1-8-2021

Using Differential Item Functioning and Anchoring Vignettes to Examine the Fairness of Achievement Motivation Items

Jacquelyn Bialo
GSU

ACCEPTANCE

This dissertation, USING DIFFERENTIAL ITEM FUNCTIONING AND ANCHORING VIGNETTES TO EXAMINE THE FAIRNESS OF ACHIEVEMENT MOTIVATION ITEMS, by JACQUELYN A. BIALO, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree, Doctor of Education, in the College of Education & Human Development, Georgia State University.

The Dissertation Advisory Committee and the student's Department Chairperson, as representatives of the faculty, certify that this dissertation has met all standards of excellence and scholarship as determined by the faculty.

_____
Hongli Li, Ph.D.
Committee Chair

_____
T. Chris Oshima, Ph.D.
Committee Member

_____
Sarah Carlson, Ph.D.
Committee Member

_____
Randy W. Kamphaus, Ph.D.
Committee Member

_____
Date

_____
Jennifer Esposito, Ph.D.
Chairperson, Department of Educational Policy Studies

_____
Paul A. Alberto, Ph.D.
Dean
College of Education & Human Development

AUTHOR'S STATEMENT

By presenting this dissertation as a partial fulfillment of the requirements for the advanced degree from Georgia State University, I agree that the library of Georgia State University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote, to copy from, or to publish this dissertation may be granted by the professor under whose direction it was written, by the College of Education & Human Development's Director of Graduate Studies, or by me. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without my written permission.

_____

JACQUELYN A. BIALO

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance with the stipulations prescribed by the author in the preceding statement. The author of this dissertation is:

Jacquelyn A. Bialo
Educational Policy Studies
College of Education & Human Development
Georgia State University

The director of this dissertation is:

Dr. Hongli Li
Department of Educational Policy Studies
College of Education & Human Development
Georgia State University
Atlanta, GA 30303

CURRICULUM VITAE

Jacquelyn A. Bialo

ADDRESS: 791 Inman Mews Drive
Atlanta, GA 30307

EDUCATION:

Ph.D.	2020	Georgia State University Educational Policy Studies
M.P.H	2010	Tufts University Global Health and Health Communication
B.A.	2004	Brown University Community Health

PROFESSIONAL EXPERIENCE:

2013-present	Graduate Research Assistant Georgia State University, Atlanta, GA
2016-2019	Psychometrist, DreamBig Center for Learning and Development Alpharetta, GA
2011-2013	Research Study Coordinator University of Massachusetts, Boston, MA

PRESENTATIONS AND PUBLICATIONS:

Li, H., Bialo, J., Xiong, Y., Hunter, C. V., & Guo, X. (2020, April). *The effect of peer assessment on noncognitive outcomes: A meta-analysis*. Paper to be presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA. (canceled due to COVID-19)

Bialo, J. A., Li, H. (2019, April). *Differential response processes among adults from different age groups on PIAAC*. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, Ontario, Canada.

PROFESSIONAL SOCIETIES AND ORGANIZATIONS

2018	American Education Research Association
2018	National Council on Measurement in Education

USING DIFFERENTIAL ITEM FUNCTIONING AND ANCHORING VIGNETTES TO EXAMINE THE FAIRNESS OF ACHIEVEMENT MOTIVATION ITEMS

by

JACQUELYN A. BIALO

Under the Direction of Hongli Li, Ph.D.

ABSTRACT

Achievement motivation is a well-documented predictor of a variety of positive student outcomes. However, researchers have also found threats to fairness and measurement scale comparability in motivation items, including group differences in response scale use and response styles. As such, the measurement comparability of achievement motivation items was evaluated before and after using anchoring vignettes to account for the effect of group-specific response scale use as a source of differential item functioning (DIF) across gender and ethnicity. Within a combined item response theory/ordinal logistic regression DIF framework, gender DIF was assessed using pairwise comparisons and ethnicity DIF was tested using both multiple-group DIF with a common base group as the reference group and all possible pairwise comparisons.

Overall, using the vignettes changed both the form of DIF within items and the pattern of DIF between groups across items. Results indicated the presence of DIF between genders, but the DIF was unrelated to group differences in response scale use. Across ethnic groups, Black/African American students and Asian students demonstrated group-specific response scale

use. When groups showed response tendencies, accounting for such scale use with the vignettes had a greater effect on reducing DIF in base group comparisons than in pairwise comparisons. Despite that DIF was identified in multiple items, the magnitude of all DIF was negligible and had little practical implication. Therefore, achievement motivation items appeared to demonstrate measurement comparability. As sources of DIF often go unidentified, a contribution of this study was the novel use of anchoring vignettes to account for group differences in response scale use as the source of DIF and to clarify the effect of those differences on measurement scale comparability and DIF.

INDEX WORDS: differential item functioning, achievement motivation, anchoring vignettes, response styles, multiple-group DIF

USING DIFFERENTIAL ITEM FUNCTIONING AND ANCHORING VIGNETTES TO EXAMINE THE FAIRNESS OF ACHIEVEMENT MOTIVATION ITEMS

by

JACQUELYN A. BIALO

A Dissertation

Presented in Partial Fulfillment of Requirements for the

Degree of

Doctor of Philosophy

in

Research, Measurement, & Statistics

in

Educational Policy Studies

in

the College of Education & Human Development

Georgia State University

Atlanta, GA
2020

DEDICATION

This dissertation is dedicated to anyone in the world who is afraid of statistics.

ACKNOWLEDGMENTS

The first time I took statistics in college, my sophomore year, no regression model could have ever predicted that someday statistics would be my chosen career path. For that, I have so many people to thank. To Dr. Li, you have given more to me than I can ever express. Your guidance held my course throughout this journey, never letting me go too far astray. You are a brilliant scholar, and I am lucky to have been your student. Thank you, Dr. Oshima, for your kindness and patience and for your base group methodology because it allowed me to do the type of analysis I wanted to for this study. Dr. Carlson, you reminded me to stop and take a break at the moments I needed it most. And finally, Dr. Kamphaus, we have come full circle. Thank you for teaching me how to write an integrated psych report my first semester at Georgia State, because that was what I modeled my dissertation after during my very last semester. Thank you to Jeff Stockwell, Kimberly Moore, and Carla Woods for your tireless administrative support and help with navigating layers of paperwork and policies. Your dedication is amazing, and students in the Educational Policy Studies Department are lucky to have you.

To my mom and dad, thank you not just for your unwavering love and support but also for teaching me to always finish strong. Caralyn, thank you for being my emergency contact throughout graduate school and answering any and every question I ever had. Matthew, thank you for reading my dissertation at 5:00 a.m. You were the perfect person for that task, and I am forever grateful for your editor's eye. And to Darren, Shara, Kelsey, Kate, Sylvester, and Marva, thank you for letting me disappear for a year, and I can't wait to reconnect. Bosco, you have been my four-legged wingman through all of this. Thank you for coming with me to the dog park every morning so I could talk to other people before going home to work all day. I am so sorry for all those nights I kept you awake, but thank you for always keeping my toes and heart warm.

And finally, to John, this dissertation is as much mine as it is yours. It takes a special person to walk step by step with someone as they go through this. You held my hand when I needed grounding, and you breathed with me when I needed calming. You reminded me to do the two most important things: answer the question and eat some chocolate. Thank you for being my rock.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

CRF	Category Response Function
DIF	Differential Item Functioning
GRM	Graded Response Model
ICF	Item Characteristic Function
IRT	Item Response Theory
LR	Logistic Regression
OLR	Ordinal Logistic Regression

1 INTRODUCTION

Motivation is perhaps at the center of any educational endeavor (Covington, 2000; Maehr & Meyer, 1997). As Ronald Reagan's Secretary of Education, Terrel Bell once said, "There are three things to remember about education. The first one is motivation. The second one is motivation. The third one is motivation" (as cited in C. A. Ames, 1990, p. 409). Bell's emphasis on motivation comes with good reason—motivation is related to a variety of positive student outcomes, including higher academic achievement, and graduation and retention rates (Hulleman et al., 2010; Robbins et al., 2004). Students who are motivated to learn are more likely to seek knowledge, use elaborative learning strategies, and persist when faced with difficulties (Senko & Dawson, 2017; Wigfield et al., 2008). These behaviors, in turn, lead to more learning and achievement, which reinforce motivation and encourage continued involvement (Senko & Dawson, 2017; Wigfield et al., 2008). In their meta-analysis, Kriegbaum and colleagues (2018) found that motivation predicted school achievement, above and beyond intelligence.

The specific type of motivation that is relevant to performance and competence in evaluative settings is referred to as achievement motivation[1] (Elliot et al., 2017; Wigfield et al., 2008). While robust evidence has shown the importance of motivation in education, salient differences across groups, such as gender and ethnicity, in motivational constructs have also emerged (Wigfield et al., 2008). For example, middle and high school students report different types of motivation and achievement goals by gender (Dupeyrat et al., 2011; Vantieghem & Van Houtte, 2018). Perceptions of academic self-concept appear to vary by ethnic groups for both undergraduate students and adolescents (Edman & Brazil, 2009; Hong et al., 2020; Shernoff & Schmidt, 2008). However, researchers have also found threats to measurement scale comparability in motivation

[1] Motivation and achievement motivation will be used interchangeably unless otherwise noted.

items, including group differences in response scale use and response styles (e.g., Gnambs & Hanfstingl, 2014; He & Van de Vijver, 2016). Therefore, before drawing any conclusions about group differences in achievement motivation, instruments should be checked for comparable measurement scale functioning across groups (Campbell et al., 2008; Dever & Kim, 2016; van der Sluis et al., 2010).

Because measurement scale comparability is fundamentally a fairness issue, it can be evaluated using differential item functioning (DIF) methods (Sireci & Rios, 2013). When using DIF to examine measurement scales, testing a hypothesized source of DIF can provide greater clarity into DIF results (Finch et al., 2016; Sandilands et al., 2013). As such, the purpose of this study was to identify if achievement motivation items demonstrated measurement comparability across gender and ethnicity before and after accounting for the effect of group differences in response scale use as the source of DIF. First, DIF was detected within a hybrid framework that combined elements from item response theory and ordinal logistic regression DIF (Millsap, 2006) using pairwise and multiple-group comparisons. Next, group-specific scale use was tested as the source of DIF by applying anchoring vignette methodology. Then, unadjusted and vignette-adjusted DIF results were compared to assess the effect of group differences in response scale use on DIF.

Background

Achievement Motivation

Despite nearly 100 years of research, motivation theorists continue to disagree over the exact nature of motivation and motivational processes (Schunk et al., 2014). Motivation is not an outcome, but rather is an unobservable process that is inferred from observable actions (e.g., task

choice, effort; Schunk et al., 2014). Achievement motivation is distinguished from general motivation by competence forming the conceptual cornerstone for achievement (Elliot & Dweck, 2005) such that the "achievement" in achievement motivation is attaining competence (Elliot & Dweck, 2005). From this perspective, achievement motivation can be defined as "the energization and direction of competence-based affect, cognition, and behavior" (Elliot, 1999, p. 169). Theories of motivation address what drives students and their reasons for engaging in a particular task (Eccles & Wigfield, 2002; Kaplan & Maehr, 2007; Liem & Elliot, 2018).

Achievement motivation is highly contextualized and therefore is best viewed through a systems framework that accounts for individual-level traits as well as group-level and environmental characteristics, such as the Socio-Cultural Hierarchical Model of Achievement Motivation (Kitayama, 2002; Liem et al., 2012; Maehr & Meyer, 1997; Nolen et al., 2015). At the individual level of the model, motivational dispositions (i.e., approach/avoidance), competence expectancies (i.e., expectations for success or failure), and achievement goals (i.e., mastery/performance) explain how students regulate their competence-relevant pursuits (Liem et al., 2012; Liem & Elliot, 2018; Michou et al., 2014). In the model, achievement goals are the cognitive representations of what students hope to attain and they direct students' behaviors (Liem & Elliot, 2018). Because higher-order motivational dispositions are channeled through lower-order achievement goals, achievement motivation can be measured with achievement goals.

At the group level, the socio-cultural model holds that students' motivational patterns and goals are socialized by their environment and cultural milieu (Elliot & Church, 1997; Liem & Elliot, 2018; Wigfield et al., 2008). Motivational experiences are also mediated by different factors, such as racial identity or cultural orientation (e.g., collectivist vs. individualist; Hill & Torres, 2010; Liem et al., 2012; Miller-Cotto & Byrnes, 2016). As such, researchers have found group

differences in achievement motivation and related constructs (Dupeyrat et al., 2011; Hong et al., 2020; Liem et al., 2012; Murdock, 2009). For example, female adolescents have reported more mastery goals than their male peers, who endorse more performance goals (Hong et al., 2020; Hyde & Durik, 2005; Senko & Hulleman, 2013). Similarly, Asian American undergraduates have shown more avoidance goals than their White counterparts (Zusho et al., 2005), with students from collectivist cultures in general being more driven by social goals than students from individualist cultures (King et al., 2017). Black/African American students have demonstrated both higher intrinsic motivation than White students and higher school engagement than Hispanic/Latinx students (Johnson et al., 2001; Lee et al., 2016; Shernoff & Schmidt, 2008).

Measurement Scale Comparability and Sources of Differential Item Functioning

On the one hand, these group differences in motivation could reflect "true" group differences (Elder, 1997; Gnambs & Hanfstingl, 2014; Sireci & Rios, 2013). From a socio-cultural perspective, such differences are expected by virtue of the different ways in which motivation is socialized (Liem et al., 2012; Liem & Elliot, 2018). On the other hand, researchers have found the presence of group-specific response scale use and response styles in motivation measures. That is, individuals and groups show particular response scale tendencies and reporting patterns when presented with Likert-style categorical response options on self-report measures (Böckenholt & Meiser, 2017; Bolt et al., 2014; Kemmelmeier, 2016). For example, boys have been found to endorse the top of the response scale, while girls trend toward options in the middle of the scale (Butler & Hasenfratz, 2017; Harzing, 2006). In their meta-analysis, Batchelor and Miao (2016) showed that Black/African American and Hispanic/Latinx Americans use the extreme upper and lower ends of scales more than European-descended North Americans (e.g., Hamamura et al., 2009; Hui & Triandis, 1989; McDaniel et al., 2011). In contrast, individuals

from East Asian cultures tend to respond modestly in the middle of the response scale (Grimm & Church, 1999; Hamamura et al., 2009; Min et al., 2016).

In motivation measurement, however, there is an overall dearth of literature on how groups use response scales. One reason for this may be that given the relationship between self-presentation and achievement motivation, response scale use in motivation has typically been studied at the individual level as a function of socially desirable responding (Elliot et al., 2018; Tan & Hall, 2005). Nonetheless, existing evidence indicates that groups show particular reporting patterns in motivation and related attitudinal constructs. For example, He and Van de Vijver (2016a) interpreted a country-level motivation-achievement paradox to partially reflect cultural differences in response scale use among 15-year-old students. On a self-esteem measure, White American college students used more extreme responding than their peers in China (Song et al., 2011). Gender has also been tied to scale use on achievement striving and self-regulation items (Gnambs & Hanfstingl, 2014; Wetzel, Böhnke, et al., 2013). Furthermore, although female students get higher grades than males at school (Voyer & Voyer, 2014), Dupeyrat et al. (2011) found that adolescent males were more likely than females to attribute personal success to their ability and competence. This difference between self-perceptions and observed performance can be taken as evidence that more than just the construct of interest is being captured in self-report motivation items.

Response styles are problematic in measurement and a threat to cross-group comparisons because observed scores reflect not only the construct of interest, but also variance due to response scale use (Bolt et al., 2014). For example, extreme response scale use will naturally lead to inflated observed scores at the top of the response scale and underestimates of the construct at

the bottom of the scale (Bolt et al., 2014; Steinberg & Thissen, 2006). In such situations, the observed score is contaminated by response-style variance (i.e., the "difference" between the true score and the response category selected), and response scale use becomes a source of systematic measurement error (Baumgartner & Steenkamp, 2001; Bolt et al., 2014). At the group level, the cumulative effect of response-style variance is problematic because it can not only lead to groups being on different measurement scales and scores taking on different meanings for different groups, but it can also induce DIF (Böckenholt & Meiser, 2017; Gnambs & Hanfstingl, 2014; Wetzel, Böhnke, et al., 2013).

DIF is present in an item if the probability of an item response varies by group after respondents are matched by construct level (Clauser & Mazor, 2005). Because DIF controls for the relationship between an item response and construct level by matching participants, DIF reveals if groups are on different measurement scales (Millsap, 2006). Research on response scale use as a source of DIF is sparse and has yielded mixed findings. For example, neither Wetzel and colleagues (2013) nor Gnambs and Hanfstingl (2014) found that controlling for response styles had an appreciable impact on DIF across German adults and adolescents. However, in both studies, samples were fairly homogenous and response styles were derived through latent class analysis, with the classes being dichotomized to reflect either extreme or non-extreme reporting behaviors. In contrast, modeling response styles as an explicit statistical dimension have yielded more accurate DIF detection (Bolt & Johnson, 2009; Chen et al., 2017; Jin & Chen, 2019).

Anchoring Vignettes

King and colleagues (2004) developed a combined survey design/statistical methodology that improves cross-group comparisons by controlling for individual response scale use called

anchoring vignettes. In anchoring vignette methodology, individuals use the same categorical response scale (e.g., agree/disagree) to respond to two sets of survey items. The first set of items is a self-report inventory measuring the construct of interest; the second set of items asks the person to rate a series of hypothetical situations (i.e., vignettes) depicting the same construct. Individual responses to the self-report items are then rescored relative to that person's ratings of the vignettes. The measurement theory behind the vignettes is that because the vignettes are written to reflect different absolute levels of the construct (e.g., low, medium, high), by rescoring a person's self-report items based on their vignette ratings, their response-style variance can be "subtracted" off (He, Buchholz, et al., 2017; King et al., 2004). In turn, by accounting for scale use, vignette-adjusted scores should have less measurement distortion and yield a "purer" measure of the construct, with more equivalent measurement scales and less DIF (Dever & Kim, 2016; Primi et al., 2018; Weiss & Roberts, 2018).

Researchers have previously drawn on anchoring vignettes to address the effects of response scale tendencies on measurement properties in motivation items and other related constructs. For example, Marksteiner et al. (2019) found that applying the vignettes to correct for culture-specific scale use slightly increased internal consistency and factor loadings. Weiss and Roberts (2018) similarly saw improved model fit, higher reliability, and higher factor loadings after adjusting a personality measure with the vignettes. He and Van de Vijver (2016) used the vignettes to confirm the presence of country-specific response styles in motivation items, but they did not use the vignettes to correct for group differences in response scale use as a source of DIF in a DIF analysis.

Methodological Framework

Fairness in Measurement

Measurement scale comparability is central to making fair comparisons across groups (Morren et al., 2011; Wu & Ercikan, 2006). In the *Standards for Educational and Psychological Testing* (hereafter referred to as "the *Standards*"), a fair test is one in which scores have the same meaning for all test-takers. Fairness is operationalized in the *Standards* as fairness in treatment during the testing process, fairness in access to the constructs measured, fairness in the validity of individual test score interpretations for the intended uses, and fairness as the absence of measurement bias (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). Fairness as the absence of measurement bias refers to the comparability of measurement scales (AERA et al., 2014). More specifically, measurement scale comparability implies the presence of measurement equivalence (Gnambs & Hanfstingl, 2014). Conceptually, measurement equivalence denotes that constructs have the same psychological meaning across groups (Rios & Wells, 2014; Svetina & Rutkowski, 2017). Statistically, measurement equivalence indicates that items and tests have the same statistical properties and relationships across groups (Camilli, 2013; Svetina & Rutkowski, 2017). When an instrument's measurement properties are consistent across test-takers, it is psychometrically fair to compare groups (Boer et al., 2018; Poortinga, 1989).

Because scale comparability is a fairness issue, it can be evaluated using DIF (AERA et al., 2014; Zumbo, 2007). In a DIF analysis, measurement scales are comparable (i.e., an absence of measurement bias) when respondents with the same level of the construct have the same probability of endorsing an item response, regardless of group membership (Clauser & Mazor, 2005; Zumbo, 2007). In a typical DIF study, two groups are selected for comparison, with one group

designated as the reference group, usually the majority, and the other designated as the focal group, usually the minority group of interest (Clauser & Mazor, 2005; Zumbo, 2007). The key in DIF is that if respondents are matched by construct level, but statistical parameters that should otherwise be equal vary based on group membership, then that would indicate that something in the measurement instrument was not functioning equally across the reference and focal groups and that groups were not on the same measurement scale (Reise et al., 1993; Walker, 2011). However, merely identifying the presence of DIF does little to inform its causes (Finch et al., 2016). Moreover, given that the source of DIF can affect how the practical implications of DIF are interpreted, it is also equally important to investigate a source of DIF (Bolt & Johnson, 2009; Finch et al., 2016; Sandilands et al., 2013).

