TORR can be fairly used to measure relational reasoning across gender, language, and race/ethnicity groups, allowing TORR scores to be validity interpreted when drawn from a diverse group of students. This finding—that no significant DIF exists on the TORR across various sub-groups of the undergraduate population—is therefore an important step in the development of the TORR, and the research literature on relational reasoning.

Appendix A: Demographic Questionnaire

1. Age: _____

2. Sex: Male Female Other

3. Gender: Man Woman Other

4. Ethnicity:

– Non-Hispanic White
– Hispanic
– Black
– America Indian
– Asian/Pacific Islander
– Other (Please specify): __________________

5. Native English-speaker: Yes No

6. Year in School:

– Freshman
– Sophomore
– Junior
– Senior
– Other: ________

7. Major(s): _________________
8. Minor(s): _________________
9. Overall GPA: _______

Appendix B: Annotated Sample FlexMIRT Code

```
<Project>
Title = "Gender DIF";
Description = "Gender_DIF";

<Options>

// Main settings used across groups

Mode = Calibration;
NewThreadModel = Yes;
Quadrature = 21,5.0;
MaxE = 1000;
MaxM = 300;
Etol = 1e-3;
Mtol = 1e-3;
GOF = Complete;
SE = SEM;
Factorloadings = Yes;
NormalMetric3PL = Yes;

<Groups>

//Each group needs separate file and N, but identical specifications

%Male%
File ="Male_DIF.csv";
Varnames = v1-v32;
N = 700;
Ncats(v1-v32)=2;
Model(v1-v32) = ThreePL;
BetaPriors(v1-v32)= 2.0;
Dimensions = 5;
Primary = 1;

%Female%
File ="Female_DIF.csv";
Varnames = v1-v32;
N = 679;
Ncats(v1-v32)=2;
Model(v1-v32) = ThreePL;
BetaPriors(v1-v32)= 2.0;
Dimensions = 5;
Primary = 1;

<Constraints>

// Priors same for both groups

Prior Male, (v1-v32), Guessing: Normal(-1.09,0.5);
Prior Female, (v1-v32), Guessing: Normal(-1.09,0.5);

// Bifactor structure for males

Fix Male, (v1-v32),slope(1,2,3,4,5);
Free Male, (v1-v32),slope(1);
Free Male, (v1-v8),slope(2);
Free Male, (v9-v16),slope(3);
Free Male, (v17-v24),slope(4);
Free Male, (v25-v32),slope(5);
```

```
// Bifactor structure for females

Fix Female, (v1-v32),slope(1,2,3,4,5);
Free Female, (v1-v32),slope(1);
Free Female, (v1-v8),slope(2);
Free Female, (v9-v16),slope(3);
Free Female, (v17-v24),slope(4);
Free Female, (v25-v32),slope(5);

// Estimating latent means and variances for females (Males set to 0,1 by
default)

Free Female, Mean(1,2,3,4,5);
Free Female, Cov (1,1);
Free Female, Cov (2,2);
Free Female, Cov (3,3);
Free Female, Cov (4,4);
Free Female, Cov (5,5);

//Constraining parameters for referent items

Equal Male, (v1), Guessing : Female, (v1), Guessing;
Equal Male, (v1), Intercept : Female, (v1), Intercept;
Equal Male, (v1), Slope(1,2) : Female, (v1), Slope(1,2);

Equal Male, (v15), Guessing : Female, (v15), Guessing;
Equal Male, (v15), Intercept : Female, (v15), Intercept;
Equal Male, (v15), Slope(1,3) : Female, (v15), Slope(1,3);

Equal Male, (v24), Guessing : Female, (v24), Guessing;
Equal Male, (v24), Intercept : Female, (v24), Intercept;
Equal Male, (v24), Slope(1,4) : Female, (v24), Slope(1,4);

Equal Male, (v28), Guessing : Female, (v28), Guessing;
Equal Male, (v28), Intercept : Female, (v28), Intercept;
Equal Male, (v28), Slope(1,5) : Female, (v28), Slope(1,5);


// Constraining parameters of item being tested

Equal Male, (v2), Guessing : Female, (v2), Guessing;
Equal Male, (v2), Intercept : Female, (v2), Intercept;
Equal Male, (v2), Slope(1,2) : Female, (v2), Slope(1,2);
```

REFERENCES

Abad, F. J., Colom, R., Rebollo, I., & Escorial, S. (2004). Sex differential item functioning in the Raven's Advanced Progressive Matrices: evidence for bias. *Personality and Individual Differences*, *36*(6), 1459-1470. http://doi.org/10.1016/S0191-8869(03)00241-1

Acredolo, C., & Horobin, K. (1987). Development of relational reasoning and avoidance of premature closure. *Developmental Psychology*, *23*(1), 13-21. doi:10.1037/0012-1649.23.1.13

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*, *36*(3), 185-198. http://doi.org/10.1111/j.1745-3984.1999.tb00553.x

Alderman, D. L., & Powers, D. E. (1980). The effects of special preparation on SAT-Verbal scores. *American Educational Research Journal*, *17*(2), 239-251. http://doi.org/10.2307/1162485

Alexander, P. A. (2012). *The Test of Relational Reasoning*. College Park, MD: Disciplined Reading and Learning Research Laboratory.

Alexander, P. A., & The Disciplined Reading and Learning Research Laboratory. (2012). Reading into the future: Competence for the 21st century. *Educational Psychologist, 47*(4), 259-280. doi:10.1080/ 00461520.2012.722511.

Alexander, P. A., Dumas, D., Grossnickle, E. M**.,** List, A., & Firetto, C. (2015). Measuring relational reasoning. *Journal of Experimental Education, 83,* 1-33.

Alexander, P. A., White, C. S., Haensly, P. A., & Crimmins-Jeanes, M. (1987). Training in analogical reasoning. *American Educational Research Journal*, *24*(3), 387–404. http://doi.org/10.3102/00028312024003387

Ashcraft, M. H., & Moore, A. M. (2009). Mathematics anxiety and the affective drop in performance. *Journal of Psychoeducational Assessment*, *27*(3), 197–205. http://doi.org/10.1177/0734282908330580

Ardila, A. (2005). Cultural values underlying psychometric cognitive testing. *Neuropsychology Review*, *15*(4), 185–195. http://doi.org/10.1007/s11065-005-9180-y

Auyeung, B., Knickmeyer, R., Ashwin, E., Taylor, K., Hackett, G., & Baron-Cohen, S. (2011). Effects of fetal testosterone on visuospatial ability. *Archives of Sexual Behavior*, *41*(3), 571-581. http://doi.org/10.1007/s10508-011-9864-8

American Association for the Advancement of Science *(1993). Project 2061: Benchmarks for science literacy.* New York*:* Oxford University Press.