Problem Statement

Motivation is fundamental to understanding academic achievement (Lee et al., 2016). Because achievement motivation predicts general educational outcomes (e.g., GPA) and more specific learning behaviors (e.g., deep learning strategies; Day et al., 2003) as well as group differences in achievement (Robbins et al., 2004; Steinmayr & Spinath, 2008), motivation is an important area for educational research (Linnenbrink-Garcia et al., 2016). When conclusions regarding group differences in motivation are used to target educational practices or to inform educational policies, it is important that the measures and items on which those conclusions are based capture substantive and meaningful group differences (Linnenbrink-Garcia et al., 2016). This is fundamentally a fairness issue as without evidence that an instrument's measurement properties are consistent across test-takers, the extent to which observed group differences reflect true group differences or differences in measurement models or measurement scales is unclear (Camilli, 2013; Campbell et al., 2008; Dever & Kim, 2016; Poortinga, 1989). If groups were on

different measurement scales, then scores would no longer have the same psychological meaning, and the fairness of cross-group comparisons and conclusions about group differences would be questionable (Kane, 2013; Messick, 1995; Osterlind & Everson, 2010).

Although fairness is a central concern in assessment (Sireci & Rios, 2013), motivation items have primarily been assessed for structural, metric, and scalar invariance (e.g., Campbell et al., 2008), with measurement scale comparability having largely gone unexamined. The absence of DIF testing in motivation is not necessarily atypical as the use of DIF in psychological measurement is rarer than its use in educational and cognitive assessment (Johanson, 1997; Wetzel, Böhnke, et al., 2013). Nonetheless, given that observed group differences in achievement motivation constructs and related outcomes have practical implications (e.g., instruction design, resource allotment; Linnenbrink-Garcia et al., 2016), researchers need to provide evidence that achievement motivation instruments are of high quality and are psychometrically fair to all students. Moreover, because groups have demonstrated response scale tendencies in motivation items, and those response tendencies can impact scale comparability and induce DIF, the effect of such differences as a source of DIF should be investigated. To that end, the purpose of this dissertation was to identify if achievement motivation items demonstrated measurement scale comparability across gender and ethnicity before and after using anchoring vignettes to account for the effect of group-level differences in response scale use as a source of DIF.

Research Questions

Two research questions will be addressed in this study:

1. Do achievement motivation items show DIF by gender or ethnicity?
2. Do achievement motivation items show DIF by gender or ethnicity after using anchoring vignettes to account for the impact of group-level differences in response

scale use? Is there a difference in the magnitude and direction of DIF between the un-adjusted and vignette-adjusted item scores?

Hypothesis 1_A: Because vignettes were designed to account for response styles, the magnitude of DIF (i.e., effect sizes) will decrease after applying the nonparametric vignette scoring.

Overview of the Study

This study used data from the 2015 iteration of the Programme for International Student Assessment (PISA), an international large-scale assessment administered triennially to 15-year-old students in over 70 countries (OECD, 2017). In each iteration, students are tested on their reading, science, and math literacy, with one of those areas being assessed more in-depth; in PISA 2015, science was the main focus (OECD, 2017). In addition to cognitive testing, the PISA 2015 background questionnaire included a variety of attitudinal items as well as three anchoring vignettes related to achievement motivation (OECD, 2017). PISA's two-stage stratified sampling method is designed to yield nationally representative samples of 15-year-old students in school (OECD, 2017; Rutkowski & Rutkowski, 2016). In the US, PISA 2015 was implemented at 177 schools, with a total sample size of 5712 students.

To answer Research Question 1, DIF was detected within a hybrid framework that combined elements from latent trait (i.e., item response theory) and observed-score (i.e., logistic regression) DIF detection methods (Millsap, 2006). Logistic regression DIF has several advantages, including that it can accommodate both uniform and nonuniform DIF as well as polytomous items (French & Miller, 1996; Miller & Spray, 1993; Swaminathan & Rogers, 1990; Zumbo, 1999). In a traditional logistic regression DIF, participants are matched using the total observed test score (Zumbo, 1999). However, the hybrid DIF method used in this study involved

matching respondents with a graded response model estimate of achievement motivation, and then detecting DIF with ordinal logistic regression (Choi et al., 2011). The magnitude of DIF was quantified using both McFadden's pseudo R^2 and regression coefficients (Zumbo, 1999).

The groups being compared for DIF were based on gender and ethnicity. Gender DIF was evaluated using pairwise comparisons; ethnicity DIF was investigated using both multiple-group DIF with a common base group as the reference group (Oshima et al., 2015) and all possible pairwise comparisons. The base group was comprised of an average subsample of the total multiple-group sample (Oshima et al., 2015). In the multiple-group DIF, each ethnic focal group was compared to the base group (Oshima et al., 2015); in the pairwise comparisons, each ethnic group was compared to the other. Using a common base group as the reference group is advantageous with multiple-level groups, such as ethnicity, as it avoids making a values statement based on the reference group (Martinková et al., 2017; Sari & Huggins, 2015). However, because groups may be on different measurement scales such that they do not evidence DIF in relation to a base group but do evidence DIF when compared to one another, all pairwise comparisons were also evaluated for DIF (Ellis & Kimmel, 1992; Sari & Huggins, 2015).

To answer Research Question 2, self-reported motivation items were first rescored using the nonparametric scoring approach to anchoring vignettes. The nonparametric approach utilizes respondent-dependent scoring in which individual scores are "rescaled" relative to that person's vignette ratings (von Davier et al., 2018). Given the presence of group-specific response scale use in self-reports of motivation (e.g., He & Van de Vijver, 2016a; Wetzel, Böhnke, et al., 2013) and that the nonparametric scoring of anchoring vignettes was specifically developed to correct for response styles (King et al., 2004), it was expected that applying the nonparametric adjustment to self-report items would mitigate the impact of response styles on DIF (King et al., 2004).

After adjusting self-report items with the vignettes, the same set of DIF analyses from Research Question 1 was re-run using the vignette-adjusted item responses. Finally, the effect of response scale use on DIF was evaluated by examining changes to DIF outcomes as well as changes in the size and direction of DIF between unadjusted and vignette-adjusted items.

Significance of the Study

Findings from this study can contribute independently to as well as at the intersection of the research literature on motivation measurement and the fairness of motivation items, multiple-group DIF methods and anchoring vignettes, and DIF sources. First, motivation items and PISA noncognitive items (or a combination thereof) have rarely undergone DIF (Hopfenbeck et al., 2018), and using anchoring vignettes to account for response scale use in a DIF analysis appears to be even rarer. To that end, findings from this study add to the limited literature on measurement scale comparability and DIF in motivation items. Results can further inform fairness and DIF in PISA items and the effect of anchoring vignettes on DIF. Second, despite the utility of multiple-group DIF (with or without a base group), researchers have not readily implemented this method (Oshima et al., 2015). This study illustrates how multiple-group methods function in an applied analysis with a short test.

Finally, a major challenge in DIF research is that DIF methods do not readily lend themselves to identifying sources of DIF, with the root cause often remaining unknown (Gierl et al., 2003; Hopfenbeck et al., 2018; Sandilands et al., 2013). Studies regarding systematic sources of DIF, and in particular, the impact of group-specific response scale use on DIF, are infrequent (e.g., Bolt & Johnson, 2009; Gnambs & Hanfstingl, 2014; Wetzel, Carstensen, et al., 2013). Moreover, despite the ubiquity of differential scale use as a general threat to fairness in cross-group comparisons (AERA et al., 2014; Bolt & Johnson, 2009; Chen et al., 2019; Ziegler, 2015),

Bolt and Johnson (2009) note that, “Relatively little is often done in practice to investigate their implications or to control for their effects” (p. 350). As such, results from this study can be used to further clarify the effect of group differences in response scale use as a source of DIF as well as to provide evidence of scale use in motivation items and any subsequent measurement effects on scale comparability or DIF.

2 REVIEW OF THE LITERATURE

The purpose of this study was to identify if achievement motivation items demonstrated measurement scale comparability across gender and ethnicity before and after using anchoring vignettes to account for the effect of group-level differences in response scale use as a source of differential item functioning (DIF). To that end, the first research question guiding this study was: do achievement motivation items show DIF by gender or ethnicity? The second question guiding this study was: do achievement motivation items show DIF by gender or ethnicity after using anchoring vignettes to account for the impact of group-level differences in response scale use? Is there a difference in the magnitude and direction of DIF between the unadjusted and vignette-adjusted item scores?

In this chapter, the background for the analysis is first provided by elaborating on the conceptual model for achievement motivation and the effects of response scale use on measurement. Next, anchoring vignettes and how they can account for response styles as a source of DIF are described. Finally, the methodological conceptual framework for the study is outlined, and ordinal logistic regression DIF is reviewed.

Background and Theoretical Framework

Achievement Motivation

Motivation is derived from the Latin verb *movere* (to move; Eccles & Wigfield, 2002; Fulmer & Frijters, 2009). The idea of motivation as the study of movement or action denotes that motivation is a process, not an outcome (Eccles & Wigfield, 2002; Schunk et al., 2014). Achievement motivation is differentiated from general motivation by grounding achievement in competence such that achievement motivation can be defined as "the energization and direction of competence-based affect, cognition, and behavior" (Elliot, 1999, p. 169). Although motivation

was originally theorized to reside within an individual's needs and drives systems, contemporary theories of achievement motivation are primarily social cognitive in nature (Cury et al., 2006; Elliot, 2006). A socio-cultural systems perspective contends that motivation is highly situated and that what students want to achieve (i.e., goals) and the reasons for why they want to achieve it are best understood as a product of individual characteristics as well as environmental surroundings (Kitayama, 2002; Liem et al., 2012; Liem & Elliot, 2018; Nolen et al., 2015).

In the socio-cultural hierarchical model of achievement motivation seen in Figure 1, higher-order motivational dispositions (e.g., competence-based motives, individual and social-relational motives), competence expectancies (i.e., expectations for success or failure), and achievement goals (i.e., mastery/performance) explain how students regulate their choice, commitment, and effort in the pursuit of competence (Liem & Elliot, 2018; Michou et al., 2014; Wigfield et al., 2008). Motivational dispositions function as broad cognitive schema that energize affective-based approach and avoidance behaviors. Competence describes a state of effectiveness that can be operationalized relative to a given set of standards and competence expectancies orient an individual's expectations of competence either towards success or failure (Elliot, 2006; Elliot & Church, 1997; Elliot & Dweck, 2005; Fryer & Elliot, 2007). Researchers have found, for example, that students focused on positive outcomes have higher academic achievement than students who focus on avoiding negative consequences (Huang, 2012).

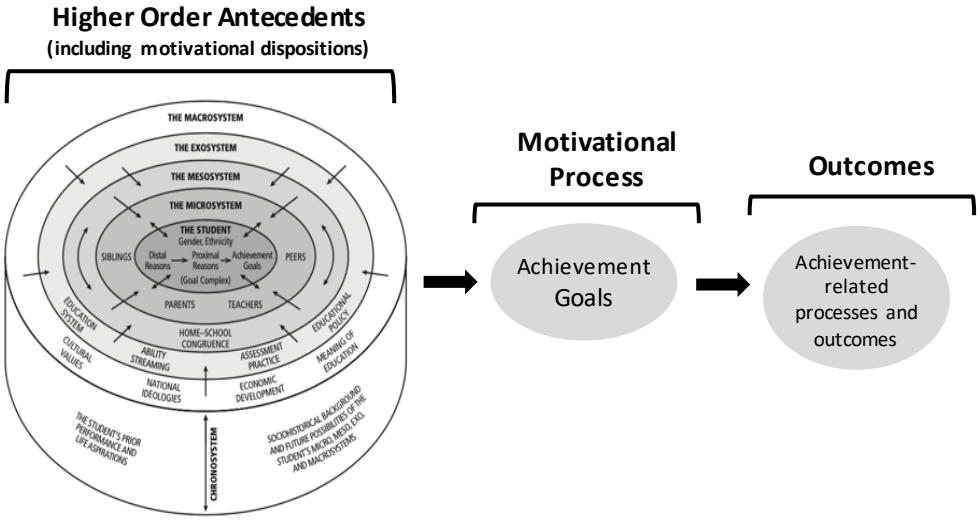
In the socio-cultural model, motivational dispositions are channeled into achievement goals, which are the cognitive representations of what students hope to attain, and the function of those goals is to direct and regulate students' competence-relevant strivings in academic settings (Elliot, 2006; Elliot & Thrash, 2001). The two major types of achievement goals—mastery and performance—differ with respect to their standards for competence (Elliot & Thrash, 2001; Liem

& Elliot, 2018). Mastery goals are grounded in self-improvement as individuals with those goals strive to either develop their own ability (i.e., mastery-approach) or avoid their own incompetence (i.e., mastery-avoidance; Baranik et al., 2010; Liem & Elliot, 2018). In contrast, performance goals are grounded in an ethos of self-presentation and a desire to either demonstrate ability (i.e., performance-approach) or to avoid the appearance of incompetence (i.e., performance-avoidance; Liem & Elliot, 2018). Taken together, individuals with either mastery- or performance-approach goals focus on success, and they adopt behaviors that facilitate positive outcomes; conversely, those with avoidance goals adopt behaviors that avoid failure and eschew negative outcomes (Conroy, 2017; Covington, 2000; Elliot & Church, 1997).

In academic settings, the relative value of mastery versus performance goals continues to be highly debated among achievement goal researchers (Hulleman et al., 2010; Senko et al., 2011; Senko & Dawson, 2017; Wormington & Linnenbrink-Garcia, 2017). Although mastery goals facilitate a variety of beneficial outcomes, including deep learning strategies (Day et al., 2003; Hulleman et al., 2010; Senko & Dawson, 2017), it is performance goals and *not* mastery goals, that actually predict academic achievement (Huang, 2012; Hulleman et al., 2010; Senko & Dawson, 2017; Senko & Hulleman, 2013). Nonetheless, because achievement goals are the channels through which higher-order motivational dispositions are manifested, achievement motivation can be measured as a function of achievement goals (Elliot & Thrash, 2001; Liem & Elliot, 2018).

Figure 1

Socio-Cultural Hierarchical Model of Achievement Motivation



Note. Adapted from "Sociocultural influences on achievement goal adoption and regulation: A goal complex perspective," by G. A. D. Liem and A. J. Elliot, 2018, in G. A. D. Liem and D. M. McInerney, *Big theories revisited 2*, p. 50.

Group Differences in Motivation

The socio-cultural hierarchical model of achievement motivation utilizes ecological systems to organize the various environmental structures, life experiences (e.g., socio-historical marginalization and oppression), and actors that socialize a student's identity, values, beliefs, and goals (Liem & Elliot, 2018; Wood & Graham, 2010). As such, researchers have found that students' motivational patterns and goals vary across gender and ethnic groups (Butler & Hasenfratz, 2017; Dupeyrat et al., 2011; Hong et al., 2020; Liem et al., 2012). Before discussing the relationship between motivation and those groups, however, it is prudent to define gender and ethnicity. *Gender* is used to describe cisgender individuals, or those who identify with the same sex and gender categories that were assigned to them at birth (Westbrook & Saperstein, 2015). *Ethnicity* is understood as a socially constructed group label ascribed to individuals based on commonalities in history, nation or region of origin, customs, and ways of being (Markus, 2008).

As there is significant overlap between ethnicity and culture, with ethnicity often being defined in relation to culture because culture describes values, beliefs, customs, and experiences shared by a group of people (Urdan & Bruchmann, 2018), the two terms are often used interchangeably. From a measurement perspective, though, it is important to note that dichotomizing gender or grouping individuals by ethnicity labels may mask important and meaningful within-group differences (Borsboom, 2006; Elder, 1997; Urdan & Bruchmann, 2018).

Regarding gender differences in motivation, for example, girls report more mastery goals and boys endorse more performance goals (Conley, 2012; Meece et al., 2006; Steinmayr & Spinath, 2008; Wilson et al., 2016). Female students can be more autonomously and extrinsically motivated than boys (Vantieghem & Van Houtte, 2018), but boys hold higher self-efficacy and self-concept beliefs than girls (Huang, 2013; Min et al., 2016). Even though boys earn lower grades than girls in school (Voyer & Voyer, 2014), boys report higher self-perceptions of competence than girls (Butler & Hasenfratz, 2017; D'Lima et al., 2014; Dupeyrat et al., 2011).

Across ethnic groups, Black/African American students have shown higher intrinsic motivation and higher academic self-efficacy as well as more differentiated academic self-concept than White students (Cokley et al., 2003; Edman & Brazil, 2009; Shernoff & Schmidt, 2008). Asian American students demonstrated lower self-esteem, more performance-avoidance goals, and higher fear of failure than their White American peers (Shernoff & Schmidt, 2008; Zusho et al., 2005). Johnson et al. (2001) showed that Hispanic/Latinx students were more engaged at school than White students but less engaged than Black/African American students. In addition to group differences in motivation, consistent with predictions from socio-cultural theories, motivation is mediated by different socialized and lived experiences. The historical oppression, stig-

matization, and disenfranchisement of Black/African Americans has made racial identity, cultural mistrust, and stereotype threat experience salient to their motivation (Caldwell & Obasi, 2010; Gray et al., 2018; Smith, 2004; Thoman et al., 2013). Social relationships and cultural values of communalism and interdependence mediate motivation for Hispanic/Latinx students (Hill & Torres, 2010). Students from Eastern societies are more motivated by socially-oriented achievement and social goals than students from individually achievement-oriented Western societies (Liem et al., 2012).

Measurement Scale Comparability and Sources of Differential Item functioning

Response Styles and Group Differences in Scale Use

On the one hand, observed group differences may reflect true group differences in motivation and motivational experiences. Based on the socio-cultural model, such group differences are predicted based on the different ways in which motivation is socialized (Liem et al., 2012; Liem & Elliot, 2018). On the other hand, researchers have found evidence that groups may differ in their use of categorical response scales when on self-report measures of achievement motivation (Dever & Kim, 2016; Gnambs & Hanfstingl, 2014; Wetzel, Böhnke, et al., 2013). More specifically, in psychological measurement, one well-known threat to measurement comparability and cross-group comparisons on assessments that have Likert-type scales are response styles (Bolt et al., 2014; Wetzel et al., 2016). That is, when presented with categorical response scales on self-report measures, individuals and groups demonstrate particular response scale preferences and patterns in their reporting behaviors (Baumgartner & Steenkamp, 2001; Bolt et al., 2014; Wetzel et al., 2016). These differences can be triggered by personality traits, idiosyncratic interpretations of item content, differential interpretations of response scale terms (e.g., *a lot* vs.

a little), and varying frames of reference (Bolt et al., 2014; He, Buchholz, et al., 2017; Morren et al., 2011; Primi et al., 2018).

Tourangeau (2018) decomposed the cognitive process of responding to survey items into roughly four steps. First, a person must comprehend and interpret the meaning of an item. Next, long-term memory and working memory are searched for relevant information. Then, information must be integrated into a summary judgment or estimate and, finally, that judgment or estimate has to be translated (i.e., mapped) onto one of the given response category options (Tourangeau, 2018). Krosnick (1991) noted that the survey response process is cognitively demanding, so he theorized that response styles reflect a satisficing strategy that reduces the cognitive burden of responding, particularly when the item demands exceeds a person's ability or motivation to respond to the item.

Some response styles are independent of item content (Baumgartner & Steenkamp, 2001; Kemmelmeier, 2016). For example, individuals with an extreme response style use the upper and lower ends of the response scale, ostensibly to reduce cognitive load because scale endpoints are more precise (Bolt et al., 2014; Johnson et al., 2005). A meta-analysis of extreme response style showed that Black/African Americans used extreme responding most frequently, followed by Hispanic/Latinx Americans and White Americans, with Asian Americans displaying the least extreme responding (Batchelor & Miao, 2016). Individuals with an acquiescent or disacquiescent response style overuse agree and disagree categories (Baumgartner & Steenkamp, 2001; Bolt et al., 2014). One theory of acquiescent responding is that most survey questions are reasonable enough, and it is more mental effort to think of reasons to disagree with a statement than to just agree (Krosnick, 1991).

Other individuals prefer more modest responses, opting for a middling response style (Baumgartner & Steenkamp, 2001; Kemmelmeier, 2016). Both acquiescent and middling response styles have been tied to cultural orientation (e.g., collectivist vs. individualistic; Johnson et al., 2005). For example, students from East Asian cultures tend to respond in the middle of the scale and to avoid the positive end (Hamamura et al., 2009; Min et al., 2016), possibly due to their dialectic thinking and collectivist emphasis (Hamamura et al., 2009; Johnson et al., 2005). In contrast to extreme or acquiescent scale use, socially desirable responding is explicitly related to item content (Kemmelmeier, 2016). Individuals specifically select responses that portray them in a more favorable or positive way than they actually are (Kemmelmeier, 2016; Lalwani et al., 2009; Paulhus, 1991). Hispanic/Latinx individuals have been found to score higher on social desirability measures than White individuals, which researchers attributed to collectivist versus individualist cultural mechanisms (Hopwood et al., 2009).

Response Scale Use in Motivation Measures

Research on response styles in motivation items is fairly limited, particularly at the group level. At the individual level, response styles have typically been studied as a function of socially desirable responding because achievement goals are argued to reflect substantive personality traits related to self-presentation and approval (Day et al., 2003; Elliot et al., 2018; Tan & Hall, 2005). Although some researchers have found that controlling for social desirability had little effect on the achievement goals students selected (Elliot et al., 2011, 2016), socially desirable responding has yielded inflated factor loadings and biased estimates between goal orientation constructs (Day et al., 2003; Tan & Hall, 2005). In achievement striving, Wetzel and Carstensen (2017) found the presence of other types of response styles, including a correlation of 0.824 with acquiescent response style and a correlation of -0.707 with disacquiescent response style.

Though there is a dearth of literature on group-specific response scale use in achievement motivation, existing evidence does show that groups have specific reporting patterns on measures of motivation and other related constructs. For example, He and Van de Vijver (2016a) attributed the negative correlation between students' self-reported motivation at the country level and their academic achievement to country and culture-specific response styles. Likewise, although Black/African American and Hispanic/Latinx students do not perform academically as well as their White counterparts, the reasons for which are highly complex and beyond the scope of this review, those students nonetheless rate themselves as having higher academic self-efficacy than their White peers (e.g., Edman & Brazil, 2009). This "motivation-achievement paradox" has been taken as evidence of group-specific scale use (e.g., He & Van de Vijver, 2016a). Boys are positively biased in their estimation of math abilities and self-concept, while girls modestly underestimate their abilities in relation to how they actually perform (Butler & Hasenfratz, 2017; Dupeyrat et al., 2011; Min et al., 2016). Cognitive validity studies also reveal that achievement goal items are not always interpreted as intended (Warnecke et al., 1997). Taken together, these findings show that groups demonstrate different response behaviors in motivation items.

Measurement Effects of Group Differences in Response Scale Use

Response styles are problematic in measurement because they can distort distributions and estimates (Bolt et al., 2014; Kemmelmeier, 2016). For example, extreme response style or acquiescent response style yields scores that may be inflated compared to the respondent's true level of the latent construct (Bolt et al., 2014). If observed scores are interpreted as reflecting the "amount" of the latent construct an individual has, then construct estimates will by extension be confounded by the presence of response styles and response styles become a source of systematic measurement error (Bolt et al., 2014; Steinberg & Thissen, 2006). This is problematic because

distorted measurements decrease score precision and can lead to inaccurate inferences (Osterlind & Everson, 2010). Moreover, variance due to response scale use can lead to groups being on different measurement scales, which threatens the comparability of score meaning and cross-group comparisons, and can induce DIF (Baumgartner & Steenkamp, 2001; Böckenholt & Meiser, 2017; Gnambs & Hanfstingl, 2014; Grimm & Church, 1999).

Evidence of the impact of group-level response scale use as a source of DIF is limited, and the research that has been conducted has yielded conflicting results. For example, Wetzel, Carstensen, et al. (2013) used latent class analyses to identify different types of reporting behaviors across men and women. Although controlling for scale use changed DIF classifications to varying degrees, with some items even only showing DIF after controlling for response styles, there was little overall practical impact of changes to DIF (Wetzel, Carstensen, et al., 2013). Gnambs and Hanfstingl (2014) followed the same approach used by Wetzel and colleagues and similarly found that response scale use had a negligible impact on DIF. However, in both studies, response styles were derived through latent class analyses and classes were dichotomized to reflect either extreme or non-extreme reporting behaviors. In contrast, simulation studies that controlled for response scale with methods such as logistic regression and multiple-indicators multiple causes models, more accurately detected DIF (Bolt & Johnson, 2009; Chen et al., 2017; Jin & Chen, 2019).

Researchers have used a variety of methods to account for the measurement effects of response styles on self-report assessments, including explicit response style measures (e.g., Balanced Inventory of Desirable Responding), alternative response scale formats, and statistical techniques (He, Van de Vijver, et al., 2017; Hopwood et al., 2009; Kyllonen & Bertling, 2013). He and colleagues derived a general response style factor, which they used to control for scale

use in regression analyses (He, Van de Vijver, et al., 2017; He & Van de Vijver, 2016b). Response styles have also been modeled as an explicit statistical dimension (e.g., Bolt et al., 2014; Ferrando, 2014; Jonas & Markon, 2019) and estimated with latent class factor analysis and mixed Rasch models (e.g., Chen et al., 2017; Morren et al., 2011; Wetzel, Carstensen, et al., 2013). These modeling approaches actually reflect a particular view of how response styles function. When viewed categorically, individual response styles reflect global, qualitative differences that an individual either does or does not apply to all items (Austin et al., 2006; Wetzel, Böhnke, et al., 2013). The continuous perspective allows respondents to display different degrees of a particular response style (Bolt et al., 2014; Bolt & Johnson, 2009), which is the perspective espoused in this study.

Anchoring vignettes are a combined survey design/statistical tool that can be used to statistically control for the effects of response styles on cross-group comparisons (Dever & Kim, 2016; King et al., 2004). In the next section, the scoring approaches and measurement assumptions for anchoring vignettes are described as well as how the vignettes can correct for response scale use in DIF.

Anchoring Vignettes

Anchoring vignettes were initially developed to improve survey comparability in political science and political self-efficacy research (King et al., 2004). They have most often been used on measures of political attitudes, quality of life, and self-rated health (Au & Lorgelly, 2014; Grol-Prokopczyk et al., 2015; Knott et al., 2017; Mojtabai, 2016; Peracchi & Rossetti, 2012). The vignettes were introduced into the Programme for International Student Assessment (PISA) in 2012 to correct for the impact of individual differences in response scale use on cross-country comparisons in attitudinal items (Cheung et al., 2018). Anchoring vignettes have improved item

discrimination and widened category thresholds as well as increased test information, reliability, and internal consistency (He, Buchholz, et al., 2017; Marksteiner et al., 2019; Primi et al., 2016; Weiss & Roberts, 2018).

In anchoring vignette procedures, a respondent is presented with a set of self-report items and a set of hypothetical scenarios depicting a particular level of the construct of interest; individual self-report items are then rescored relative to a person's ratings of the vignettes (Kapteyn et al., 2011). The measurement theory behind the vignettes is that because they are written to reflect an absolute level of the construct (e.g., low, medium, high), any systematic differences in responses to the same vignettes are supposed to reflect response styles that can be "subtracted" to yield a response-style variance free estimate of the latent construct (Hopkins & King, 2010; King & Wand, 2007). In this study, the methodological rationale behind the use of the vignettes was that if response scale use was the source of DIF, then accounting for individual response scale use, and by extension group-level scale use, would decrease group differences in measurement scales and therefore decrease DIF.

There are two approaches to rescoring self-report items with the vignettes. In the parametric approach, self-assessment equations are modeled relative to a set of vignette equations in a hierarchical ordered probit (HOPIT) model (Paccagnella, 2013). Group differences in reporting behaviors are accounted for by allowing self-report item thresholds to vary relative to fixed vignette item thresholds (von Davier et al., 2018; Weiss & Roberts, 2018). Although the HOPIT model reveals group differences, an important limitation is that individual responses are not adjusted for scale use (Grol-Prokopczyk et al., 2011). In contrast to the parametric approach, the nonparametric method utilizes respondent-dependent scoring to account for individual scale use (King & Wand, 2007; von Davier et al., 2018). Specifically, a person's self-report responses are

"rescaled" relative to their personal ratings of the vignettes (King & Wand, 2007; von Davier et al., 2018). This process "stretches" the response scale and increases item discrimination (King et al., 2004; Primi et al., 2016; von Davier et al., 2018).

The nonparametric scoring procedure produces a new set of vignette-adjusted item responses that have accounted for response scale use and can be modeled the same way as the unadjusted item response (King & Wand, 2007; Weiss & Roberts, 2018). The formula is based on *2J + 1* scale, where *J* is the number of vignettes. With three vignettes, depending on a person's ratings of Z_{ij} vignettes, their individual self-assessment responses, $\mathrm{X}_i \in \{1,2,3,4\}$, are transformed onto a 7-point vignette-adjusted variable on the *C*-scale, $\mathrm{Y}_i \in \{1,2,3,4,5,6,7\}$ as (von Davier et al., 2018):

	Vignette-Adjusted Score	Relationship Between Self-Report and Vignette Ratings
$C =$	$\mathrm{Y}_i = 1$ if $\mathrm{X}_i < \mathrm{Z}_1$	(lower than low-level vignette)
	$\mathrm{Y}_i = 2$ if $\mathrm{X}_i = \mathrm{Z}_1$	(same as the low-level vignette)
	$\mathrm{Y}_i = 3$ if $\mathrm{Z}_1 < \mathrm{X}_i < \mathrm{Z}_2$	(in-between the low and middle-level vignette)
	$\mathrm{Y}_i = 4$ if $\mathrm{X}_i = \mathrm{Z}_2$	(same as the middle-level vignette)
	$\mathrm{Y}_i = 5$ if $\mathrm{Z}_2 < \mathrm{X}_i < \mathrm{Z}_3$	(in-between middle and high-level vignette)
	$\mathrm{Y}_i = 6$ if $\mathrm{X}_i = \mathrm{Z}_3$	(same as the high-level vignette)
	$\mathrm{Y}_i = 7$ if $\mathrm{X}_i > \mathrm{Z}_3$	(higher than the high-level vignette)

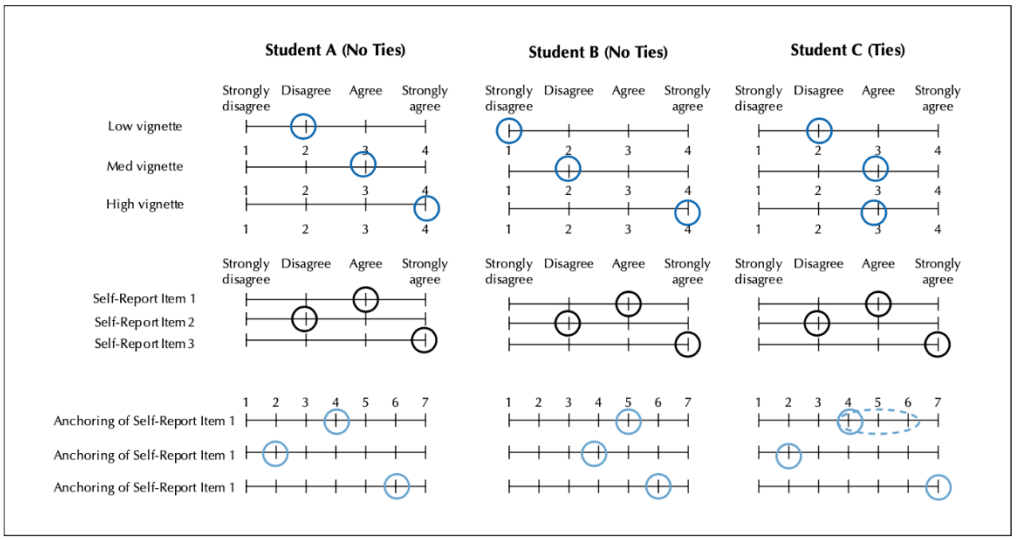
For example, if Respondent A rated items x_1 to x_4 as 1 (strongly disagree), 2 (disagree), 3 (agree), and 4 (strongly agree), and rated vignettes Z_1 as 2 (*disagree*), Z_2 as 3 (*agree*), Z_3 as 4 (*strongly agree*), then Respondent A's self-report item responses would be recoded as $(x_i = 1, y_i = 1)$, $(x_i = 2, y_i = 2)$, $(x_i = 3, y_i = 4)$, and $(x_i = 4, y_i = 6)$. See Figure 2 for a visual example of how the nonparametric formula is implemented. By expanding the response scale from four to seven points, item discrimination for vignette-adjusted scores is presumed to increase (He, Buchholz, et al., 2017), and those adjusted scores are expected to be free from response

style variance, which ultimately yields a more accurate measure of the construct (King et al., 2004; Primi et al., 2016) and is hypothesized therefore to reduce DIF.

However, the nonparametric method is not without disadvantages. In particular, the formula above shows that in order for the nonparametric method to yield a single scalar point value (i.e., 1, 2, etcetera) for vignette-adjusted self-report responses, the vignettes must be strictly ordered (i.e., $Z_1 < Z_2 < Z_3 \ldots < Z_x$), with no allowances for tied ratings. These requirements are directly related to the measurement assumptions for the vignettes, which are described in the next section.

Figure 2

Examples of Nonparametric Anchoring Vignette Adjustments



Note. From "PISA 2012 Technical Report," by the Organisation for Economic Co-operation and Development, 2014, (https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf).

Measurement Assumptions for Anchoring Vignettes

Anchoring vignettes rest on two measurement assumptions: response consistency and vignette equivalence. Response consistency is a within-person assumption that relies on respondents applying the same intrapersonal category threshold to both their self-report responses and to

their ratings of the vignette characters (King et al., 2004; von Davier et al., 2018). For example, a respondent must have the same internal cutoff for *agree* for their self-assessment and vignette ratings. Instead of explicitly testing for response consistency, it is usually assumed given the likelihood of a study's findings (Weiss & Roberts, 2018). Evidence that response consistencies violations occur, for example, when vignettes reflecting different construct levels are rated equally (i.e., tied) by a respondent (von Davier et al., 2018). With tied vignette responses, instead of a single value, rescored self-report responses actually take on a vector of values (King et al., 2004; Mõttus et al., 2012). For example, in Figure 2, Student C rated $Z_2 = Z_3$; therefore, Student C's vignette-adjusted score for Item 1 took on the range of values from four to six. See Appendix A for the complete distribution of nonparametric responses with three vignettes.

Because vectors yield less precise results than a single point estimate, researchers using the nonparametric scoring have developed various approaches to selecting a single point from the vector (King et al., 2004; Mõttus et al., 2012). One approach involves estimating the proportion of each response category and allocating a single value to the vector responses (King & Wand, 2007). King and Wand's (2007) solution was the HOPIT model that directly models threshold parameters. The OECD (2013) recommends using the lowest score from the vector as the minimum value "clearly pertains to the respondent rather than a higher value that just might pertain to the respondent" (p. 356). Using the lowest interval value has been shown to yield higher reliability (Primi et al., 2016, 2018; von Davier et al., 2018; Weiss & Roberts, 2018). However, after finding a relationship between vignette ordering and academic achievement with a sample of Brazilian adolescents, Primi et al. (2018) cautioned that recoding with the lowest interval value could lead to lower scores for respondents with order violations.

Vignette equivalence is a between-person assumption and refers to the idea that all respondents interpret the construct level depicted in a given vignette the same way (King et al., 2004). Vignette equivalence is usually tested by examining the order of vignette ratings, because if the vignettes are understood the same way by all respondents, they should rank-order the vignettes the same way (Bzostek et al., 2016; Grol-Prokopczyk, 2014; He, Buchholz, et al., 2017). Conversely, if respondents misorder the vignettes (i.e., rate a lower level vignette higher than a higher level vignette) then, by definition, they are not interpreting the vignettes the same way (Bzostek et al., 2016; He, Buchholz, et al., 2017). Although misordered vignettes are logically uninterpretable, if vignettes are poorly constructed or have low discrimination, discarding observations from students with ties or misordering could result in significant loss of information (Bago d'Uva et al., 2008). Therefore, PISA recommends that order violations are re-classified into ties at the highest category (e.g., if $Z_1 < Z_3 < Z_2$, Z = {1, 4, 2} is converted into Z = {1, 4, 4}) and then treated according to procedures for ties (OECD, 2013). Although the assumptions of response consistency and vignette equivalence are not always easy to meet, researchers have proceeded with analyses even when these assumptions have been violated (He, Van de Vijver, et al., 2017; Stankov et al., 2018; Weiss & Roberts, 2018).

Summary

In this section, group differences in achievement motivation, group differences in response scale use and response styles, and the impact of such differences on measurement scale comparability were described. Methods to address response scale use and how anchoring vignettes can be used to account for group differences in response scale as a source of DIF on DIF were also discussed. In the next section, the methodological conceptual framework and DIF method for this study are reviewed.

Methodological Conceptual Framework

Fairness in Assessment

Fairness in assessment is complicated and can be conceptualized in a variety of ways (Helms, 2006; Kane, 2010; McNamara & Roever, 2006; Xi, 2010; Zwick, 2019). From a philosophical perspective, fairness has to do with justice and the fairness of tests for particular groups (Kane & Bridgeman, 2017; McNamara & Roever, 2006; Nisbet & Shaw, 2019). This is especially relevant given the history of assessment and the adverse impact testing has had on groups that have been traditionally stigmatized or marginalized (McNamara & Roever, 2006; Nisbet & Shaw, 2019; Sireci & Rios, 2013). Persistent concerns about systematic inequalities in U.S. educational institutions demand that tests should not be used in biased (e.g., selection) or harmful ways (e.g., to discriminate, to reinforce stereotypes; Camilli, 2013; Kane & Bridgeman, 2017; Nisbet & Shaw, 2019). Tests should be logically and ethically defensible; test-takers should be treated equitably and with respect and sensitivity (Camilli, 2013; Kane, 2010). Such a broad perspective of fairness captures the historical roots of fairness and how it is fundamentally driven by concerns regarding the social, ethical, political, and legal implications of testing (Camilli, 2013; Nisbet & Shaw, 2019; Xi, 2010).

Fairness in Measurement

Fairness can also be defined more narrowly in relation to measurement and measurement scales (Boer et al., 2018; Poortinga, 1989). In the *Standards for Educational and Psychological Testing* (hereafter referred to as "the *Standards*"), a fair test is one that "reflects the same construct(s) for all test-takers, and scores from it have the same meaning for all individuals in the intended population" (AERA et al., 2014, p. 50). Fairness is operationalized in the *Standards* as fairness in treatment during the testing process, fairness in access to the constructs measured

(i.e., equal opportunity to show standing on construct), fairness in the validity of individual test score interpretations for the intended uses, and fairness as the absence of measurement bias.

Fairness as the absence of measurement bias has to do with the comparability of measurement scales (AERA et al., 2014). More specifically, in order to interpret an observed test score for its intended use, measurement scales must be equivalent such that the "same attribute…relate[s] to the same set of observations in the same way in each group" (Borsboom, 2006, p. S 176). When measurement scales are comparable (i.e., equivalent/invariant) for all respondents, then scores have the same psychological meaning and it is psychometrically fair to compare groups (Borsboom, 2006; McNamara & Roever, 2006; Poortinga, 1989; Ziegler & Brunner, 2016). However, if characteristics of items and tests function differently across groups such that groups were on different measurement scales, then that would threaten the fairness of cross-group comparisons (Kane, 2013; Messick, 1995; Osterlind & Everson, 2010). As Borsboom (2006) contended, "unless measurement invariance holds, fairness and equity cannot exist in principle" (p. 179).

Fairness and Differential Item Functioning

Within a fairness framework, measurement scale comparability can be evaluated using DIF (AERA et al., 2014; Zumbo, 2007). In a typical DIF analysis, respondents are separated into two groups, referred to as the reference group and focal group, and then they are matched by construct level (Oshima et al., 2015). The reference group usually reflects the majority, while the focal group includes the minority population for whom there is concern that an item might be unfair (Sireci & Rios, 2013; Turner & Keiffer, 2019). Matching groups by construct level is necessary to avoid confounding DIF with mean group differences (DeMars, 2010). After matching re-

spondents by construct level, if the statistical parameters for an item or test are the same, regardless of group membership, then that would indicate that groups are on the same measurement scale; that is, it would indicate fairness as the absence of measurement bias (Clauser & Mazor, 2005). If groups were not on the same scale, then DIF would uncover the presence of those group differences (Osterlind & Everson, 2010).

When evaluating two groups, such as gender, which group is assigned to serve as the reference group is arbitrary for statistical purposes (Ellis & Kimmel, 1992; Oshima et al., 2015). However, with a multilevel group such as ethnicity that lends itself to a multiple-group DIF (MG-DIF) analysis, the choice of reference group becomes more important (Ellis & Kimmel, 1992; Sari & Huggins, 2015). In particular, even if it is inadvertent, researchers make an underlying value statement about fairness and standards for comparison when they select a specific group as the reference group (e.g., using White students as the reference group implies that White students should be the reference point; Martinková et al., 2017; Sari & Huggins, 2015; Sue, 1996). To avoid this predicament, a composite, or common base group that reflects an average sample of the total multiple-group sample size, can be constructed to serve as the reference group in an MG-DIF (Ellis & Kimmel, 1992; Oshima et al., 2015; Sari & Huggins, 2015). In relation to a common base group, fairness reflects the absence of DIF between the base group and the focal ethnic group.