Atkins, S. M., Sprenger, A. M., Colflesh, G. J. H., Briner, T. L., Buchanan, J. B., Chavis, S. E., … Dougherty, M. R. (2014). Measuring working memory is all fun and games: A four-dimensional spatial game predicts cognitive task performance. *Experimental Psychology*, *61*(6), 417-438. doi:10.1027/1618-3169/a000262

Aubusson, P., Harrison, A. G., & Ritchie, S. (2006). *Metaphor and analogy in science education*. Dordrecht, The Netherlands: Springer.

Baker, S. T., Friedman, O., & Leslie, A. M. (2010). The opposites task: Using general rules to test cognitive flexibility in preschoolers. *Journal of Cognition and Development*, *11*(2), 240-254. doi:10.1080/15248371003699944

Baillargeon, R., & Graber, M. (1987). Where’s the rabbit? 5.5-month-old infants’ representation of the height of a hidden object. *Cognitive Development*, *2*(4), 375-392. doi:10.1016/S0885-2014(87)80014-X

Baldo, J. V., Bunge, S. A., Wilson, S. M., & Dronkers, N. F. (2010). Is relational reasoning dependent on language? A voxel-based lesion symptom mapping study. *Brain and Language*, *113*(2), 59-64. doi:10.1016/j.bandl.2010.01.004

Bashford, A., & Levine, P. (2010). *The Oxford handbook of the history of eugenics*. Oxford, UK: Oxford University Press.

Begolli, K. N., & Richland, L. E. (2015). Teaching mathematics by comparison: Analog visibility as a double-edged sword. *Journal of Educational Psychology*. http://doi.org/10.1037/edu0000056

Bellocchi, A., & Ritchie, S. M. (2011). Investigating and theorizing discourse during analogy writing in chemistry. *Journal of Research in Science Teaching*, *48*(7), 771-792. doi:10.1002/tea.20428

Bassok, M., Dunbar, K. N., & Holyoak, K. J. (2012). Introduction to the special section on the neural substrate of analogical reasoning and metaphor comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 261-263.

Bianchi, I., Savardi, U., & Kubovy, M. (2011). Dimensions and their poles: A metric and topological approach to opposites. *Language and Cognitive Processes, 26,* 1232-1265. doi: 10.1080/01690965.2010.520943

Bialystok, E., & Martin, M. M. (2004). Attention and inhibition in bilingual children: Evidence from the dimensional change card sort task. *Developmental Science*, *7*(3), 325-339. http://doi.org/10.1111/j.1467-7687.2004.00351.x

Binet, A., Simon, T., & Town, C. H. (1913). *A method of measuring the development of the intelligence of young children*. Chicago: Chicago Medical Book Company.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*(3), 261-280. http://doi.org/10.1177/014662168801200305

Borghese, P., & Gronau, R. C. (2005). Convergent and discriminant validity of the universal nonverbal intelligence test with limited English proficient Mexican-American elementary students. *Journal of Psychoeducational Assessment*, *23*(2), 128-139.

Bohn, R. E., & Short, J. E. (2009). *How much information? 2009 report on American consumers*. La Jolla, CA: UC San Diego Global Information Industry Center. Retrieved from http://hmi.ucsd.edu/pdf/HMI_2009_ConsumerReport_Dec9_2009.pd

Broughton, S. H., Sinatra, G. M., & Reynolds, R. E. (2010). The nature of the refutation text effect: An investigation of attention allocation. *The Journal of Educational Research*, *103*(6), 407-423. doi:10.1080/00220670903383101

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science*, *2*(2), 155–192. http://doi.org/10.1207/s15516709cog0202_4

Brown, A., & Croudace, T. J. (2015). Scoring and estimating score precision using multidimensional IRT models. In S. P. Reise, D. A., Revicki, (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment.* (pp. 307–333). New York: Routledge.

Bureau, US Census. (2010). Statistical Abstract of the United States: 2010. Retrieved February 25, 2016, from https://www.census.gov/library/publications/2009/compendia/statab/129ed.html

Burke, B. A., & Sunal, D. W. (2010). A framework to support Hispanic undergraduate women in STEM majors. In D. W. Sunal, C. S. Sunal, E. L. Wright, (Eds.), *Teaching science with Hispanic ELLs in K-16 classrooms.* (pp. 273-312). Charlotte: Information Age Publishing.

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 309-329. http://doi.org/10.1348/000711007X249603

Cai, L. (2013). *flexMIRT version 2: Flexible multilevel multidimensional item analysis and test scoring* [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical & Statistical Psychology*, *66*(2), 245-276. http://doi.org/10.1111/j.2044-8317.2012.02050.x

Cameron, I. M., Crawford, J. R., Lawton, K., & Reid, I. C. (2013). Differential item functioning of the HADS and PHQ-9: An investigation of age, gender and educational background in a clinical UK primary care sample. *Journal of Affective Disorders*, *147*(1-3), 262-268. http://doi.org/10.1016/j.jad.2012.11.015

Cattell, R. B. (1940). A culture-free intelligence test. *Journal of Educational Psychology*, *31*(3), 161-179. doi: 10.1037/h0059043

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1-22. http://doi.org/10.1037/h0046743

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action*. New York: Elsevier.

Camilli, G. (2013). Ongoing issues in test fairness. *Educational Research and Evaluation*, *19*(2-3), 104-120. http://doi.org/10.1080/13803611.2013.767602

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*, 404-431.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive psychology*, *4*(1), 55-81. doi:10.1016/0010-0285(73)90004-2

Chan, J., & Schunn, C. (2015). The impact of analogies on creative concept generation: Lessons from an in vivo study in engineering design. *Cognitive Science*, *39*(1), 126-155. http://doi.org/10.1111/cogs.12127

Chi, M. T. H. (2013, April). Thinking about relations in learning. In J. M. Kulikowich (Chair), *Exploring and leveraging relational thinking for academic performance.* Symposium conducted at the meeting of the American Educational Research Association, San Francisco.

Chi, M. T. H., & Roscoe, R. D. (2002). The processes and challenges of conceptual change. In M. Limon & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 3-27). Amsterdam: Kluwer.