Ellis and Kimmel (1992) further argued that when the reference group reflects the average, any difference between the base group and the focal group captures group-specific response patterns. While it is true that DIF in relation to an average group reveals the difference in measurement scales between the focal group and the base group, it cannot necessarily be assumed that those differences are *only* due to response scale use. Correcting for response scale use with the

vignettes can provide additional evidence that the source of DIF in relation to the base group is response scale use. To that end, if groups show reduced DIF in base group comparisons after self-report responses are adjusted with the vignettes, then that could indicate the presence of group-specific differences in response scale use.

Hybrid DIF Framework

There are myriad DIF detection procedures and almost as many taxonomies to classify such procedures. Within the context of fairness testing, Camilli (2006) divided DIF procedures into those based on item response models (IRT) and those based on observed-score analysis, such as Mantel-Haenszel methods and logistic regression (LR). Logistic regression DIF is advantageous over other observed-score-based DIF methods because LR models can accommodate polytomous items, use either a discrete or continuous matching variable, be extended to multiple groups, and identify both uniform and nonuniform DIF (Magis et al., 2011; Millsap & Everson, 1993; Swaminathan & Rogers, 1990; Zumbo, 2007). Uniform DIF indicates that one group is consistently favored by an item; that is, after groups are matched by construct level, the item is easier for one group to endorse (Jodoin & Gierl, 2001; Sireci & Rios, 2013). Nonuniform DIF reflects an interaction between groups and the construct such that the magnitude and direction of group differences in response probabilities vary along the ability spectrum (Jodoin & Gierl, 2001; Sireci & Rios, 2013).

Because IRT estimates are modeled as a function of both item characteristics and person characteristics, IRT estimates are more precise than the total score typically used to match participants in observed-score-based DIF methods like LR (Sharkness, 2014). With short scales where the score from the item being investigated for DIF is included in the matching variable, the accuracy of the total score becomes even more important given that total scores yielded from fewer

items are generally less reliable under the assumptions of classical test theory (Balluerka et al., 2014; Crane et al., 2006; DeVellis, 2006; Millsap & Everson, 1993). As such, modern DIF methods have begun to integrate different DIF frameworks into "hybrid" approaches that include elements from latent trait (e.g., IRT) and observed-score DIF methods (e.g., LR; Choi et al., 2011; Millsap, 2006). In this study, each item was fit to a graded response model (GRM), and then the difference in likelihood ratios tests between GRMs for each group was tested for DIF using ordinal logistic regression. Consequently, instead of the observed score that is usually the matching criterion for an LR DIF, respondents are matched with a GRM-derived trait estimate (Choi et al., 2011).

In the next section, IRT modeling and the GRM are introduced to provide a foundation for the detailed description of ordinal logistic regression DIF.

Graded Response Model

In IRT, the relationship between the construct of interest and probability of an item response is captured through a nonlinear, monotonically increasing logistic curve called an item response function (IRF; Raju et al., 2002; Reise, 2014). The mathematical formula depicting the IRF for a 2-parameter binary logistic model is:

$$P(x_i = 1|\theta) = \frac{\exp\left(a_i(\theta - b_i)\right)}{1+\exp\left(a_i(\theta - b_i)\right)}, \quad (1)$$

where $P(x_i = 1|\theta)$ is the probability of a correct response to item i conditioned on examinee trait level θ, a is the discrimination (i.e., slope) parameter, and b is the difficulty (i.e., location) parameter. Item discrimination describes the strength of the relationship between θ and the item (Teresi et al., 2012). The steeper the slope of an IRF, the more differentiating the item is, with item discrimination generally falling between 0 and 2 (Hambleton et al., 1991; Reise, 2014).

Higher slopes provide more discrimination and information around the middle of the trait continuum, so items with larger slopes discriminate better between individuals in the average range as compared to items with smaller slopes (Reise, 2014). The *b* parameter provides item location and indicates the point along the latent construct spectrum where the probability of endorsing an item response is .50 (Hambleton et al., 1991). Item difficulty generally ranges from -2.0 to +2.0, with IRFs for the easiest items shifted negatively to the left on the x-axis and the IRFs for harder items shifted positively to the right on the x-axis (Hambleton et al., 1991; Reise, 2014).

The GRM is one polytomous item extension of IRT models (Samejima, 2010). For dichotomous items, two IRFs are calculated to reflect the probability of a correct or incorrect answer (Cohen et al., 1993). For polytomous items, multiple IRFs must be estimated to reflect the probability of selecting a particular response category (Reise, 2014). Specifically, the GRM estimates category response functions (CRF) for *k* ordered categories, where the CRF captures the relationship between the construct level and the cumulative probability of selecting at or above response category *k* (Cohen et al., 1993; Reise, 2014). CRFs for *k*-1 (*b*) thresholds are estimated, with each response category having a boundary response function (BRF) that is estimated as a function of one common slope. The BRF indicates the construct level needed to have a 50% or higher chance of responding above the threshold between adjacent categories (Reise, 2014). Items with high discrimination yield steeper BRFs, and narrower, more peaked CRFs (Oshima & Morris, 2008; Reise, 2014). When the distance between boundary functions is large, an item discriminates well across all levels of the construct; if the distance is narrow, the opposite is true (Reise, 2014; Song et al., 2011). The BRF is functionally equivalent to the IRF for a dichotomously scored item (Raju et al., 2002), and the sum of the IRFs for a polytomous item yields an

item true score function that is equivalent to the IRF for the second category of a dichotomously scored item (Cohen et al., 1993).

With the GRM, DIF can be visualized by plotting the IRFs and CRFs for each group and evaluating the difference between groups across the θ range (Oshima & Morris, 2008). However, although IRFs can inform the type of DIF (i.e., uniform or nonuniform), Penfield (2007) argued that studying the patterns of between-group differences in response probabilities for each score level within a differential step functioning framework can provide insight into the causes of DIF in a polytomous item. In step taxonomy, causes of DIF can be identified by examining the location of DIF in an item (i.e., *pervasive* or *non-pervasive*) and the consistency of DIF magnitude/sign across score levels (i.e., *constant*, *convergent*, or *divergent*; Penfield et al., 2009).

The source of DIF is *constant* and *pervasive* when all score levels show equally "substantial" differences in category thresholds between groups and indicate that the cause of DIF is located at the item level (e.g., item prompt, item content). In contrast, *non-pervasive* effects are present when only one or a few response categories evidence threshold differences (Penfield et al., 2009); non-pervasive patterns indicate that the source of DIF is not at the item level, but rather is caused by something at a particular score level (Penfield et al., 2009). When category thresholds consistently favor one group, but the size of the threshold differences between groups varies according to score level (i.e., *convergent* effects), the source of DIF is due to either the unequal distribution of item-level DIF effects at each score level or the presence of multiple sources of DIF affecting different score levels (Penfield et al., 2009). Conversely, a shift in category thresholds from favoring one group to the other group across score levels (i.e., *divergent* effects) offers evidence of multiple sources of DIF that are local to a particular score level (Penfield et al., 2009).

Because the GRM is the polytomous item extension of the 2-parameter logistic response model, when the GRM is used to generate the construct estimate, the GRM and ordinal logistic regression models are nearly equivalent, with both models reflecting the same hypothesis about the relationship between items and persons (Crane et al., 2006; Gnambs & Hanfstingl, 2014). As such, the following description of ordinal logistic regression models is assumed to also apply to the GRM formulation used in the hybrid DIF framework.

Ordinal Logistic Regression DIF Models

Logistic Regression DIF. In the original formulation for an LR DIF with dichotomous items, the probability of a correct response takes the same form as Equation 1, and the conditional difference (i.e., DIF) between groups is modeled as (Swaminathan & Rogers, 1990):

$$P(u_{ij} = 1|\theta_{ij}) = \frac{e^{(\beta_{0j} + \beta_{1j}\theta_{1j})}}{[1 + e^{(\beta_{0j} + \beta_{1j}\theta_{1j})}]}, i = 1, \ldots, n_j, j = 1, 2, \quad (2)$$

where u_{ij} is the probability of item response *u* by person *i* in group *j*, θ is the construct level, β_{0j} is the intercept parameter, and β_{1j} is the slope parameter for group *j*. According to Equation 2, if the logistic regression curves for the two groups are identical (i.e., $\beta_{01} = \beta_{02}$ and $\beta_{11} = \beta_{12}$), then DIF is not flagged. If $\beta_{01} \neq \beta_{02}$ (i.e., unequal intercepts) but $\beta_{11} = \beta_{12}$ (i.e., equal slopes), then the difference in intercepts indicates uniform DIF, with one group consistently scoring higher (i.e., favored) than the other across θ. If $\beta_{01} = \beta_{02}$, but $\beta_{11} \neq \beta_{12}$, then the logistic curves for the two groups are not parallel, which indicates nonuniform DIF and that the difference between groups varies along the ability spectrum. The difference between models yields an omnibus likelihood ratio χ^2 statistic that can be tested for statistical significance against a χ^2 distribution with 2 *df* (Swaminathan & Rogers, 1990).

Ordinal Logistic Regression DIF. Logistic regression DIF models can be extended to accommodate ordered polytomous items using a variety of methods. For example, the response

probabilities for adjacent categories can be compared, or polytomous items can be recoded into a series of dichotomous logistic regressions (Agresti, 2002; Miller & Spray, 1993). Another approach involves fitting a sequence of cumulative probabilities, or logits (Agresti, 2002; French & Miller, 1996; Miller & Spray, 1993). With polytomous items, logits are the natural logarithm of the ratio reflecting the probability of selecting a response category to the probability of not selecting a response category (French & Miller, 1996). In a cumulative logit model, ordered response categories *C* become *C* – 1 cumulative logits, with each logit having its own intercept. The cumulative logit in a regression model of ordinal responses is (Zumbo, 1999):

$$\text{logit}[P(Y \leq j)] = \alpha_j + \beta(\text{X}), j = 1, \ldots C - 1, \quad (3)$$

where *j* is the category selected (e.g., agree, strongly agree), *C* is the number of response options (e.g., 3-point scale), and X is a set of predictors (Agresti, 2002; Zumbo, 1999). The effect of a given predictor in Equation 3 is assumed to be constant for each cumulative probability across all *j* such that the odds of selecting a response category is the same for all possible combinations when *C*-category responses are collapsed to a binary variable. Because β is constant, logits can be incorporated into a proportional odds model that simultaneously uses all cumulative logits and indicates the cumulative odds of selecting a response category (Agresti, 2002).

Equation 3 can be re-written to express the full ordinal logistic regression DIF model as (Osterlind & Everson, 2010; Zumbo, 1999):

$$Logit\ P(u_i \geq k) = \alpha_k + \beta_1\theta \quad \text{(Model 1)}$$

$$Logit\ P(u_i \geq k) = \alpha_k + \beta_1\theta + \beta_2 group \quad \text{(Model 2)}$$

$$Logit\ P(u_i \geq k) = \alpha_k + \beta_1\theta + \beta_2 group + \beta_3(\theta * grou), \quad \text{(Model 3)}$$

where $P(u_i \geq k)$ is the probability of item response *u* to item *i* is category *k* or higher, α_k is the category intercept and reflects the maximum likelihood probability of selecting response category *k*, θ is the construct estimate derived from the GRM, *group* indicates the reference group and captures the dependency between an item response and group membership, and $(\theta * group)$ describes the interaction between *group* and θ.

Multiple-Group DIF. Although multiple-group DIF has not been well-studied (Magis et al., 2011; Oshima et al., 2015), ordinal logistic regression DIF models can also be extended to accommodate multiple groups by including common and group-specific parameters as (Finch, 2016; Magis et al., 2011):

$$Logit(\pi_{ig}) = \alpha + \beta S_i + \alpha_g + \beta_g S_i, \quad (4)$$

where α is the common intercept across groups and β is the common slope across groups. Binary indicators can be added to each β parameter for all groups except one to reveal group-specific effects (Choi et al., 2011).

Ordinal Logistic Regression DIF Detection

Ordinal logistic regression (OLR) DIF begins by first fitting Models 1, 2, and 3 as hierarchically structured logistic regressions (Choi et al., 2011). The models evaluate if the separate (i.e., uniform DIF) and combined effects (i.e., nonuniform) of group and group x ability interaction on the probability of an item response are statistically significant, above and beyond the ability estimate (Zumbo, 1999). The dependent variable is the probability of selecting a response category (i.e., item response). The independent variables are the matching criterion, the group term, and the interaction term. The models are implemented sequentially as a likelihood ratio, beginning with the matching criterion, followed by the group variable, followed by the interaction term (Zumbo, 1999). In the hybrid DIF framework, each GRM yields a likelihood ratio χ^2

test, and the difference in likelihood ratio χ^2 tests between nested models is tested for significance against a χ^2 distribution using OLR to flag DIF (Choi et al., 2011). That is, the *p*-value from the χ^2 test associated with the -2 log-likelihood difference between nested models is used as the DIF flagging criterion. A significant *p*-value indicates that the more complex model fits the data as well as the model to which it was being compared (Segeritz & Pant, 2013).

DIF is statistically detected by comparing hierarchically nested models in two stages (Crane et al., 2006; Lai et al., 2005; Oliveri et al., 2016). In the first stage, the difference between Model 3 and Model 1 is tested for omnibus DIF or "total DIF effect" (Choi et al., 2011). Under the omnibus DIF test (*df* = 2), the aggregated effects of uniform and nonuniform DIF are reflected in $H_0: \beta_2 = \beta_3 = 0$ (Swaminathan & Rogers, 1990) and $H_1: \beta_2 \neq 0$ (uniform DIF) and $H_1: \beta_3 \neq 0$ (nonuniform DIF; Jodoin & Gierl, 2001; Scott et al., 2010). This test is interpreted as: after controlling for the level of the construct, there is or is not a significant relationship between group membership and the probability of an item response. If the relationship is significant, then the item is flagged for DIF.

The second stage of DIF detection is partitioned into two steps, each with a *df* = 1 likelihood ratio χ^2 test (Crane et al., 2006). In the first step, nonuniform DIF is assessed by comparing Model 3 to Model 2 (Crane et al., 2006). The null hypothesis is $H_0: \beta_3 = 0$; if β_3 is significant, then nonuniform DIF is detected and group differences are assumed to vary by construct level (Sireci & Rios, 2013). If $\beta_3 > 0$, then the item favors students with higher levels of the construct in the reference group and lower construct levels in the focal group; if $\beta_3 < 0$, the opposite pattern occurs (Jodoin & Gierl, 2001). In the second step, uniform DIF is detected by comparing Model 2 to Model 1 (Choi et al., 2011; Crane et al., 2006). The null hypothesis tested is $H_0: \beta_2 = 0$; if β_2 is significant, then uniform DIF is detected. If $\beta_2 > 0$, then the item favors the reference

groups; if $\beta_2 < 0$, then the item favors the focal group (Jodoin & Gierl, 2001). β_2 reveals if there is a significant difference between group responses after controlling for overall construct level (i.e., the effect of group membership), while β_3 reveals the consistency of those differences across θ (Teresi & Jones, 2016; Zumbo, 2007).

Effect Sizes and Power. Because statistical power is a function of sample size, DIF is often statistically detected in large samples (Jodoin & Gierl, 2001; Kim et al., 2007). Further compounding the problem is that when logistic regression is used for DIF analyses, the multiple comparisons can lead to inflated type I errors (Jodoin & Gierl, 2001). The combined effect of a large sample size with multiple comparisons means that DIF will be flagged frequently, but the DIF may have little practical meaning (Suh, 2016). As Crane et al. (2007) noted, "Especially when there are large sample sizes, it is possible to have a statistically significant but practically irrelevant relationship between a demographic covariate and the probability of item responses" (p. 12). Therefore, to control for type I errors, the combined use of both an effect size that quantifies the magnitude of DIF as well as criteria to determine if any DIF is substantively meaningful or practically significant is necessary (Jodoin & Gierl, 2001; Stark et al., 2004; Suh, 2016).

Although logistic regression DIF is somewhat disadvantaged by the absence of an associated effect size, pseudo R^2 values[2] are often used to quantify the magnitude of OLR DIF (Jodoin & Gierl, 2001; Zumbo, 1999). Differences in pseudo R^2 values between models describe the amount of variance added to a model when a new term is introduced to the regression equation (Oliveri et al., 2012; Zumbo, 1999). Each OLR DIF yields a corresponding pseudo R^2 value that

[2] This is the same R^2 that is typically reported in an ordinary least squares regression. Note that in logistic regression, R^2 is called pseudo R^2.

can be used as the DIF effect size (Jodoin & Gierl, 2001; Zumbo, 1999). For example, the difference in pseudo R^2 values between Models 1 and 2 can be defined as (Jodoin & Gierl, 2001):

$$\Delta R^2 = {R^2}_{\mathrm{M1}} - {R^2}_{\mathrm{M2}} \quad (4)$$

The magnitude of nonuniform DIF is reflected in the difference between Model 3 and Model 2, which describes the unique variation attributed to the group-ability interaction; therefore, it is considered a measure of how much nonuniform DIF is present in an item (Zumbo, 1999). The magnitude of uniform DIF can be calculated as the change in R^2 between Model 2 and Model 1 (Equation 4), which reflects the variation attributable to group membership (Zumbo, 1999).

There are a variety of pseudo R^2 statistics, including the Cox & Snell, Nagelkerke, and McFadden, that are all calculated slightly differently and therefore have their respective advantages and disadvantages (Choi et al., 2011; Menard, 2000). One of the benefits of reporting McFadden's pseudo R^2 is that it can be meaningfully interpreted as the proportional reduction in the -2 log-likelihood ratio statistic (Choi et al., 2011; Menard, 2000). The value indicates the likelihood of improving model fit and reducing error variation when predictors are used in the model as compared to when predictors are not used in the model (i.e., baseline or intercept model; Menard, 2000). Various criteria have been used to classify the "practical" meaning of pseudo R^2 values (Jodoin & Gierl, 2001; Suh, 2016; Zumbo, 1999). From their simulation study, Jodoin and Gierl (2001) proposed threshold values of .035 and .070. However, those recommendations were based on simulation studies in which the total observed score, and not an IRT-generated estimate, was used as the matching criterion.

In addition to pseudo R^2, Jodoin and Gierl (2001) suggested that changes to logistic regression β coefficients can also be used as a uniform DIF effect size. Based on the proportional

odds assumption, the magnitude of uniform DIF can be quantified by the percentage change in β_1 from Model 1 to Model 2 as (Crane et al., 2006):

$$|(\beta_1 - \beta_2)/\beta_2|. \quad (5)$$

To determine how much of a percentage change is practically meaningful, various cutoffs have been used, ranging from a change of 1% to 10% (Choi et al., 2011; Crane et al., 2006, 2007; Feuerherd et al., 2014; Lambert, Garcia, Epstein, & Cullinan, 2018). However, Crane et al. (2007) found that a 1% change was not clinically relevant. As such, researchers have generally used either a 5% or 10% to indicate the practical meaning of the DIF (i.e., negligible, slight, moderate; Crane et al., 2006; Lambert, Garcia, Epstein, & Cullinan, 2018; Lambert, Garcia, January, & Epstein, 2018).

In terms of power and sample size, based on their simulation study of OLR DIF, in which an IRT-generated ability estimate was used as the matching criterion, Scott et al. (2009) recommended a sample size of at least 200 respondents per group to ensure about 80% power; with short scales (i.e., only two items), they suggest 300 respondents per group. OLR DIF is also robust to small rates of missing completely at random responses (i.e., 10%; Robitzsch & Rupp, 2009). However, one concern regarding power and OLR DIF has to do with scale length (Scott et al., 2007). In DIF analyses of short scales (i.e., no more than five items), particularly with polytomous items, concerns arise regarding "pseudo-DIF" and power (Hidalgo et al., 2016; Scott et al., 2009). Pseudo-DIF describes how sometimes when true DIF is seen in one item, other items will demonstrate DIF effects in the opposite direction; on short scales, it can be hard to detect which item is causing the pseudo-DIF, particularly in the presence of crossing DIF or DIF cancellation (Scott et al., 2007).

3 METHODOLOGY

The purpose of this study was to identify if achievement motivation items demonstrated measurement scale comparability across gender and ethnicity before and after using anchoring vignettes to account for the effect of group-level differences in response scale use as a source of differential item functioning (DIF). To that end, a hybrid DIF framework that combined item response theory and ordinal logistic regression was used to address the following research questions: Do achievement motivation items show DIF by gender or ethnicity? Do achievement motivation items show DIF by gender or ethnicity after using anchoring vignettes to account for the impact of response style differences?