Chi, M. T. H., & Slotta, J. D. (1993). The ontological coherence of intuitive physics. *Cognition and Instruction*, *10*(2-3), 249-260. doi: 10.1207/s1532690xci1002&3_5

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, *63*(1), 1-49. doi: 10.2307/1170558

Chen, J. (2005). *Effects of test anxiety, time pressure, ability and gender on response aberrance*. Dissertation abstracts international. ProQuest Information & Learning, US.

Chen, C., Zhou, X., Chen, C., Dong, Q., Zang, Y., Qiao, S., Yang, T., et al. (2007). The neural basis of processing anomalous information. *NeuroReport: For Rapid Communication of Neuroscience Research*, *18*(8), 747-751. doi:10.1097/WNR.0b013e3280ebb49b

Cho, S., Holyoak, K. J., & Cannon, T. D. (2007). Analogical reasoning in working memory: Resources shared among relational integration, interference resolution, and maintenance. *Memory & Cognition*, *35*(6), 1445-1455. doi:10.3758/BF03193614

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, *5*(2), 115-124.

Cole, N. S., & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement*, *38*(4), 369-382. http://doi.org/10.1111/j.1745-3984.2001.tb01132.x

Cole, M., & Wertsch, J. V. (1996). Beyond the individual-social antinomy in discussions of Piaget and Vygotsky. *Human Development*, *39*(5), 250-256. doi:10.1159/000278475

Collins, M. A., & Laski, E. V. (2015). Preschoolers' strategies for solving visual pattern tasks. *Early Childhood Research Quarterly*, *32*, 204-214. http://doi.org/10.1016/j.ecresq.2015.04.004

Crone, E. A., Wendelken, C., van Leijenhorst, L., Honomichl, R. D., Christoff, K., & Bunge, S. A. (2009). Neurocognitive development of relational reasoning. *Developmental Science*, *12*(1), 55-66. doi: 10.1111/j.1467-7687.2008.00743.x

DeVellis, R. F. (2006). Classical test theory. *Medical Care*, *44*(11), 50-59.

DeWolf, M., Bassok, M., & Holyoak, K. J. (2015). Conceptual structure and the procedural affordances of rational numbers: Relational reasoning with fractions and decimals. *Journal of Experimental Psychology: General*, *144*(1), 127-150. http://doi.org/10.1037/xge0000034

Diamond, M. (2004). Sex, gender, and identity over the years: A changing perspective. *Child and Adolescent Psychiatric Clinics of North America*, *13*(3), 591-607. http://doi.org/10.1016/j.chc.2004.02.008

Dixon-Román, E. J., Everson, H. T., & McArdle, J. J. (2013). Race, poverty and SAT scores: Modeling the influences of family income on black and white high school students' SAT performance. *Teachers College Record*, *115*(4), 1-19.

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

DuBois, W. E. B. (2013). *W. E. B. DuBois on sociology and the Black community*. Chicago: University of Chicago Press.

Dumas, D., Alexander, P. A., (2016). Calibration of the Test of Relational Reasoning. *Psychological Assessment*.

Dumas, D., Alexander, P. A., & Grossnickle, E. M. (2013). Relational reasoning and its manifestations in the educational context: A systematic review of the literature. *Educational Psychology Review, 25*, 391-427. doi: 10.1007/s10648-013-9224-4.

Dumas, D., & Schmidt, L. (2015). Relational reasoning as predictor for engineering ideation success using analogies in TRIZ. *Journal of Engineering Design*.

Dumas, D., Alexander, P. A., & Schmidt, L. (under review). Predicting creative problem-solving in mechanical engineering. *Thinking Skills and Creativity.*

Dumas, D., Alexander, P. A., Baker, L. M., Jablansky, S., & Dunbar, K. N. (2014). Relational reasoning in medical education: Patterns in discourse and diagnosis. *Journal of Educational Psychology, 106,* 1021-1035. doi: 10.1037/a003677

Dumontheil, I., Houlton, R., Christoff, K., & Blakemore, S. J. (2010). Development of relational reasoning during adolescence. *Developmental Science*, *13*(6), 15-24. doi:10.1111/j.1467-7687.2010.01014.x

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R.J. Sternberg, & J. Davidson (Eds.). *The nature of insight* (pp. 365-396). Cambridge MA: MIT press.

Ehri, L. C., Satlow, E., & Gaskins, I. (2009). Grapho-phonemic enrichment strengthens keyword analogy instruction for struggling young readers. *Reading and Writing Quarterly: Overcoming Learning Difficulties*, *25*(2-3), 162-191. doi: 10.1080/10573560802683549

Elliott, R. (1987). *Litigating intelligence: IQ tests, special education, and social science in the courtroom*. New York: Auburn House Publishing Group.

Elosua, P. (2011). Assessing measurement equivalence in ordered-categorical data. *Psicológica*, *32*(2), 403–421.

Embretson, S. E., & Reise, S. P. (2012). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fast, K. V., & Campbell, G. (2004). "I still like Google:" University students' perceptions of searching OPACs and the web. *Proceedings of the American Society for Information Science and Technology, 41*, 138-146.

Faustmann, A., Murdoch, B. E., Finnigan, S. P., & Copland, D. A. (2007). Effects of advancing age on the processing of semantic anomalies in adults: Evidence from event-related brain potentials. *Experimental Aging Research*, *33*(4), 439-460. doi:10.1080/03610730701525378

Farrington-Flint, L., & Wood, C. (2007). The role of lexical analogies in beginning reading: Insights from children's self-reports. *Journal of Educational Psychology, 99*, 326-338. doi: 10.1037/0022-0663.99.2.326.

Filik, R. (2008). Contextual override of pragmatic anomalies: Evidence from eye movements. *Cognition*, *106*(2), 1038-1046. doi:10.1016/j.cognition.2007.04.006

Filik, R., & Leuthold, H. (2008). Processing local pragmatic anomalies in fictional contexts: Evidence from the N400. *Psychophysiology*, *45*(4), 554-558. doi:10.1111/j.1469-8986.2008.00656.x

Fischer, B. A. (2012). Maltreatment of people with serious mental illness in the early 20th century: A focus on Nazi Germany and eugenics in America. *Journal of Nervous and Mental Disease*, *200*(12), 1096-1100.