To answer these questions, first, gender DIF was evaluated using pairwise comparisons, and ethnicity DIF was tested using both multiple-group DIF and pairwise comparisons. Next, group differences in response scale use was accounted for by applying the nonparametric scoring of anchoring vignettes to self-reported motivation items. Finally, the same set of DIF analyses from the first research question were re-run with the vignette-adjusted items. To evaluate the effect of the vignette correction on DIF, results from unadjusted and vignette-adjusted DIF were compared for changes in DIF outcome as well as changes to DIF magnitude and form.

Data Source

Data for this study came from the 2015 iteration of the Programme for International Student Assessment (PISA)[3]. PISA is an international large-scale assessment administered triennially to 15-year-old students in over 70 countries (OECD, 2017). PISA is a 2-hour test designed to be independent of school curricula; that is, rather than evaluate domain-specific content, PISA assesses if students can apply what they learned to real-life situations (Hopfenbeck et al., 2018;

[3] Data are publicly available for download at https://www.oecd.org/pisa/data/2015database/.

Kaplan & Kuger, 2016). To reduce the testing burden on students, PISA's cognitive scales are administered using multiple-matrix sampling and balanced incomplete block design (Kaplan & Kuger, 2016; Rutkowski et al., 2013). The PISA 2015 background questionnaire included a variety of items asking students about their beliefs and attitudes as well as three anchoring vignettes related to achievement motivation. In contrast to the matrix sampling used for cognitive items, the PISA background questionnaire is administered to all students. Therefore, any missing responses to achievement motivation items were considered missing-at-random (OECD, 2017).

PISA's two-stage stratified sampling design is supposed to yield nationally representative samples of 15-year-olds in school when analyses are weighted (Kaplan & Kuger, 2016; OECD, 2017; Rutkowski & Rutkowski, 2016). The OECD (2017) recommends using sampling weights for any inferential analysis with PISA data. However, the decision to use sampling weights with analytical models that make inferences about item and model parameters is less straightforward (Cai, 2013; Solon et al., 2015). If a specified model correctly reflects the relationship between the response variable and covariates and the probability of inclusion is uncorrelated with the response variable, then the absence of sampling weights does not bias model estimates (Cai, 2013). Moreover, in the context of latent trait modeling, survey weights are not attributes of an individual, but rather they are survey-dependent and constructed from survey data (Gelman, 2007). Furthermore, as model complexity increases, the use of sampling weights makes the interpretation of model parameters less straightforward (Gelman, 2007). Finally, a review of DIF analyses with PISA data did not find that researchers consistently used (e.g., Oliveri et al., 2014) or addressed sampling weights (e.g., Hopfenbeck et al., 2018). For these reasons, all DIF analyses were run as unweighted.

Participants

In the U.S., PISA 2015 was administered in 177 schools across Massachusetts, North Carolina, and Puerto Rico, with a total sample size of 5,712 (OECD, 2017). Sample demographics can be seen in Table 1. A total of 58 students (0.01%) were excluded from the ethnicity analyses because their response was missing or they did not self-identify with one of the response options. The final analytic sample size was 5,712 for gender and 5,654 for ethnicity.

Table 1

Sample Characteristics

Group	n	%
Gender		
Male	2854	50.0
Female	2858	50.0
Ethnicity		
White, not Hispanic	2498	43.7
Black/African American	790	13.8
Hispanic/Latinx	1761	30.8
Asian	207	3.6
Multi-racial/Other	398	7.0

Measures

Achievement Motivation

Achievement motivation was measured in PISA 2015 background questionnaire with five items (see Table 2). Response options were based on a 4-point scale, ranging from *strongly disagree* to *strongly agree.* The items included content about students' goals; however, the standards for competence were unclear (OECD, 2019). The items also assessed a competitive desire to outperform others (OECD, 2019). Therefore, students with a high score on this variable were presumed to regulate their motivational disposition to approach success and to compete with others by adopting achievement goals with the aim of outperforming peers (Marsh et al., 2006; Mu-

rayama & Elliot, 2012; OECD, 2019). Cronbach's α was calculated in the *ltm* package (Rizopoulos, 2006) in R software, with a cutoff of .80 as the criteria for acceptable reliability (Peterson, 1994; Vaske et al., 2017).

Table 2

PISA 2015 Achievement Motivation Items

Variable	Achievement Motivation Item
Item 1	I want top grades in most or all of my courses.
Item 2	I want to be able to select from among the best opportunities available when I graduate.
Item 3	I want to be the best, whatever I do.
Item 4	I see myself as an ambitious person.
Item 5	I want to be one of the best students in my class.

Anchoring Vignettes

In PISA 2015, three achievement motivation anchoring vignettes (see Table 3) were administered directly following the self-report items using the same 4-point response scale ranging from *strongly disagree* to *strongly agree*. The vignettes were written to reflect low, medium, and high levels of motivation and intended to be ordered as $Z_1 < Z_2 < Z_3$.

Table 3

PISA 2015 Achievement Motivation Anchoring Vignettes

Vignette (Z_i)	Achievement Motivation Anchoring Vignettes
Z_1	Mario gives up easily when confronted with a problem and is often not prepared for his classes. Mario is motivated. (Z_1)
Z_2	Olivia mostly remains interested in the tasks she starts and sometimes does more than what is expected from her. Olivia is motivated. (Z_2)
Z_3	John wants to get top grades at school and continues working on tasks until everything is perfect. John is motivated. (Z_3)

Procedures

To test for DIF in achievement motivation items across gender and ethnicity, this study fit a series of hierarchically nested ordinal logistic regression models for each of the five PISA items using the *lordif* package (Choi et al., 2011) in R software. *Lordif* estimates omnibus, uniform, and nonuniform DIF by implementing a hybrid item response theory (IRT)/ordinal logistic regression (OLR) DIF framework in which the matching criterion is derived from the graded response model (GRM). DIF values are obtained from an iterative algorithm that generates trait estimates using group-specific item parameters and DIF-free items. Those trait estimates are used in subsequent DIF analyses until DIF is flagged across successive iterations (Choi et al., 2011). For a description of the *lordif* algorithm, see Appendix B.

Research Question 1

Overview

To answer Research Question 1, two primary steps were taken. First, the psychometric properties of the achievement motivation items were evaluated. This included calculating descriptive statistics for the items (mean, standard deviation, and percentage of missing responses), using exploratory factor analysis to evaluate dimensionality, and fitting items to a GRM to assess model assumptions and item fit. Second, OLR DIF models were used to detect gender DIF in pairwise comparisons and ethnicity DIF in multiple-groups and pairwise comparisons. Procedures to complete these steps are described in further detail below.

Psychometric Assessment of Motivation Items

Graded Response Model Assumptions and Item Fit Testing. To test for unidimensionality, exploratory factor analysis (EFA) was applied to the five motivation items in MPlus soft-

ware (Muthén & Muthén, 2017) using weighted least squares mean and variance, which is relatively robust to latent trait distribution violations in the sample (Marsh et al., 2003). The single-factor structure for the EFA was determined based on the first eigenvalue being greater than 1 and containing at least 20% of the variance as well as EFA residual variance and the root mean square error of approximation (RMSEA; Finch, 2020; Reeve et al., 2007; Zopluoglu & Davenport, 2017). In his simulation study of EFA with categorical variables, Finch (2020) identified an RMSEA cutoff of .015 as indicating fit for a single-factor model, and Rutkowski and Svetina (2014) recommend an RMSEA cutoff of .030 with large samples and multiple groups when assessing unit (i.e., factor) equivalence. To test for local independence, the motivation items were also fit to a GRM in the *mirt* package (Chalmers, 2012) in R software, with an absolute correlation of greater than .20 between item pairs as the cutoff for dependence (Forrest et al., 2014; Reeve et al., 2007). The ascending order of response category thresholds was examined for monotonicity (Forrest et al., 2014; Zhong et al., 2014). Given the large sample size and the sensitivity of χ^2 tests to sample sizes and multiple-groups (Rutkowski & Svetina, 2014), the RMSEA associated with the S-χ^2 for each item was used to assess item fit.

Ordinal Logistic Regression Models

Gender DIF Model. To identify if achievement motivation items show DIF by gender or ethnicity (Research Question 1), the OLR DIF models for gender and ethnicity were constructed as follows. Gender is defined in PISA as male/female (OECD, 2017). In this DIF analysis, boys were coded as the reference group, with girls assigned to the focal group (males = 0, female = 1). However, it is important to note that dichotomizing gender as male/female does not capture the complexities of gender as a social construct (Westbrook & Saperstein, 2015). Nonetheless, Westbrook and Saperstein (2015) found that defining gender this way is the current protocol in social

sciences research. Gender DIF was evaluated with pairwise comparisons using the following models:

$$Logit\ P(u_i \geq k) = \alpha_k + \beta_1\theta \quad \text{(GM 1)}$$

$$Logit\ P(u_i \geq k) = \alpha_k + \beta_1\theta + \beta_2 Bos \quad \text{(GM 2)}$$

$$Logit\ P(u_i \geq k) = \alpha_k + \beta_1\theta + \beta_2 Boys + \beta_3(\theta * Boys), \quad \text{(GM 3)}$$

where $P(u_i \geq k)$ is the probability of item response *u* to item *i* is at category *k* or higher, α_k is category intercept, θ is the GRM estimate of achievement motivation, *group* reflects boys as the reference group, and $(\theta * Boys)$ describes the interaction between θ and group membership.

Ethnicity DIF Model. In the PISA 2015 background survey, the response options to describe racial/ethnic background were: White, Black or African American, Asian, American Indian or Alaska Native, and Native Hawaiian or other Pacific Islander (OECD, 2017). Students were also asked to report their Hispanic status. Responses to these two questions were collapsed into a combined[4] variable (RACETHC) with the following categories: White, not Hispanic (1), Black or African American (2), Hispanic or Latino (3), Asian (4), Multi-Racial (5), and Other (6). Due to the small size of the other group (n = 62), the Other and Multi-racial categories were combined into a single category of Multi-racial/Other.

Ethnicity DIF was evaluated using both multiple-group DIF (MG-DIF) and all possible pairwise comparisons. The DIF model for ethnicity was:

$$Logit\ P(u_i \geq k) = \alpha_k + \beta_1\theta \quad \text{(EM 1)}$$

$$Logit\ P(u_i \geq k) = \alpha_k + \beta_1\theta + \beta_2 Group \quad \text{(EM 2)}$$

$$Logit\ P(u_i \geq k) = \alpha_k + \beta_1\theta + \beta_2 Group + \beta_3(\theta * Group), \quad \text{(EM 3)}$$

[4] The PISA 2015 Technical Manual does not describe how the two questions were combined to derive the RACETHC variable.

where $P(u_i \geq k)$ is the probability of item response u to item i at category k or higher, α_k was the category intercept, and θ was the GRM estimate of motivation. For the pairwise comparisons, *group* reflects the reference group, and $(\theta * group)$ describes the interaction between θ and the reference group. For example, in the comparison between Black/African American and Asian students, Black/African American students would be the reference group, with their parameters being reflected in the *group* term in EM 1, EM 2, and EM 3.

For the MG-DIF, *group* reflects the common base group that served as the reference group against which each focal ethnic group was compared (e.g., Asian vs. base group; Hispanic/Latinx vs. base group; Oshima et al., 2015). To create a base group reflecting the sample average, an unweighted random sample was drawn from the total sample of multiple groups and divided by the total number of groups in the analysis (Oshima et al., 2015). In this case, the sample of $N = 5712$ was divided by five (i.e., the number of groups) to yield a base group of $n = 1142$ students (5712/5 = 1142.2). Next, the ethnicity variable for students in the base group was recoded to reflect their base group status. Once the base group was created, the ethnicity variable had six levels—five ethnic groups and the base group. In the MG-DIF, there were $n = 2003$ White students, $n = 639$ Black/African American students, $n = 1396$ Hispanic/Latinx students, $n = 166$ Asian students, and $n = 319$ students in the Multi-racial/Other group. The OLR models were extended to accommodate the multiple groups by including a binary indicator to β_2 and β_3 for all groups except for one (Choi et al., 2011). To detect DIF, each ethnic group was compared to the base group using EM 1, EM 2, and EM 3.

DIF Detection. After participants were matched by θ, the same two-stage DIF detection process was applied to all items. First, Model 1 was nested in Model 3 to identify the omnibus, or "total," DIF effect ($df = 2$). Second, the difference between Model 3 and 2 was evaluated for

nonuniform DIF (*df* = 1), followed by comparing Model 2 to Model 1 to detect uniform DIF (*df* = 1). DIF was flagged based on the *p*-value associated with the -2 log-likelihood difference in χ^2 tests between nested models, with a significant *p*-value indicating the presence of DIF. Item response functions (IRF) were visually examined for crossing or nonparallel curves that would indicate nonuniform DIF (Bolt & Gierl, 2006; Su & Wang, 2005) and the absolute summed difference between IRFs, weighted by focal group, was calculated to reflect the impact of DIF on group differences in response probabilities. The pattern of group differences in response category thresholds was examined within a differential step functioning framework to provide information about the location and source of DIF (Penfield et al., 2008, 2009).

Effect Sizes. To control for the increased risk in Type I errors associated with large sample sizes and the multiple comparisons involved in testing for nonuniform DIF, significance was set to $p < .01$ for all analyses (Oliveri et al., 2016). Additionally, the practical significance of DIF effect sizes, reported as McFadden's pseudo R^2 because it reflects the proportional reduction in the -2 log-likelihood ratio statistic (Choi et al., 2011; Menard, 2000), was classified according to the following criteria: negligible DIF if $\Delta R^2 < .035$, moderate DIF if $.035 \leq \Delta R^2 \leq .070$, and large DIF if $\Delta R^2 > .070$ (Hidalgo & López-Pina, 2004; Jodoin & Gierl, 2001; Oliveri et al., 2012). Regression β coefficients were also used as an effect size for uniform DIF by comparing the percentage change in β_1 from Model 1 to Model 2 (Crane et al., 2006; Jodoin & Gierl, 2001); a change of greater than 5% was classified as slight to moderate DIF (Crane et al., 2006; Lambert, Garcia, January, & Epstein, 2018).

Sample Size, Power, and Handling of Missing Data

Based on their simulation study of OLR DIF, in which an IRT-generated ability estimate was used as the matching variable, Scott et al. (2009) recommended a sample size of at least 200

respondents per group to ensure about 80% power; with short scales (i.e., fewer than five items), they suggested 300 respondents per groups. Table 1 shows that there were just over 200 Asian students in the sample; therefore, power could be problematic in comparisons with that group. Additionally, the minimum cell size per response category was set to five to maintain adequate power for item fit indices; categories that did meet that minimum were collapsed with adjacent cells (Kang & Chen, 2008; Orlando & Thissen, 2000).

In PISA 2015, *valid skip*, *not reached*, *not applicable*, *invalid*, and *no response* were coded as missing. Because all students were administered the background questionnaire, missing responses were considered to be missing-at-random (OECD, 2017). Furthermore, OLR DIF has been shown to be relatively robust to missing items (Finch, 2011). Therefore, omitted responses were treated as not present for each separate analysis (Choi et al., 2011).

Research Question 2

Overview

The purpose of the second research question in this study was to evaluate if motivation items showed DIF by gender or ethnicity after using anchoring vignettes to account for response scale use as the source of DIF. To answer Research Question 2, three main steps were taken. First, item characteristics and measurement assumptions of the vignettes were evaluated. Second, the nonparametric vignette scoring was applied to self-reported motivation items to account for response scale use. Finally, the same OLR DIF models from Research Question 1 were re-run using the vignette-adjusted item scores. Detailed procedures for these steps are described next.

Psychometric Assessment of Motivation Items

Vignette Assumption Testing and Item Analysis. To evaluate the psychometric properties of the vignettes, the three items were fit to a GRM and the same criteria for item fit from the

motivation items were applied to the vignettes. To assess the vignette measurement assumption of response consistency, the confidence intervals for the response category thresholds from the self-report items and the vignettes were examined for overlap (Weiss & Roberts, 2018); vignette equivalence was investigated by examining vignette ordering patterns. To deal with ties, this study used the lowest score from the vector range (OECD, 2014). Likewise, misordered vignettes were converted into ties at the highest category (e.g., $Z = \{1,4,2\}$ converted into $Z = \{1,4,4\}$) and then responses were treated according to procedures for ties (OECD, 2014). The *anchors* package (Wand et al., 2011) in R software was used to analyse ratings patterns for the vignettes and to rescore self-report responses.

Nonparametric Vignette Scoring of Self-Report Items

Because PISA specifically introduced anchoring vignettes to address the effects of response scale use on cross-group comparisons and because group differences in response scale use are present in motivation items (He & Van de Vijver, 2016a; Tan & Hall, 2005), the nonparametric method was selected as the vignette scoring approach (OECD, 2014). Relative to their individual ratings of Z_{ij}, students' responses to the self-assessment items, $\mathrm{X}_i \in \{1,2,3,4\}$, were recoded into the 7-point vignette-adjusted *C*-scale variable, $\mathrm{Y}_i \in \{1,2,3,4,5,6,7\}$ using the following formula (von Davier et al., 2018):

$$C = \begin{cases} \mathrm{Y}_i = 1 \text{ if } \mathrm{X}_i < \mathrm{Z}_1 \\ \mathrm{Y}_i = 2 \text{ if } \mathrm{X}_i = \mathrm{Z}_1 \\ \mathrm{Y}_i = 3 \text{ if } \mathrm{Z}_1 < \mathrm{X}_i < \mathrm{Z}_2 \\ \mathrm{Y}_i = 4 \text{ if } \mathrm{X}_i = \mathrm{Z}_2 \\ \mathrm{Y}_i = 5 \text{ if } \mathrm{Z}_2 < \mathrm{X}_i < \mathrm{Z}_3 \\ \mathrm{Y}_i = 6 \text{ if } \mathrm{X}_i = \mathrm{Z}_3 \\ \mathrm{Y}_i = 7 \text{ if } \mathrm{X}_i > \mathrm{Z}_3 \end{cases}$$

After new achievement motivation item scores were generated by applying the nonparametric formula, the vignette-adjusted item scores were used to re-run the same OLR DIF models

from Research Question 1. Then, to answer Research Question 2, unadjusted DIF results from Research Question 1 were compared to vignette-adjusted DIF results from Research Question 2 to identify changes to DIF outcomes and changes in the magnitude or form of DIF. Finally, to assess the effect of the vignettes on DIF, the difference in McFadden's pseudo R^2 between unadjusted and vignette-adjusted DIF was calculated as:

$$\Delta R^2 = R^2_{UnA} - R^2_A \tag{5}$$

To classify the effect of the vignettes on DIF, the same criteria used to classify DIF effect sizes were also applied to the vignette effects.

4 RESULTS

This chapter summarizes the results of analyses that tested for the presence of differential item functioning (DIF) and group differences in response scale use as a source of DIF across gender and ethnicity in the PISA 2015 achievement motivation items. As described in Chapter 3, this study involved three major steps. First, the five PISA achievement motivation items were analysed for DIF within a hybrid DIF framework that used a graded response model estimate of motivation as the matching criterion. DIF was detected within a 2-stage framework in which first omnibus DIF is tested, followed by testing for nonuniform and then uniform DIF. Gender DIF was evaluated with pairwise comparisons, and ethnicity DIF was evaluated using both multiple-group DIF methods with a base group as the reference group and all pairwise comparisons. Second, responses to self-reported achievement motivation items were adjusted with the nonparametric anchoring vignette scoring to account for response scale use as the source of DIF. Third, the same set of DIF analyses from the first step were re-run using the vignette-adjusted item responses.

Given the interrelated nature of the research questions in this study (i.e., Research Question 2 cannot be answered without Research Question 1), and that results indicated different response processes across groups, this chapter is organized as follows. First, the psychometric characteristics for both the achievement motivation items and the anchoring vignettes are presented for the full analytic sample and separately by group as necessary. Next, results are presented for gender DIF before and after using the vignettes to adjust self-report scores. Finally, results for the base group multiple-group ethnicity DIF are described, followed by findings from ethnicity pairwise comparisons.

Psychometric Analysis of Achievement Motivation Items

Descriptive Characteristics

Descriptive characteristics for the five PISA achievement motivation items were obtained from SPSS version 25.0 (IBM Corp., 2017) and can be seen in Table 4 (see Appendix C for the response distribution to items by group). These items were included in the PISA 2015 background questionnaire, which was administered to all students. Therefore, missing responses were considered to be missing-at-random and not due to PISA's matrix sampling. Because no item was missing more than 3.7% of responses, missing values were not imputed (He, Buchholz, et al., 2017). Rather, for each separate DIF analysis, missing responses were treated as not present and were excluded listwise (Choi et al., 2011). All analyses were run as unweighted, with higher scores interpreted as reflecting competitive-based, performance-approach motivation. Statistical significance was set to $p < .01$ for the entire study.