Frane, A. V. (2015). Power and Type I error control for univariate comparisons in multivariate two-group designs. *Multivariate Behavioral Research*, *50*(2), 233-247. http://doi.org/10.1080/00273171.2014.968836

Freedle, R., & Kostin, I. (1997). Predicting black and white differential item functioning in verbal analogy performance. *Intelligence*, *24*(3), 417-444. http://doi.org/10.1016/S0160-2896(97)90058-1

Freeman, B. (2012). Using digital technologies to redress inequities for English language learners in the English speaking mathematics classroom. *Computers & Education*, *59*(1), 50-62. http://doi.org/10.1016/j.compedu.2011.11.003

Gagné, P., & Hancock, G. R. (2010). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*, *41*(1), 65-83. doi:10.1207/s15327906mbr4101_5

Galton, F. (1869). *Hereditary genius: an inquiry into its laws and consequences*. London: Macmillan.

Gardner, H. (1995). Perennial antinomies and perpetual redrawings: Is there progress in the study of mind? In R. Solso & D. Massaro (Eds.), *The science of the mind: 2001 and beyond* (pp. 65-78). New York: Oxford University Press.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155-170. doi:10.1207/s15516709cog0702_3

Gentner, D., & Gentner, D. R. (1983). Flowing waters or teeming crowds: Mental models of electricity. In D. Gentner & A. L. Stevens (Eds.), *Mental models*. Hillsdale, NJ: Erlbaum.

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, *95*(2), 393-405. doi:10.1037/0022-0663.95.2.393

Goel, V., & Dolan, R. J. (2001). Functional neuroanatomy of three-term relational reasoning. *Neuropsychologia*, *39*(9), 901-909. doi:10.1016/S0028-3932(01)00024-0

Goswami, U. (1992). *Analogical reasoning in children.* Hove, East Sussex, UK: Lawrence Erlbaum Associates.

Goswami, U. (2013). The development of reasoning by analogy. In P. Barrouillet, & C. Gauffroy, (Eds.), *The development of thinking and reasoning.* (pp. 49-70). New York: Psychology Press.

Goswami, U., & Bryant, P. (1992). Rhyme, analogy, and children’s reading. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition.* (pp. 49-63). Hillsdale, NJ: Lawrence Erlbaum Associates.

Goswami, U., & Mead, F. (1992). Onset and rime awareness and analogies in reading. *Reading Research Quarterly*, *27*(2), 152-162. doi:10.2307/747684

Gould, S. J. (1996). *The mismeasure of man*. New York: W. W. Norton & Company.

Grant, D. A., & Berg, E. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of Experimental Psychology*, *38*(4), 404-411. http://doi.org/10.1037/h0059831

Green, A. E., Kraemer, D. J. M., Fugelsang, J. A., Gray, J. R., & Dunbar, K. N. (2010). Connecting long distance: Semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cerebral Cortex*, *20*(1), 70–76. http://doi.org/10.1093/cercor/bhp081

Green, A. E., Kraemer, D. J. M., Fugelsang, J. A., Gray, J. R., & Dunbar, K. N. (2012). Neural correlates of creativity in analogical reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 264–272. http://doi.org/10.1037/a0025764

Greenfield, P. M. (1997). You can’t take it with you: Why ability assessments don’t cross cultures. *American Psychologist*, *52*(10), 1115-1124. http://doi.org/10.1037/0003-066X.52.10.1115

Greiff, S., Stadler, M., Sonnleitner, P., Wolff, C., & Martin, R. (2015). Sometimes less is more: Comparing the validity of complex problem solving measures. *Intelligence*, *50*, 100-113. http://doi.org/10.1016/j.intell.2015.02.007

Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational contexts—Something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology*, *105*(2), 364-379. http://doi.org/10.1037/a0031856

Grossnickle, E. M.**,** Dumas, D.**,** & Alexander, P. A. (2015, August). *Getting to the source: What contributes to relational reasoning performance?* Paper to be presented at the annual meeting of the European Association for Research on Learning and Instruction, Limassol, Cyprus.

Hancock, G. R. (1999). A sequential Scheffé-type respecification procedure for controlling Type I error in exploratory structural equation model modification. *Structural Equation Modeling*, *6*(2), 158-168. http://doi.org/10.1080/10705519909540126

Hancock, G. R., & Freeman, M. J. (2001). Power and sample size for the root mean square error of approximation test of not close fit in structural equation modeling. *Educational and Psychological Measurement*, *61*(5), 741-758.

Hancock, G. R., & French, B. F. (2013). Power analysis in structural equation modeling. In G. R. Hancock, R. O. Mueller, (Eds.), *Structural equation modeling: A second course (2nd ed.).* (pp. 117-159). Charlotte: Information Age Publishing.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudek, S.H.C du Toit, & D. Sörbum (Eds.), *Structural equation modeling: Past and present. A festchrift in honor of Karl G. Jörestog.* (pp. 195-261). Chicago: Scientific Software International.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H. I. Braun, (Eds.), *Test Validity*. (pp.129-145). Hillsdale, NJ: Erlbaum.

Holyoak, K. J. (1985). The pragmatics of analogical transfer. In G. H. Bower (Ed.), *The psychology of learning and motivation*, (Vol. 19, pp. 59-87). New York: Academic Press.

Griffiths, J. R., & Brophy, P. (2005). Student searching behavior and the web: use of academic resources and Google. *Library Trends*, *53*(4), 539-554

Hein, S., Reich, J., Thuma, P. E., & Grigorenko, E. L. (2014). Physical growth and nonverbal intelligence: Associations in Zambia. *The Journal of Pediatrics, 165*(5), 1017-1023. doi.org: 10.1016/j.jpeds.2014.07.058

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds), *The Oxford handbook of thinking and reasoning* (pp. 234-259). New York: Oxford University Press.

Hofstadter, D .R., (2001). Epilogue: Analogy as the core of cognition. In D. Gentner, K. J. Holyoak, & B. N. Kokinov, (Eds.) *The analogical mind: Perspectives from cognitive science*, (pp. 499-538). Cambridge, MA: MIT Press.

Hummel, J. E., & Holyoak, K. J. (2005). Relational reasoning in a neurally plausible cognitive architecture: An overview of the LISA project. *Current Directions in Psychological Science*, *14*(3), 153-157. doi:10.1111/j.0963-7214.2005.00350.x

Iverson, S. V. (2007). Camouflaging power and privilege: a critical race analysis of university diversity policies. *Educational Administration Quarterly*, *43*(5), 586-611. http://doi.org/10.1177/0013161X07307794

James, W. (1890). *The principles of psychology*. New York: Henry Holt and Company.