Across the five achievement motivation items, mean item scores were all above 3, which indicates that students had little difficulty endorsing these items and they reported generally high levels of achievement motivation. Items 1, 2, and 3 yielded higher group averages than Items 4 and 5. For gender, girls had higher mean scores than boys on the five items. For ethnicity, the lowest item score was from Hispanic/Latinx youth to Item 4 (M = 3.14), and the highest mean response was from Black/African American students to Item 2 (M = 3.72). Across ethnic groups, Item 4 had the largest range (range = 0.34), followed by Item 5 (range = 0.26) and Item 3 (range = 0.21). Black/African American students had the highest mean responses and smallest ranges; in contrast, Hispanic/Latinx students had the lowest mean responses and largest range of item responses. Cronbach's α was calculated using the *ltm* package (Rizopoulos, 2006) in R and was acceptable at .854 (95% CI = .846, .863).

Table 4

Descriptive Statistics for Achievement Motivation Items, by Group

Achievement Motivation Item		Gender (n = 5712)			Ethnic Group (n = 5654)					
		# RC	Male (50%)	Female (50%)	# RC	White, not Hispanic (43.7%)	Black/ African American (13.8%)	Hispanic/ Latinx (30.8%)	Asian (3.6%)	Multi-racial/ Other (6.9%))
1. I want top grades in most or all of my courses.	M	4	3.43 (0.66)	3.56 (0.59)	3	3.46 (0.64)	3.63 (0.55)	3.48 (0.64)	3.58 (0.62)	3.48 (0.63)
	% msg		2.8%	1.5%		1.2%	2.3%	1.3%	0.0%	1.3%
2. I want to be able to select from among the best opportunities available when I graduate.	M	4	3.56 (0.60)	3.68 (0.52)	2	3.60 (0.57)	3.72 (0.52)	3.61 (0.57)	3.69 (0.54)	3.59 (0.57)
	% msg		3.1%	1.7%		1.4%	2.8%	1.7%	0.0%	1.5%
3. I want to be the best, whatever I do.	M	4	3.50 (0.66)	3.53 (0.63)	3	3.48 (0.65)	3.69 (0.54)	3.48 (0.66)	3.51 (0.65)	3.55 (0.63)
	% msg		3.1%	2.0%		1.5%	2.9%	2.0%	0.0%	1.5%
4. I see myself as an ambitious person.	M	4	3.22 (0.71)	3.27 (0.72)	3	3.24 (0.7)	3.48 (0.64)	3.14 (0.76)	3.26 (0.76)	3.27 (0.66)
	% msg		3.7%	2.6%		1.8%	3.3%	3.0%	0.5%	2.0%
5. I want to be one of the best students in my class.	M	4	3.22 (0.77)	3.33 (0.72)	3	3.23 (0.76)	3.49 (0.65)	3.23 (.75)	3.44 (0.70)	3.26 (0.77)
	% msg		3.3%	2.0%		1.7%	3.0%	1.8%	1.0%	1.5%

Note. Standard deviations are presented in parentheses. # RC = number of response categories remaining after collapsing categories to maintain cell minimum of 5; msg = missing.

GRM Model Assumptions

Because the graded response model (GRM) was used to estimate students' level of achievement motivation, GRM assumptions of unidimensionality, local independence, and monotonicity were evaluated. Table 5 displays the psychometric characteristics for the five items for the whole analytic sample; group-specific parameters can be found in Appendix D. To evaluate unidimensionality, an exploratory factor analysis (EFA) of the five items was run in MPlus version 8.0 using weighted least squares mean and variance (Muthén & Muthén, 2017). The EFA yielded a first eigenvalue of 3.72, and the residual correlation matrix revealed no local dependence (i.e., no absolute residual correlations ≥ .20). Residual variance was .32 or less for four out of the five items. RMSEA values for each item yielded from the *mirt* package (Chalmers, 2012)

in R were between .028 and .032. Those values were above Finch's (2020) cutoff of .015 but close to the cutoff of .030 recommended by Rutkowski and Svetina (2014) for multiple groups and large samples. Taken together, the items were considered to have met essential unidimensionality (i.e., the primary factor is large enough that trait level estimates are unaffected by the presence of smaller factors; Hattie, 1985; Stout, 1990). Finally, increasing item thresholds indicated that the assumption of monotonicity was met for the items.

Table 5

Fit Statistics for Achievement Motivation Items

Motivation Item	RMSEA S - χ^2	Factor Loading	Residual Variance	a	b_1	b_2	b_3
Item 1	.028	.88*	.23	3.39	-2.74	-1.72	-0.20
Item 2	.028	.89*	.21	3.74	-2.72	-2.05	-0.47
Item 3	.027	.82*	.32	2.67	-2.94	-1.76	-0.29
Item 4	.032	.72*	.49	1.81	-3.04	-1.54	0.35
Item 5	.032	.84*	.29	2.94	-2.44	-1.23	0.16

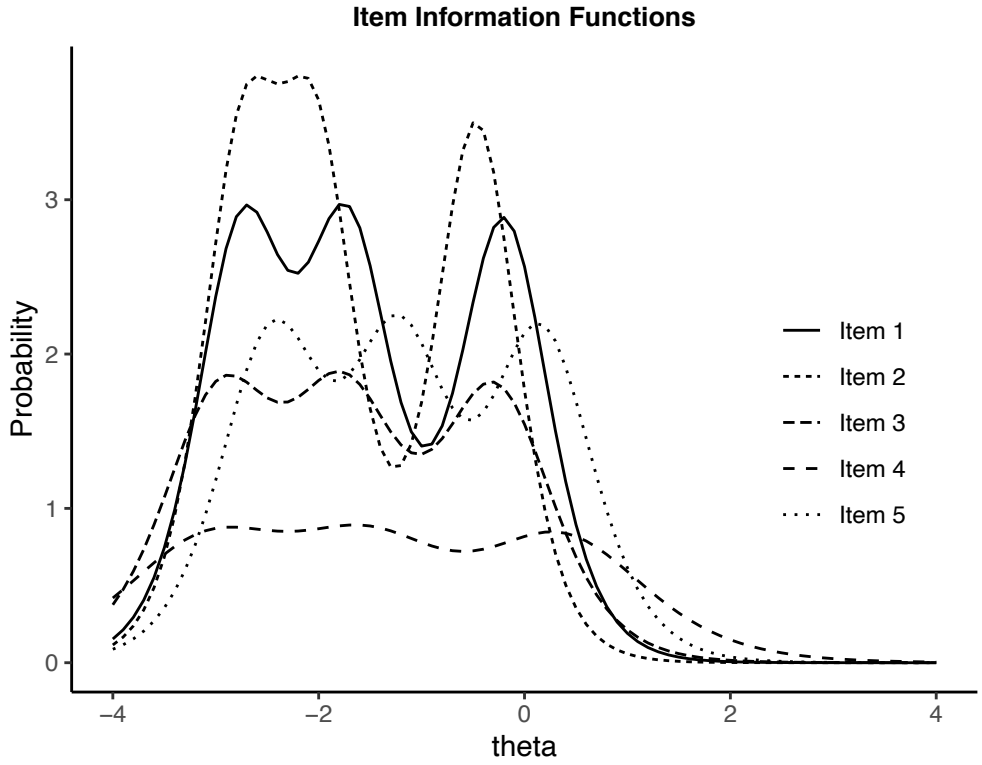
Note. a = item discrimination; b = response category threshold.
* $p < .01$.

Item Analysis

Table 5 shows that Item 2 had the highest discrimination and highest factor loading; Item 4 had the lowest discrimination and the lowest factor loading. Category thresholds fell below zero for Items 1, 2, and 3, and close to zero for Items 4 and 5, indicating that the items were "easy" (Hambleton et al., 1991). Item 2 had the highest discrimination, but that discrimination was limited to the narrowest range across the five items (range = 2.25). In contrast, Item 4 had the lowest discrimination but covered the widest range (range = 2.68). Moreover, ceiling effects evident in item information functions displayed in Figure 3 show that these items provided limited information about students with average or higher levels of achievement motivation.

Figure 3

Item Information Functions for PISA 2015 Achievement Motivation Items



Psychometric Analysis of Anchoring Vignettes

Descriptive Characteristics

The PISA 2015 anchoring vignettes can be seen in the left column of Figure 4, with descriptive statistics presented in Table 6. Cronbach's α for the three vignettes was -.342 (Bootstrap 95% CI = -.434, -.257), indicating a weak correlation between the vignettes or inconsistent responses to them by students (Vaske et al., 2017). PISA wrote the vignettes to be ordered from low to high such that vignette 1 reflected low motivation and vignette 3 reflected the most motivation. Although the mean scores for the vignettes followed the intended rank-order, Table 6 shows that for each group, the mean was close for vignettes 2 and 3 and their standard deviations

overlapped. Male and female students rated the vignettes similarly; therefore, the response distribution to the vignettes by gender are presented in Appendix E.

In contrast to gender ratings of the vignettes, Figure 4 shows that the vignette ratings by ethnic groups were varied. For vignette 1, White students had the lowest mean and Hispanic/Latinx students had the highest mean response. For vignette 2, Asian students had the lowest mean, while Black/African American and Multi-racial/Other had the highest means. For vignette 3, Black/African American students had the lowest average response, but their mean was relatively similar to Hispanic/Latinx, Asian, and Multi-racial/Other students. Across ethnic groups, the percentage of missing responses for Black/African American students was almost twice that of the other groups, with the exception of Asian students who had no missing responses to the vignettes.

Table 6

Descriptive Statistics for PISA 2015 Anchoring Vignettes, by Group

Vignette (Z_i)		Gender		Ethnic Group				
		Male	Female	White, not Hispanic	Black/ African American	Hispanic/ Latinx	Asian	Multi-racial/ Other
Low (Z_1)	M	1.35	1.30	1.26	1.37	1.39	1.29	1.34
		(0.63)	(0.57)	(0.54)	(0.66)	(0.65)	(0.59)	(0.61)
	% msg	3.5%	2.2%	1.6%	3.9%	2.1%	0.0%	2.3%
Medium (Z_2)	M	3.22	3.26	3.19	3.27	3.3	3.15	3.27
		(0.65)	(0.63)	(0.60)	(0.71)	(0.65)	(0.68)	(0.63)
	% msg	3.5%	2.0%	1.6%	3.8%	2.0%	0.0%	1.3%
High (Z_3)	M	3.76	3.81	3.84	3.72	3.73	3.78	3.79
		(0.54)	(0.48)	(0.44)	(0.59)	(0.57)	(0.48)	(0.47)
	% msg	4.0%	2.2%	2.0%	4.1%	2.5%	0.0%	2.3%

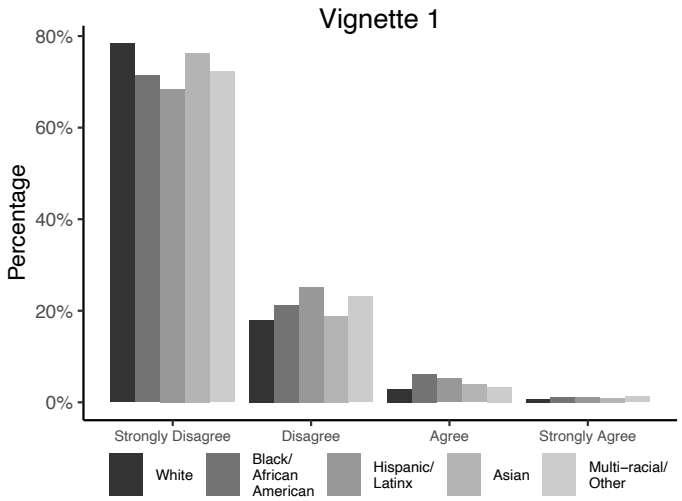
Note. Standard deviations are presented in parentheses; msg = missing.

Figure 4

Response Distribution of Vignette Ratings, by Ethnicity

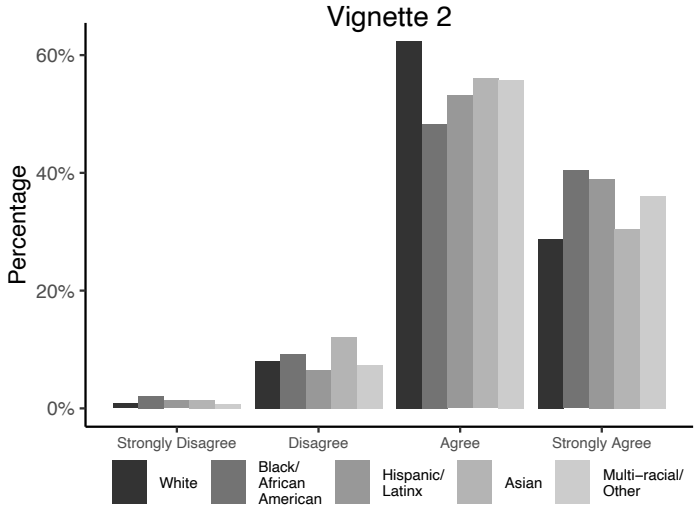
1. Mario gives up easily when confronted with a problem and is often not prepared for his classes.

   Mario is motivated. (Z_1)



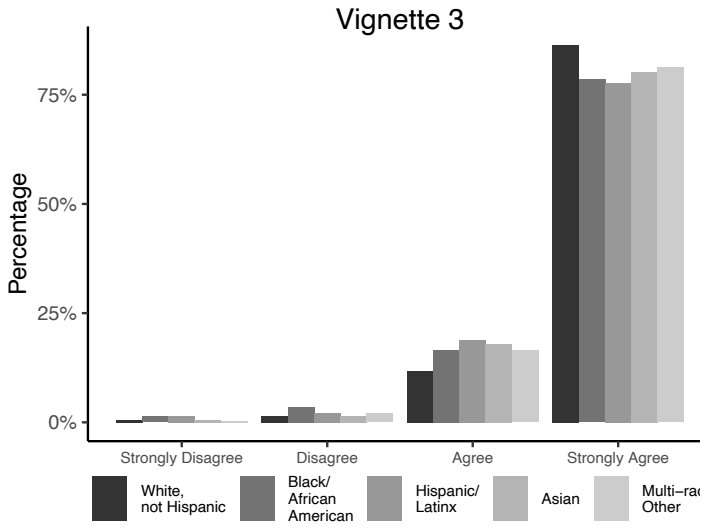
2. Olivia mostly remains interested in the tasks she starts and sometimes does more than what is expected from her.

   Olivia is motivated. (Z_2)



3. John wants to get top grades at school and continues working on tasks until everything is perfect.

   John is motivated. (Z_3)



Vignette Assumption Testing

Vignette Equivalence. Vignette equivalence was evaluated by examining tied ratings and vignette ordering patterns. Across the whole sample, Table 7 shows that only 52% of students rated the vignettes in the intended order, with 35% of students rating vignettes 2 and 3 as displaying equal levels of motivation and an additional 5% of students misordering them (e.g., $Z_3 > Z_2$). The presence of ties and misorderings shows insufficient discrimination between the two vignettes (He, Buchholz, et al., 2017). Though the percentage of tied ratings was high and discrimination was low, this was consistent with other studies of the PISA 2015 vignettes (e.g., Marksteiner et al., 2019). Figure 5 depicts vignette rating patterns by ethnicity for the three vignettes. As females and males were relatively consistent in their ordering of the vignettes, their ordering pattern can be found in Appendix F. By ethnic group, White students and Asian students had the highest percentage of students correctly order the vignettes. Black/African American students had the highest percentage of ties; within that group, more students tied vignettes 2 and 3 than correctly ordered the 3 vignettes (42% vs. 41%). Approximately 6% of students from the Hispanic/Latinx, Asian, and Multi-racial/Other groups misordered the vignettes.

King et al. (2004) noted that while participants do not have to follow the ranking order intended by the vignette authors, there does have to be a consensus order across respondents. In other words, respondents do not have to follow the intended ranking pattern, but all respondents do have to rank-order the vignettes the same way to meet vignette equivalence. Therefore, despite the ties and misorderings in this sample, as over 80% of students ordered the vignettes correctly, including ties (Marksteiner et al., 2019), vignette equivalence was assumed to have been met.

Table 7

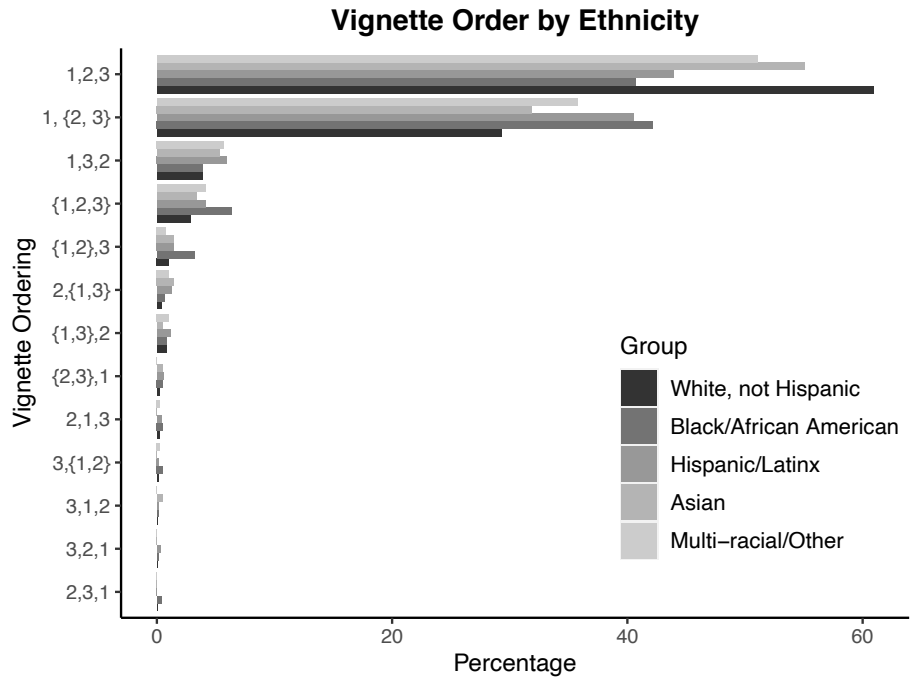
Ranking Patterns for PISA 2015 Anchoring Vignettes, All Students

Ordering Pattern	Frequency	Proportion	# of Distinct Rankings	# of Ranking Violations
1, 2, 3	2860	.52	3	0
1, {2, 3}	1937	.35	2	0
1, 3, 2	259	.05	3	1
{1, 2, 3}	213	.04	1	0
{1, 2}, 3	80	.01	2	0
{1, 3,}, 2	51	.01	2	1
2, {1, 3}	42	.01	2	1
{2, 3}, 1	21	.00	2	2
2, 1, 3	18	.00	3	1
3, {1, 2}	12	.00	2	2
3, 2, 1	7	.00	3	3
3, 1, 2	5	.00	3	2
2, 3, 1	4	.00	3	2

Note. Vignettes intended to be ordered as $Z_1 < Z_2 < Z_3$. {} = tied ratings; 1 = low vignette (Z_1), 2 = medium vignette (Z_2), 3 = high vignette (Z_3).

Figure 5

Vignette Ordering Pattern, by Ethnicity



Response Consistency. A GRM was fit with the three vignette items in the *mirt* package (Chalmers, 2012) in R software to obtain item fit statistics and item parameters, including response category thresholds. To assess response consistency, confidence intervals (CI) from GRM response category thresholds were examined for overlap across the self-report items and the vignettes. Because of the negative discrimination for vignette 1 and the extreme category thresholds for vignette 2 (see Table 8), category thresholds from vignette 3 were used to test response consistency. The 95% bootstrap CIs for item category thresholds for the five self-report items and vignette 3 overlapped such that response consistency was assumed to have been met.

Table 8

Item Parameters for PISA 2015 Anchoring Vignettes

Vignette (Z_i)	RMSEA S - χ^2	a	b_1	b_2	b_3
Z_1	.043	-1.90	-0.85	-2.26	-3.32
Z_2	.065	0.54	-8.55	-4.52	1.32
Z_3	.061	2.31	-3.07	-2.44	-1.15

Note. a = item discrimination; b = response category threshold.