Jauk, E., Benedek, M., Dunst, B., & Neubauer, A. C. (2013). The relationship between intelligence and creativity: New support for the threshold hypothesis by means of empirical breakpoint detection. *Intelligence*, *41*(4), 212–221. http://doi.org/10.1016/j.intell.2013.03.003

Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, *39*(1), 1-123.

Jiang, S., Li, J., Liu, X., Qi, S., & Yang, Y. (2015). An ERP study of advantage effect differences on task switching in proficient and non-proficient bilinguals. *Acta Psychologica Sinica*, *47*(6), 746-756.

Kaufman, S. B. (2007). Sex differences in mental rotation and spatial visualization ability: Can they be accounted for by differences in working memory capacity? *Intelligence*, *35*(3), 211-223. http://doi.org/10.1016/j.intell.2006.07.009

Kelley, K. (2013). Effect size and sample size planning. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods (Vol 1): Foundations,* (pp. 206-222). Oxford, UK: Oxford University Press.

Kendeou, P., & O'Brien**, N**. (2015, April). Antithetical reasoning with refutational texts. In G. M. Sinatra (Chair), *Relational reasoning in STEM domains: What empirical research can contribute to the national dialogue.* Symposium conducted at the annual meeting of the American Educational Research Association, Chicago, IL.

Kenneavy, K. (2013). New contours in the mediated sex/gender/sexuality landscape. *Sex Roles*, *68*(9-10), 620-622. http://doi.org/10.1007/s11199-012-0250-3

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(2), 212–228. http://doi.org/10.1080/10705511.2011.557337

Klahr, D., & Dunbar, K. (1988). The psychology of scientific discovery: Search in two problem spaces. *Cognitive Science, 12*, 1-48.

Klineberg, O. (1935). *Race differences*. Oxford, UK: Harper.

Klockars, A. J., & Hancock, G. R. (1994). Per-experiment error rates: The hidden costs of several multiple comparison procedures. *Educational and Psychological Measurement*, *54*(2), 292-298. http://doi.org/10.1177/0013164494054002004

Koo, J., Becker, B. J., & Kim, Y. S. (2014). Examining differential item functioning trends for English language learners in a reading test: A meta-analytical approach. *Language Testing*, *31*(1), 89-109. http://doi.org/10.1177/0265532213496097

Krawczyk, D. C. (2012). The cognition and neuroscience of relational reasoning. *Brain Research*, *1428*, 13-23. doi: 10.1016/j.brainres.2010.11.080

Krawczyk, D. C., McClelland, M. M., & Donovan, C. M. (2011). A hierarchy for relational reasoning in the prefrontal cortex. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior*, *47*(5), 588-597. doi: 10.1016/j.cortex.2010.04.008

Krawczyk, D. C., Morrison, R. G., Viskontas, I., Holyoak, K. J., Chow, T. W., Mendez, M. F., … Knowlton, B. J. (2008). Distraction during relational reasoning: The role of prefrontal cortex in interference control. *Neuropsychologia*, *46*(7), 2020-2032. http://doi.org/10.1016/j.neuropsychologia.2008.02.001

Kroll, J. F., & Fricke, M. (2014). What bilinguals do with language that changes their minds and their brains. *Applied Psycholinguistics*, *35*(5), 921-925. http://doi.org/10.1017/S0142716414000253

Kuhn, D., & Udell, W. (2007). Coordinating own and other perspectives in argument. *Thinking and Reasoning*, *13*(2), 90-104. doi:10.1080/13546780600625447.

Kulkarni, D., & Simon, H. A. (1988). The processes of scientific discovery: The strategy of experimentation. *Cognitive Science*, *12*(2), 139-175. doi:10.1207/s15516709cog1202_1

Kumar, S., Cervesato, I., & Gonzalez, C. (2014). How people do relational reasoning? Role of problem complexity and domain familiarity. *Computers in Human Behavior*, *41*, 319-326. doi:10.1016/j.chb.2014.09.015

Lee, H. S., & Holyoak, K. J. (2008). The role of causal models in analogical inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1111-1122. doi:10.1037/a0012581

Livingston, S. A. (2006). Item analysis. In S. M. Downing, & T. M. Haladyna, (Eds.), *Handbook of test development,* (pp. 421-441). Hillsdale, NJ: Erlbaum.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York: Routledge.

Maguire, M. J., McClelland, M. M., Donovan, C. M., Tillman, G. D., & Krawczyk, D. C. (2012). Tracking cognitive phases in analogical reasoning with event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 273-281. doi:10.1037/a0025485

Mann, H. M., Rutstein, D. W., & Hancock, G. R. (2009). The potential for differential findings among invariance testing strategies for multisample measured variable path models. *Educational and Psychological Measurement*, *69*(4), 603-612. http://doi.org/10.1177/0013164408324470

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies. *Journal of National Cancer Institute, 22*(4), 719-748.

Markman, A. B., & Gentner, D. (2000). Structure mapping in the comparison process. *The American Journal of Psychology*, *113*(4), 501-538. doi:10.2307/1423470

McCallum, R. S. (2003). The universal nonverbal intelligence test. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 87-111). Springer: New York.

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, *11*, 19-35.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130-149. http://doi.org/10.1037/1082-989X.1.2.130

McCrink, K., Spelke, E. S., Dehaene, S., & Pica, P. (2013). Non-symbolic halving in an Amazonian indigene group. *Developmental Science*, *16*(3), 451-462. http://doi.org/10.1111/desc.12037

McGuire, F. (1994). Army alpha and beta tests of intelligence. In R. Sternberg (Ed.), *Encyclopedia of human intelligence*, (pp. 125-129). New York: Macmillan.

Messick, S. (1980). Test validity and the ethics of assessment. In D. N. Bersoff (Ed.), *Ethical conflicts in psychology*, (pp. 273-275). Washington:American Psychological Association.

Miller, D. I., & Halpern, D. F. (2014). The new science of cognitive sex differences. *Trends in Cognitive Sciences*, *18*(1), 37-45. http://doi.org/10.1016/j.tics.2013.10.011

Miller W. S. (*1*960). *Technical manual for the Miller Analogies Test.* New York*:* Psychological Corporation.

Mitchell, M. (1993). *Analogy-making as perception: A computer model*. Cambridge, MA: MIT Press.

Muthén, L.K. & Muthén, B.O. (2012). *Mplus: Seventh Edition.* [Computer Software]. Los Angeles: Muthén & Muthén

Murphy, P. K., Greene, J. A., Firetto, C. M., Montalbano, C., Mengyi, L., Wei, L., & Croninger, R. M. V. (April, 2016) Promoting relational reasoning in elementary students' writing. In D. Dumas (Chair), *The malleability of relational reasoning.* Symposium conducted at the annual meeting of the American Educational Research Association, Washington, DC.

Murphy, P. K., & Mason, L. (2006). Changing knowledge and beliefs. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 305-324). Mahwah, NJ. Lawrence Erlbaum Associates.

Naglieri, J. A., & Insko, W. R. (1986). Construct validity of the Matrix Analogies Test—Expanded Form. *Journal of Psychoeducational Assessment*, *4*(3), 243-255. doi: 10.1177/073428298600400308

Nazareth, A., Herrera, A., & Pruden, S. M. (2013). Explaining sex differences in mental rotation: Role of spatial activity experience. *Cognitive Processing*, *14*(2), 201-204. http://doi.org/10.1007/s10339-013-0542-8

Nisbett, R. E. (2009). *Intelligence and How to Get it: Why Schools and Cultures Count*. W. W. Norton & Company: New York.

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 510-520.

Nye, R. A. (2010). How sex became gender. *Psychoanalysis and History*, *12*(2), 195-209. http://doi.org/10.3366/pah.2010.0005

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland, & H. Wainer, (Eds.), *Differential item functioning.* (pp. 255-276). Hillsdale, NJ: Erlbaum.

Padilla, A. M., & Borsato, G. N. (2008). Issues in culturally appropriate psychoeducational assessment. In L. A. Suzuki, & J. G. Ponterotto, (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications.* (pp. 5-21). San Francisco: Jossey-Bass.

Palumbo, M. V., & Steele-Johnson, D. (2014). Do test perceptions influence test performance? Exploring stereotype threat theory. *North American Journal of Psychology, 16*(1), 1-12.

Patrick, C. J., Hicks, B. M., Nichol, P. E., & Krueger, R. F. (2007). A bifactor approach to modeling the structure of the psychopathy checklist-revised. *Journal of Personality Disorders*, *21*(2), 118–141. http://doi.org/10.1521/pedi.2007.21.2.118

Piaget, J. (1928/1966). *Judgment and reasoning in the child*. Totowa, NJ: Littlefield, Adams, & Co.

PBS Newshour Online. (2010, December 10). Math, science, reading scores show U.S. schools slipping behind. Retrieved from http://www.pbs.org/newshour/extra/features/us/july-dec10/education_12-10.html

Pashler, H. E. (1998). *Attention*. New York: Psychology Press.

Penfield, R. D., & Camilli, G. (2006). Differential item functioning and item bias. In S. Sinharay (Ed.), *Handbook of Statistics,* (Vol. 26, pp. 125-167). Netherlands: Elsevier.

Pontius, A. A. (1997). Lack of sex differences among east Ecuadorian school children on geometric figure rotation and face drawings. *Perceptual and Motor Skills, 85*(1), 72-74. http://doi.org/10.2466/PMS.85.5.72-74

Poortinga, Y. H. (1995). Cultural bias in assessment: Historical and thematic issues. *European Journal of Psychological Assessment*, *11*(3), 140-146. http://doi.org/10.1027/1015-5759.11.3.140

Preacher, K. J., & Coffman, D. L. (2006). Computing power and minimum sample size for RMSEA [Computer software]. Available from http://quantpsy.org/

Preacher, K. J., Cai, L., & MacCallum, R. C. (2007). Alternatives to traditional model comparison strategies for covariance structure models. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 33-62). Mahwah, NJ: Lawrence Erlbaum Associates.

Raven, J. C. (1941). Standardization of progressive matrices. *British Journal of Medical Psychology*, *19*, 137-150. doi:10.1111/j.2044-8341.1941.tb00316.x

Ratiu, I., & Azuma, T. (2015). Working memory capacity: Is there a bilingual advantage? *Journal of Cognitive Psychology*, *27*(1), 1-11. http://doi.org/10.1080/20445911.2014.976226

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Reynolds, C. R., & Ramsay, M. C. (2003). Bias in psychological assessment: An empirical review and recommendations. In J. R. Graham, & J. A. Naglieri, (Eds.), *Handbook of psychology: Assessment psychology,* (pp. 67-93). Hoboken, NJ: John Wiley & Sons.

Rice, W. R. (1989). Analyzing tables of statistical tests. *Evolution*, *43*(1), 223-225. http://doi.org/10.2307/2409177

Richards, S. (2007). What strategies can I incorporate so that the English language learners in my classroom will better understand oral directions? In C. Caro-Bruce, R. Flessner, M. Klehr, & K. Zeichner, (Eds.), *Creating equitable classrooms through action research.* (pp. 59-77). Thousand Oaks, CA: Corwin Press.

Richland, L. E., Chan, T. K., Morrison, R. G., & Au, T. K. F. (2010). Young children's analogical reasoning across cultures: Similarities and differences. *Journal of Experimental Child Psychology, 105*(1-2), 146-153. doi:10.1016/j.jecp.2009.08.003

Richland, L. E., & McDonough, I. M. (2010). Learning by analogy: Discriminating between potential analogs. *Contemporary Educational Psychology*, *35*(1), 28-43. doi:10.1016/j.cedpsych.2009.09.001

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology*, *94*(3), 249-273. http://doi.org/10.1016/j.jecp.2006.02.002

Richland, L. E., & Simms, N. (2015). Analogy, higher order thinking, and education. *WIREs Cognitive Science*, *6*(2), 177-192. http://doi.org/10.1002/wcs.1336

Richland, L. E., Zur, O., & Holyoak, K. J. (2007). Cognitive supports for analogies in the mathematics classroom. *Science*, *316*(5828), 1128-1129. doi:10.1126/science.1142103

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, *47*(3), 361-372. doi:10.1111/j.1745-3984.2010.00118.x

Rindler, S. E. (1980). The effects of skipping over more difficult items on time-limited tests: Implications for test validity. *Educational and Psychological Measurement*, *40*(4), 989–998. http://doi.org/10.1177/001316448004000425

Rindermann, H., & Neubauer, A. C. (2004). Processing speed, intelligence, creativity, and school performance: Testing of causal hypotheses using structural equation models. *Intelligence*, *32*(6), 573–589. http://doi.org/10.1016/j.intell.2004.06.005

Rivas, G. E. L., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement*, *33*(4), 251–265. http://doi.org/10.1177/0146621608321760

Rosselli, M., & Ardila, A. (2003). The impact of culture and education on non-verbal neuropsychological measurements: A critical review. *Brain and Cognition*, *52*(3), 326-333. http://doi.org/10.1016/S0278-2626(03)00170-2

Russell, B., & Lackey, D. (1973). *Essays in analysis*. New York, NY: Allen & Unwin.