Vignette Item Analysis

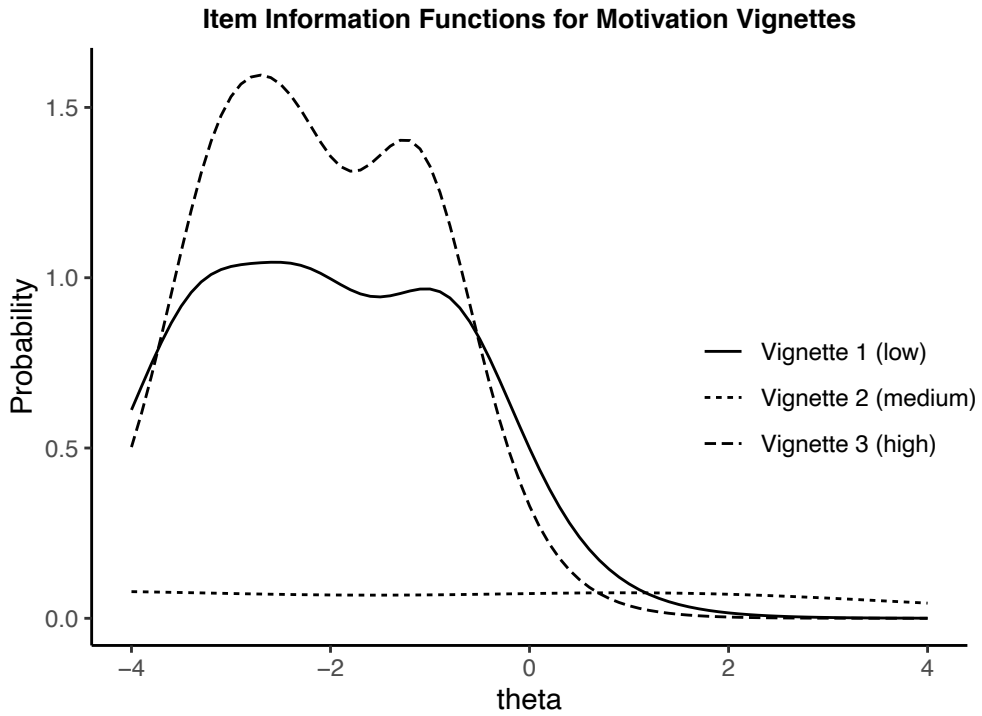
The RMSEA values associated with the S – χ^2 index from the vignette GRM ranged from .043 to.061, which were all above the recommended thresholds for item fit. Category thresholds for vignette 2 seen in Table 8 stand out as covering an unusually wide range of θ (from -8.55 to +1.32), and discrimination for vignette 1 was negative (a = -1.89). Although typically a negative a parameter in a GRM could warrant caution as that shows the opposite of monotonicity (i.e., the probability of an item response *decreases* as the level of the construct increases), vignette 1 was specifically written to elicit high levels of disagreement from students. Moreover, for the low level of motivation depicted in vignette 1, students with higher levels of

motivation may have endorsed strongly disagree more than students with lower levels of motivation, because higher motivation students may have different internal standards for what they view as high and low levels of motivation. In other words, students with low self-reported motivation might not disagree with the low-level vignette as strongly as their highly motivated peers because lower motivation students would not necessarily view that vignette character as having low motivation.

Item information functions for the three vignettes, seen in Figure 6, show that overall, the vignettes provide limited information for students with average or higher motivation levels (i.e., $\theta \geq 0$). While information functions for vignettes 1 and 3 are shifted to the left, they show some discrimination at the lower end of the θ range (i.e., $\theta < 0$). In contrast, the item information function for vignette 2 is nearly flat and indicates an absence of discrimination and information.

Figure 6

Item Information Functions for PISA 2015 Anchoring Vignettes



Given the poor psychometric characteristics for vignette 2 (i.e., unusual category thresholds and lack of discrimination), vignette 2 was removed from any further analyses, with only vignettes 1 and 3 being used to adjust the self-reported achievement motivation items for the DIF analyses. In the case of two vignettes, the self-report items are adjusted according to the following *2J + 1* formula, where *J* is the number of vignettes (King & Wand, 2007):

$$C = \begin{cases} Y_i = 1 \text{ if } X_i < Z_1 \\ Y_i = 2 \text{ if } X_i = Z_1 \\ Y_i = 3 \text{ if } Z_1 < X_i < Z_2 \\ Y_i = 4 \text{ if } X_i = Z_2 \\ Y_i = 5 \text{ if } X_i > Z_2 \end{cases}$$

After rescoring self-report items with the vignettes, the Cronbach's α for the five vignette-adjusted self-report items was .887 (Bootstrap 95% CI = .878, .895); prior to the vignette adjustment, α for the motivation items was .854. Figures 7 and 8 shows the distribution of motivation item responses after the self-report items were nonparametrically adjusted with the vignettes. For gender (Figure 7), girls had a higher percentage of self-report responses rescaled to a 4, indicating that more girls rated themselves as equal to Z_2 than boys. In every other response category, except for category 3 on Item 2, boys had a higher percentage of adjusted responses in a given category. For response category 5, this meant that a higher percentage of boys rated themselves as higher than Z_2 than did girls, though the difference was small for Items 1, 2, and 5. Overall, however, the distribution of vignette-adjusted scores was relatively similar between male and female students. Regarding ethnicity (Figure 8), Black/African American students had the highest percentage of responses rescored to a 4 or 5, followed by Asian students. Hispanic/Latinx students had the highest percentage of vignette-adjusted responses clustered at the low end of the scale.

Figure 7

Vignette-Adjusted Responses to Achievement Motivation Items, by Gender

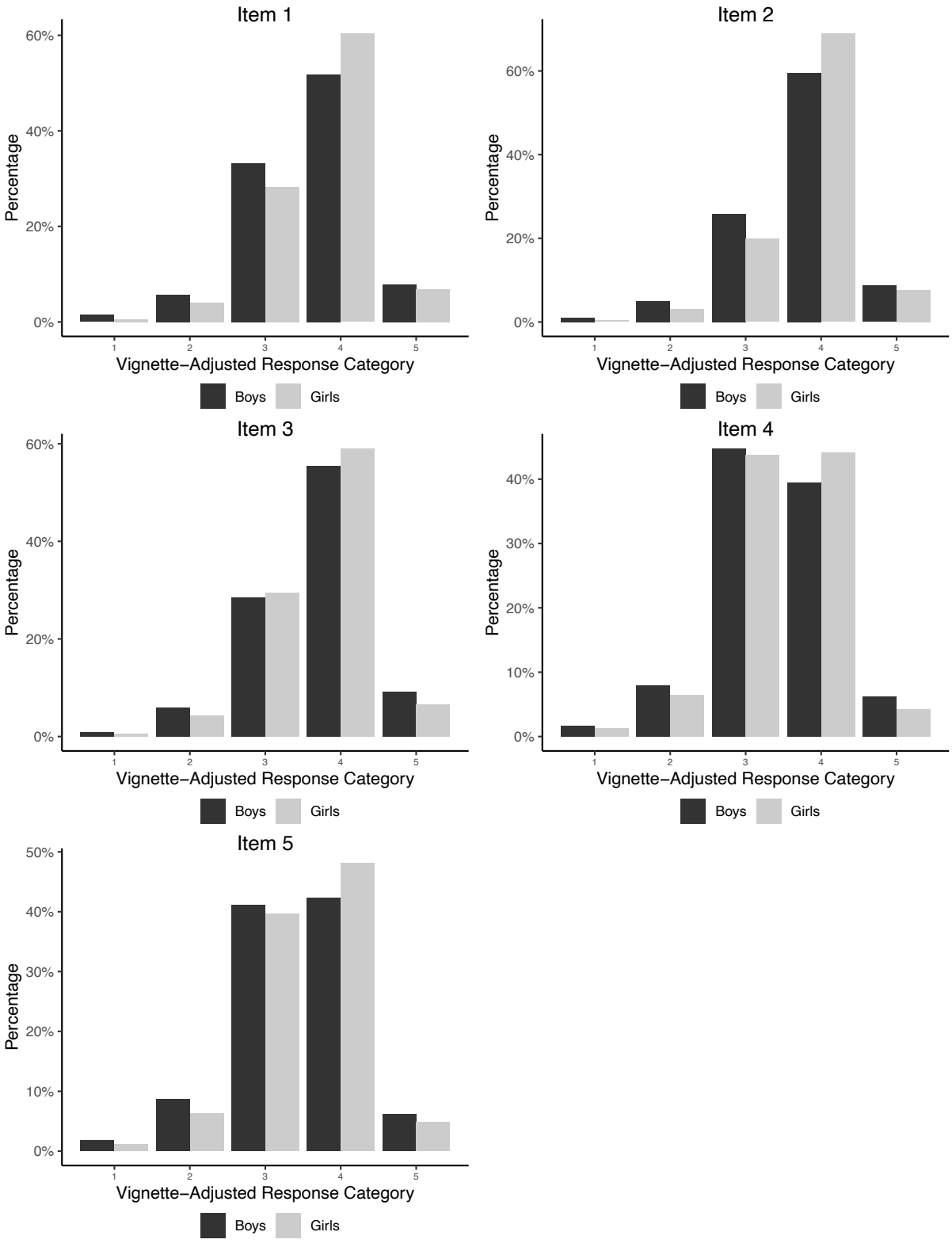
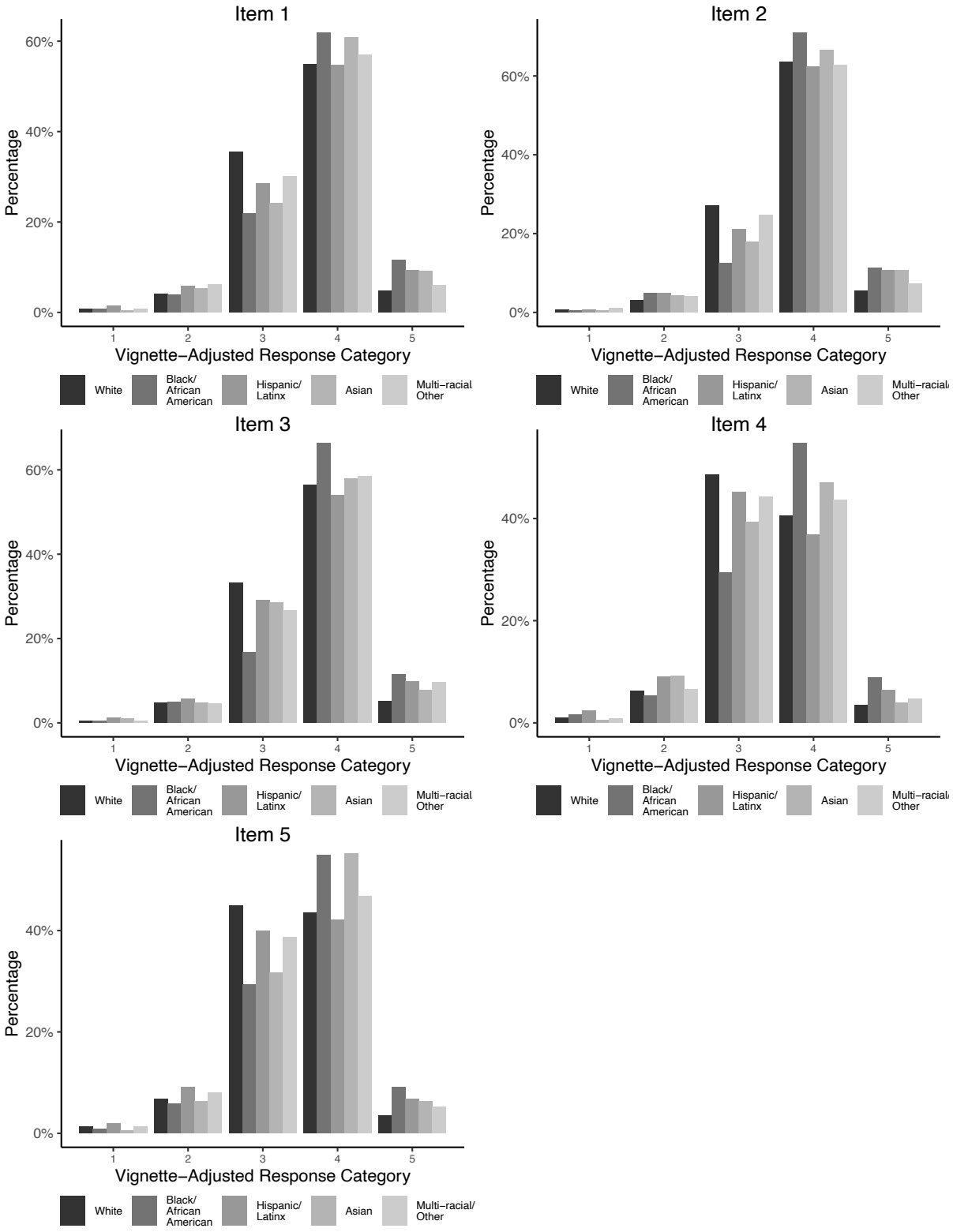


Figure 8

Vignette-Adjusted Responses to Achievement Motivation Items, by Ethnicity



DIF Results

Given the interrelated nature of the research questions in this study and because results indicated different response processes between gender and ethnicity, first, an overall summary of DIF results is presented. Next, results are described for gender DIF before and after using the vignettes to adjust self-report scores. Finally, results for the base group multiple-group ethnicity DIF are described, followed by findings from ethnicity pairwise comparisons.

Figures 9 through 14 display item response functions (IRF) and category response functions (CRF) for items flagged for gender DIF and the multiple-group DIF before and after the vignette adjustments. For each item, the top panel (panel 1) displays DIF before items were rescaled with the vignettes; the bottom panel (panel 2) displays DIF after items were rescaled. The IRFs (plot A) display the item true score response function for each group based on group-specific item parameters (Choi et al., 2011). The IRF figures also display the impact-weighted density line, which reflects the absolute summed difference between IRFs, weighted by the focal group (Choi et al., 2011). The CRFs and response category thresholds (plot B) were examined for the pattern and location of DIF in order to provide insight into sources of DIF and the effects of response scale use on DIF (Penfield et al., 2008).

Summary of DIF Results

Table 9 presents a summary of items flagged for DIF before and after using vignettes 1 and 3 to adjust the self-reported achievement motivation items to account for the effect of group differences in response scale use as a source of DIF. Overall, gender DIF was found before and after vignette adjustments in Items 1 and 2. Item 5 was flagged for gender DIF before the vignette adjustment, but not after; Item 3 demonstrated the opposite gender DIF pattern. In the

base group multiple-group ethnic DIF, Item 4 was flagged for DIF before and after being adjusted; Item 2 was only flagged after the vignette adjustment. DIF patterns also changed across some pairwise comparisons after the vignette rescoring, including that DIF was no longer flagged in base group comparisons for either Black/African American students or Asian students.

Table 9

Summary of Items Flagged for DIF, Before and After Vignette Adjustments

DIF Grouping	Unadjusted	Vignette-Adjusted
Gender	1, 2, 5	1, 2, 3
Multiple-Group Ethnic DIF	4	2, 4
Pairwise Comparison		
Black/African American vs. BG	4	-
Hispanic/Latinx vs. BG	4	4
Asian vs. BG	4	-
White vs. Black/African American	3, 4	3, 4
White vs. Hispanic/Latinx	4	4
White vs. Asian	4	-
Black/African American vs. Hispanic/Latinx	2, 4	2, 4
Black/African American vs. Asian	3, 4	4

Note. BG = base group.

Gender DIF Results

Table 10 presents results from gender DIF analyses before and after the vignette adjustment of self-report items. Per a two-stage DIF detection framework, items were first evaluated for "total" omnibus DIF effects, followed by nonuniform DIF and uniform DIF. Before rescoring items using the vignettes, omnibus DIF ($df = 2$) was flagged in Items 1 ("top grades"), 2 ("best opportunities"), and 5 ("best student"). Nonparallel IRFs between male and female students in Figure 10.1a visually confirmed the presence of nonuniform DIF in Item 2. That is, when matched by motivation level, there was an interaction between motivation and gender that varied by level of motivation. Figure 9.1a shows uniform DIF in Item 1, and Figure 12.1a shows uniform DIF in Item 5. For those items, when genders were matched by motivation level, female

students had an easier time endorsing higher response categories than their male peers (i.e., items favored girls).

The CRFs for Item 1 (Figure 9.1b), Item 2 (Figure 10.1b), and Item 5 (Figure 12.1b) show that the difference in category thresholds between genders got successively smaller as the level of motivation increased, with the largest difference occurring at the first category threshold. For Item 1, the $|\Delta b_{R-F}|$ between response category thresholds for males and females at the lowest category was 0.35 and 0.26 at the second-lowest response category. For Item 2, $|\Delta b|$ for the three category thresholds was 0.50, 0.34, and 0.25. For Item 5, the $|\Delta b|$ did not exceed 0.21. The convergent and pervasive pattern of differences in group thresholds, particularly for Items 1 and 2, point to a source of DIF at the item level. However, impact-weighted density lines, displayed in Figures 9.1a, 10.1a, and 12.1a, show that few students fell at the levels of motivation where the largest group differences in thresholds were located. This minimal impact is reflected in McFadden's pseudo R^2 effect sizes that all fell below .018; similarly, β_1 did not exceed the threshold of 5% for any of the DIF items. Therefore, all DIF was classified as negligible.

After using the vignettes to rescore motivation items, results displayed in Table 10 show that omnibus gender DIF was flagged again in Items 1 ("top grades") and 2 ("best opportunities"). Item 1 shifted from uniform to nonuniform DIF, and the nonuniform DIF in Item 2 became more pronounced. In contrast, after rescaling, Item 5 ("best student") was no longer flagged for DIF, but Item 3 ("be the best") was flagged for uniform DIF. As Item 3 had not been flagged for DIF prior to the vignette adjustment and because the DIF effect sizes for both Items 3 and 5 were negligible, the shift in DIF from Item 5 to Item 3 may have been due to pseudo-DIF. Consistent with the purpose of the vignettes to "stretch" the response scale, regardless of DIF

status, Figures 9.2a, 10.2a, 11.2a, and 12.2a all show that item discrimination increased for students with θ = 2 or higher motivation as compared to before the adjustment.

Vignette-adjusted CRFs for Item 1 (Figure 9.2b), Item 2 (Figure 10.2b), and Item 3 (Figure 11.2b) revealed a shift in response threshold patterns across the three items from favoring female students at low levels of motivation to favoring males at high levels of motivation. The distance between middle category thresholds also decreased after the rescoring, but DIF remained at the lowest score levels. In Item 1, the $|\Delta b_{R-F}|$ at the lowest category was 0.38; for Item 2, the lowest category difference was 0.50. This divergent, non-pervasive pattern indicates that the source of DIF was specific to a given score level. As the shift occurred at the highest response category in all three items, it appears that it was easier for boys to transition to the highest response category than for girls. Despite these changes in DIF patterns, however, pseudo R^2 for vignette-adjusted DIF was negligible.

Table 11 shows the difference in pseudo R^2 values between unadjusted and vignette-adjusted DIF (i.e., $\Delta R^2 = R^2_{UnA} - R^2_A$). Omnibus DIF was reduced in Item 1 by almost 70% and was reduced in Item 2 by 56%. Uniform DIF in Item 1 decreased by .010 (77%) and decreased in Item 2 by .011 (65%). However, gender DIF effect sizes were negligible to be begin with, which left little room for the vignettes to have a large effect on DIF. Nonetheless, according to the R^2 effect size criteria used in this study, the effect of the vignette correction on gender DIF was negligible. Moreover, although adjusting for group-specific response scale use with the vignettes changed DIF patterns in four items, negligible effect sizes before and after the vignette adjustment indicated measurement scale comparability in these motivation items for gender.

Table 10

Gender DIF, Before and After Vignette Adjustments

Item	Model 1:3 (Omnibus DIF)				Model 2:3 (Nonuniform DIF)				Model 1:2 (Uniform DIF)					
	UnA		A		UnA		A		UnA			A		
	p	R^2	p	R^2	p	R^2	p	R^2	p	R^2	β_1	p	R^2	β_1
1	<.001	.013	<.001	.004	.452	.000	.004	.001	<.001	.013	2.57%	<.001	.003	0.68%
2	<.001	.018	<.001	.008	.003	.001	<.001	.002	<.001	.017	3.96%	<.001	.006	1.41%
3	.968	.000	<.001	.001	.971	.000	.026	.000	.801	.000	0.02%	.003	.001	0.31%
4	.188	.000	.378	.000	.183	.000	.542	.000	.210	.000	0.06%	.210	.000	0.08%
5	<.001	.003	.151	.000	.518	.000	.691	.000	<.001	.003	0.75%	.057	.000	0.02%

Note. Boys were coded as the reference group in gender DIF comparisons. UnA = unadjusted; A = vignette-adjusted; p = p-value associated with the difference in likelihood ratio χ^2 tests between nested models; β_1 = percentage change in β_1 from GM1 to GM2, with > 5% difference being slight to moderate DIF; R^2 = difference in McFadden's pseudo R^2 between nested models. DIF classification categories: negligible DIF if $R^2 < .035$, moderate DIF if $.035 \leq R^2 \leq .070$, and large DIF if $R^2 > .070$.