Shaunessy, E., Karnes, F. A., & Cobb, Y. (2004). Assessing potentially gifted students from lower socioeconomic status with nonverbal measures of intelligence. *Perceptual and Motor Skills*, *98*(3c), 1129–1138. http://doi.org/10.2466/pms.98.3c.1129-1138

Sagi, E., Gentner, D., & Lovett, A. (2012). What difference reveals about similarity. *Cognitive Science*, *36*(6), 1019-1050. doi:10.1111/j.1551-6709.2012.01250.

Saklofske, D. H., Vijver, F. J. R., Oakland, T., Mpofu, E., & Suzuki, L. A. (2015). Intelligence and culture: history and assessment. In S. Goldstein, D. Princiotta, & J. A. Naglieri (Eds.), *Handbook of Intelligence* (pp. 341-365). Springer: New York.

Samejima, F. (1998). Efficient nonparametric approaches for estimating the operating characteristics of discrete item responses. *Psychometrika*, *63*(2), 111–130. http://doi.org/10.1007/BF02294770

Sax, L. J., Gilmartin, S. K., & Bryant, A. N. (2003). Assessing response rates and nonresponse bias in web and paper surveys. *Research in Higher Education*, *44*(4), 409–432. http://doi.org/10.1023/A:1024232915870

Shuwairi, S. M., Bainbridge, R., & Murphy, G. L. (2014). Concept formation and categorization of complex, asymmetric, and impossible figures. *Attention, Perception, & Psychophysics, 76*(6), 1789-1802. http://doi.org/10.3758/s13414-014-0691-6

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

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*(2), 199-223. http://doi.org/10.1037/1082-989X.1.2.199

Schmidt, W. H., McKnight, C. C., Cogan, L. S., Jakwerth, P. M., & Houang, R. T. (1999). *Facing the consequences: Using TIMSS for a closer look at US mathematics and science education*. New York: Kluwer Academic Publishers.

Simon, H. A., & Lea, G. (1974). Problem solving and rule induction: A unified view. In L. W. Gregg (Ed.), *Knowledge and cognition.* Oxford England: Lawrence Erlbaum.

Sinatra, G. M. (Chair, 2015, April), *Relational reasoning in STEM domains: What empirical research can contribute to the national dialogue.* Symposium conducted at the annual meeting of the American Educational Research Association, Chicago, IL.

Sinatra, G. M., & Broughton, S. H. (2011). Bridging reading comprehension and conceptual change in science education: The promise of refutation text. *Reading Research Quarterly*, *46*(4), 374-393.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31*(7), 15-21.

Sorensen, R. A. (2003). *A brief history of the paradox: Philosophy and the labyrinths of the mind*. New York: Oxford University Press.

Senturk, N., Yeniceri, N., Alp, I. E., & Altan-Atalay, A. (2014). An exploratory study on the Junior Brixton Spatial Rule Attainment Test in 6- to 8-year-olds. *Journal of Psychoeducational Assessment*, *32*(2), 123-132. doi:10.1177/0734282913490917

Spearman, C. (1904). General intelligence objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201-292. http://doi.org/10.2307/1412107

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York: MacMillan.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4-28. http://doi.org/10.1006/jesp.1998.1373

Spironelli, C., Segrè, D., Stegagno, L., & Angrilli, A. (2014). Intelligence and psychopathy: A correlational study on insane female offenders. *Psychological Medicine*, *44*(1), 111-116.

Sprenger, A. M., Atkins, S. M., Bolger, D. J., Harbison, J. I., Novick, J. M., Chrabaszcz, J. S., Dougherty, M. R. (2013). Training working memory: Limits of transfer. *Intelligence, 41*(5), 638-663. doi: 10.1016/j.intell.2013.07.013

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292-1306. http://doi.org/10.1037/0021-9010.91.6.1292

Steele, C. (2003). Race and the schooling of Black Americans. In S. Plous (Ed.), *Understanding prejudice and discrimination.* (pp. 98-107). New York: McGraw-Hill.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797-811. http://doi.org/10.1037/0022-3514.69.5.797

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. Oxford, UK: Lawrence Erlbaum Stevenson.

Sternberg, R. J. (2004). Culture and intelligence. *American Psychologist*, *59*(5), 325-338. http://doi.org/10.1037/0003-066X.59.5.325

Sternberg, R. J. (2008). Culture, instruction, and assessment. In J. Elliott, & E. Grigorenko, (Eds.), *Western psychological and educational theory in diverse contexts.* (pp. 5-22). New York: Routledge.

Sternberg, R. J. (1986). Haste makes waste versus a stitch in time? A reply to Vernon, Nador, and Kantor. *Intelligence*, *10*(3), 265–270. http://doi.org/10.1016/0160-2896(86)90020-6

Sternberg, R. J., & Rifkin, B. (1979). The development of analogical reasoning processes. *Journal of Experimental Child Psychology*, *27*(2), 195-232. http://doi.org/10.1016/0022-0965(79)90044-4

Stephens, A. C. (2006). Equivalence and relational thinking: Preservice elementary teachers' awareness of opportunities and misconceptions. *Journal of Mathematics Teacher Education*, *9*(3), 249-278. doi: 10.1007/s10857-006-9000-1

Stevenson, C. E., Bergwerff, C. E., Heiser, W. J., & Resing, W. C. M. (2014). Working memory and dynamic measures of analogical reasoning as predictors of children's math and reading achievement. *Infant and Child Development*, *23*(1), 51-66. http://doi.org/10.1002/icd.1833

Stewart, A. J., Kidd, E., & Haigh, M. (2009). Early sensitivity to discourse-level anomalies: Evidence from self-paced reading. *Discourse Processes*, *46*(1), 46-69. doi:10.1080/01638530802629091

Strømsø, H. I., & Bråten, I. (2010). The role of personal epistemology in the self-regulation of internet-based learning. *Metacognition and Learning*, *5*(1), 91-111.

Stulberg, L. M. (2015). African American school choice and the current race politics of charter schooling: Lessons from history. *Race and Social Problems*, *7*(1), 31-42. http://doi.org/10.1007/s12552-014-9133-2

Trey, L., & Khan, S. (2008). How science students can learn about unobservable phenomena using computer-based analogies. *Computers and Education*, *51*(2), 519-529. doi:10.1016/j.compedu.2007.05.019

Trickett, S. B., Trafton, J. G., & Schunn, C. D. (2009). How do scientists respond to anomalies? Different strategies used in basic and applied science. *Topics in Cognitive Science*, *1,* 711-729. doi: 10.1111/j.1756-8765.2009.01036.x

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning.* (pp. 67-113). Hillsdale, NJ: Erlbaum.

Thurstone, L. L. (1947). *Multiple-factor analysis; a development and expansion of The Vectors of Mind*. Chicago, IL, US: University of Chicago Press.

Tomes, Y. I. (2010). Culture and psychoeducational assessment: Cognition and achievement. In E. García-Vásquez, T. D. Crespi, & C. A. Riccio, (Eds.), *Handbook of education, training, and supervision of school psychologists in school and community, Vol 1: Foundations of professional practice.* (pp. 167-183). New York: Routledge.

Tunteler, E., & Resing, W. C. M. (2010). The effects of self and other scaffolding on progression and variation in children’s geometric analogy performance: A microgenetic research. *Journal of Cognitive Education and Psychology*, *9*(3), 251-272. doi:10.1891/1945-8959.9.3.251

Tzuriel, D., & Flor-Maduel, H. (2010). Prediction of early literacy by analogical thinking modifiability among kindergarten children. *Journal of Cognitive Education and Psychology*, *9*(3), 207-226. http://doi.org/10.1891/1945-8959.9.3.207

Ulrich, R., Mattes, S., & Miller, J. (1999). Donders’s assumption of pure insertion: an evaluation on the basis of response dynamics. *Acta Psychologica*, *102*(1), 43–76. http://doi.org/10.1016/S0001-6918(99)00019-0

Van der Henst, J. B., & Schaeken, W. (2005). The wording of conclusions in relational reasoning. *Cognition*, *97*(1), 1-22. doi:10.1016/j.cognition.2004.06.008

van Gog, T., Paas, F., & van Merriënboer, J. J. G. (2004). Process-oriented worked examples: Improving transfer performance through enhanced understanding. *Instructional Science*, *32*(1), 83-98. doi:10.1023/B:TRUC.0000021810.70784

Vars, F. E., & Bowen, W. G. (1998). Scholastic aptitude test scores, race, and academic performance in selective colleges and universities. In C. Jencks, & M. Phillips, (Eds.), *The Black-White test score gap.* (pp. 457-479). Washington: Brookings Institute Press.

Verney, S. P., Granholm, E., Marshall, S. P., Malcarne, V. L., & Saccuzzo, D. P. (2005). Culture-fair cognitive ability assessment: Information processing and psychophysiological approaches. *Assessment*, *12*(3), 303-319. http://doi.org/10.1177/1073191105276674

Vernon, P. A. (1986). He who doesn’t believe in speed should beware of hasty judgments: A reply to Sternberg. *Intelligence*, *10*(3), 271–275. http://doi.org/10.1016/0160-2896(86)90021-8

Vock, M., & Holling, H. (2008). The measurement of visuospatial and verbal-numerical working memory: Development of IRT-based scales. *Intelligence, 36*(2), 161-182. http://doi.org/10.1016/j.intell.2007.02.004

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*(2), 250-270. http://doi.org/10.1037/0033-2909.117.2.250

Vuoksimaa, E., Kaprio, J., Eriksson, C. J. P., & Rose, R. J. (2012). Pubertal testosterone predicts mental rotation performance of young adult males. *Psychoneuroendocrinology*, *37*(11), 1791-1800. http://doi.org/10.1016/j.psyneuen.2012.03.013

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*(3), 426-482. http://doi.org/10.2307/1990256

Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., de Menezes Santos, M., Thomas, C. R., & Miller, B. L. (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science*, *10*(2), 119-125. doi:10.1111/1467-9280.00118

Wainer, H., & Thissen, D. (2001). True score theory: The traditional method. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 23-72). Mahwah, NJ: Lawrence Erlbaum Associates.

Watson, J. M. (1988). Achievement Anxiety Test: Dimensionality and utility. *Journal of Educational Psychology*, *80*(4), 585–591. http://doi.org/10.1037/0022-0663.80.4.585

Weatherholt, T. N., Harris, R. C., Burns, B. M., & Clement, C. (2006). Analysis of attention and analogical reasoning in children of poverty. *Journal of Applied Developmental Psychology*, *27*(2), 125-135. http://doi.org/10.1016/j.appdev.2005.12.010

Weber, K., & Lavric, A. (2008). Syntactic anomaly elicits a lexico-semantic (N400) ERP effect in the second language but not the first. *Psychophysiology*, *45*(6), 920-925. doi:10.1111/j.1469-8986.2008.00691.x

Wertheimer, M. (1900). *Gestalt theory*. Raliegh, NC: Hayes Barton Press.

Wechsler, D. (1991). *WISC-III: Wechsler intelligence scale for children: Manual*. Psychological Corporation.

White, C. S., & Caropreso, E. J. (1989). Training in analogical reasoning processes: Effects on low socioeconomic status preschool children. *The Journal of Educational Research*, *83*(2), 112-118.

White, C. S., & Alexander, P. A. (1986). Effects of training on four-year-olds' ability to solve geometric analogy problems. *Cognition and Instruction*, *3*(3), 261-268. http://doi.org/10.1207/s1532690xci0303_6

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology*, *89*(5), 696-716. http://doi.org/10.1037/0022-3514.89.5.696

Woods, C. M. (2008). IRT-LR-DIF with estimation of the focal-group density as an empirical histogram. *Educational and Psychological Measurement*, *68*(4), 571-586. http://doi.org/10.1177/0013164407310133

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, *73*(3), 532-547.

Xavier, J. M., Bobbin, M., Singer, B., & Budd, E. (2005). A needs assessment of transgendered people of color living in Washington, DC. *International Journal of Transgenderism*, *8*(2-3), 31-47. http://doi.org/10.1300/J485v08n02_04

Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in item response theory scale scores. *Educational and Psychological Measurement*, *72*(2), 264-290. http://doi.org/10.1177/0013164411410056

Zurcher, R. (1998). Issues and trends in culture-fair assessment. *Intervention in School and Clinic*, *34*(2), 103-106. http://doi.org/10.1177/105345129803400206