Table 11

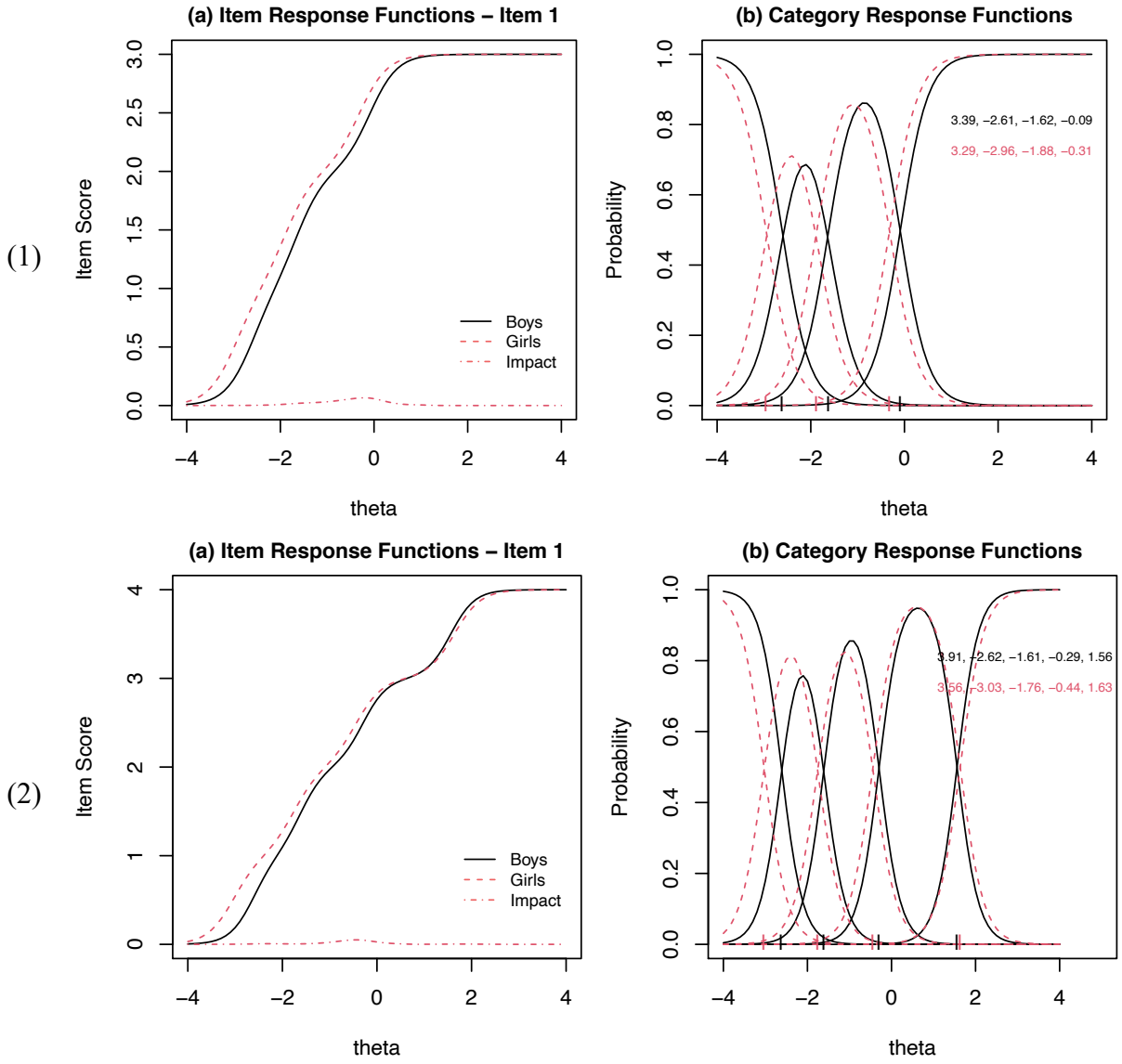
Changes in Gender DIF After Vignette Adjustments

Achievement Motivation Item	Model 1:3 (Omnibus DIF)	Model 2:3 (Nonuniform DIF)	Model 1:2 (Uniform DIF)
	ΔR^2	ΔR^2	ΔR^2
1. I want top grades in most or all of my courses.*‡	-.009	+.001	-.010
2. I want to be able to select from among the best opportunities available when I graduate.*‡	-.010	+.001	-.011
3. I want to be the best, whatever I do.*	+.001	.000	+.001
4. I see myself as an ambitious person.	.000	.000	.000
5. I want to be one of the best students in my class.‡	-.003	.000	-.003

Note. NU = Nonuniform DIF; * = item flagged for DIF before being adjusted at $p < .01$; ‡ = item flagged for DIF after vignette adjustment at $p < .01$; ΔR^2 = change in McFaddden's pseudo R^2 between unadjusted and vignette-adjusted responses; - indicates a decrease in value; + indicates an increase in value.

Figure 9

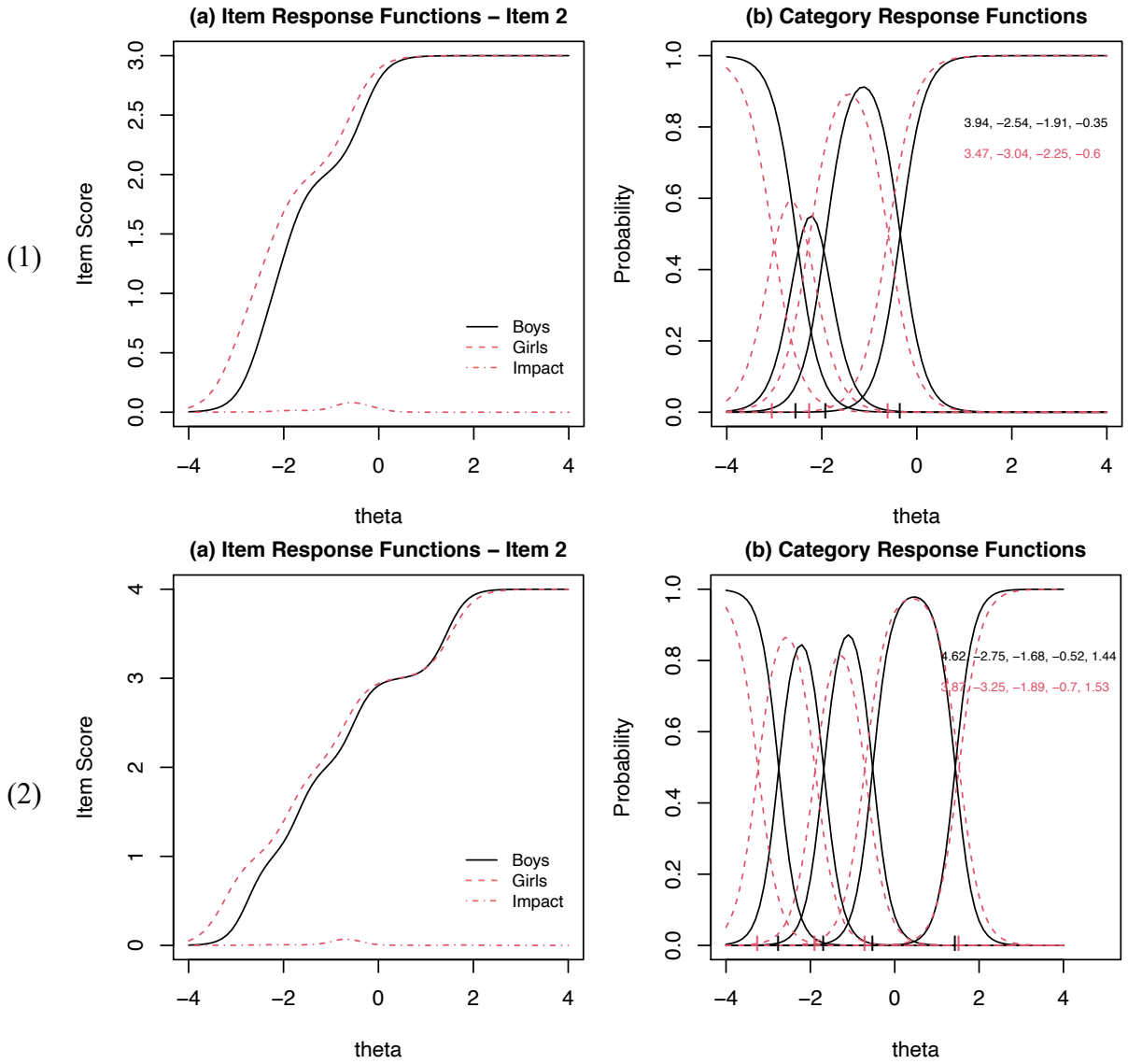
Gender DIF – Item 1



Note. Panel 1 displays DIF before applying the vignette adjustment; Panel 2 displays DIF after applying the vignettes. Item response functions (IRF) reflect the item true score function; impact = absolute summed difference in IRFs between groups, weighted by girls. The hash marks above the x-axis in category response function (CRF) plots are the category boundaries, with values for group-specific parameters displayed above ($a, b_1 \ldots b_k$).

Figure 10

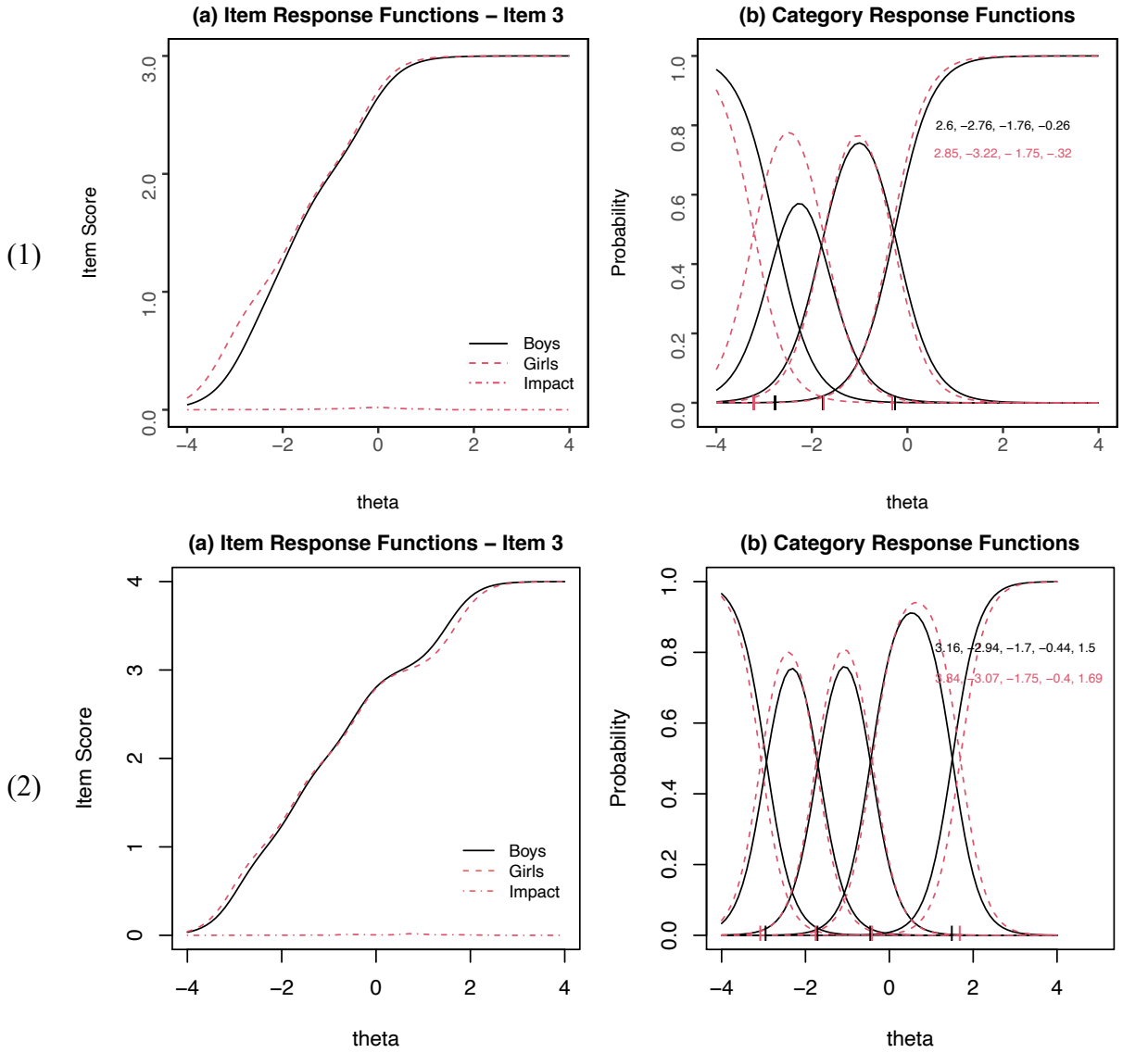
Gender DIF – Item 2



Note. Panel 1 displays DIF before applying the vignette adjustment; Panel 2 displays DIF after applying the vignettes. Item response functions (IRF) reflect the item true score function; impact = absolute summed difference in IRFs between groups, weighted by girls. The hash marks above the x-axis in category response function (CRF) plots are the category boundaries, with values for group-specific parameters displayed above ($a, b_1 \ldots b_k$).

Figure 11

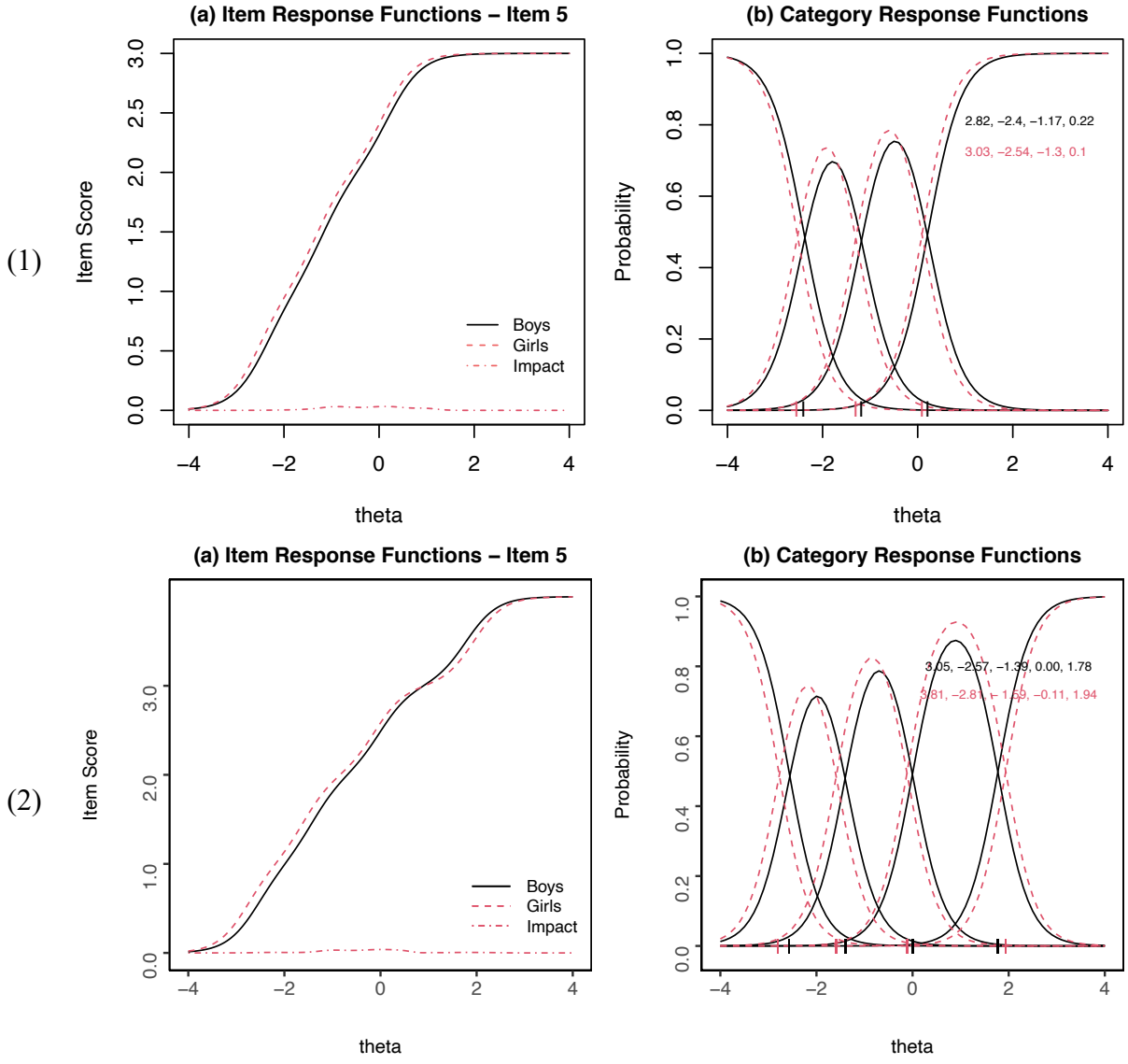
Gender DIF – Item 3



Note. Panel 1 displays DIF before applying the vignette adjustment; Panel 2 displays DIF after applying the vignettes. Item response functions (IRF) reflect the item true score function; impact = absolute summed difference in IRFs between groups, weighted by girls. The hash marks above the x-axis in category response function (CRF) plots are the category boundaries, with values for group-specific parameters displayed above ($a, b_1 \ldots b_k$).

Figure 12

Gender DIF – Item 5



Note. Panel 1 displays DIF before applying the vignette adjustment; Panel 2 displays DIF after applying the vignettes. Item response functions (IRF) reflect the item true score function; impact = absolute summed difference in IRFs between groups, weighted by girls. The hash marks above the x-axis in category response function (CRF) plots are the category boundaries, with values for group-specific parameters displayed above ($a, b_1 \ldots b_k$).

Ethnicity DIF Results

Multiple-Group DIF (MG-DIF). Results are presented in Table 12 for the MG-DIF with a base group as the reference group before and after the vignette adjustment. Table 13 shows changes in MG-DIF after vignette adjustments, and Table 14 displays results for significant base group pairwise comparisons. Before being adjusted with the vignettes, Item 4 ("ambitious") was flagged in the MG-DIF. DIF was also present in the comparisons between the base group and Black/African American students, Hispanic/Latinx students, and Asian students. DIF in the base group comparison indicates that the focal group's response scale differs in relation to the average response scale (i.e., item response function) for the item. At the bottom of the response scale, Asian and Hispanic/Latinx students fell below the average and demonstrated a need for more achievement motivation than the average student to shift from the lowest response category to the adjacent category (see Figure 14.1a). At the same location on the response scale, Black/African American students demonstrated the opposite pattern, showing that it was easier for them in comparison to the average to make the same category shift. In contrast, around $\theta = 1$, Figure 14.1a shows how the item shifted to favoring Asian students.

The category threshold pattern for the MG-DIF in Item 4 was divergent and non-pervasive, indicating a cause of DIF specific to those score levels. The largest difference in category thresholds fell at the first category between the base group and Asian students (|-1.61-1.06| = 0.55). At the second category threshold, the largest difference was between the base group and Black/African American students (0.33-0.07 = 0.26). However, both McFadden's pseudo R^2 and $\Delta\beta_1$ for Item 4 were very small and classified the MG-DIF as negligible, with impact-weighted density lines showing that few students fell at those levels of motivation.

After motivation items were adjusted with the vignettes, Item 2 (Figure 13) and Item 4 evidenced DIF in the base group MG-DIF (Table 12). For Item 2 ("best opportunities"), whereas Asian students needed less motivation than the average student before shifting from the lowest response category to the next higher one, Black/African American students needed more motivation than the average student to make the same category shift. This is the opposite pattern displayed by these groups in Item 4 before being adjusted by the vignettes. At the first, second, and third category thresholds in Item 2, the largest difference was between the base group and Black/African American students, followed by Asian and then Hispanic/Latinx students. Moreover, in contrast to the hypothesized direction of the relationship, $\Delta\beta_1$ for Item 2 actually *increased*, from 0.36% before vignette adjustments to 0.54% after vignette adjustments. The pattern of threshold differences remained divergent and pervasive, and DIF effects were classified as negligible.

For Item 4, the vignette adjustment decreased β_1 by 0.11% (Table 12). DIF was no longer detected in the comparison between the base group and Black/African American and Asian students; however, DIF remained between the base group and Hispanic/Latinx students (Table 14). Thus, adjusting for group differences in response scale use changed DIF patterns in relation to the average (reflected in the base group) for two out of three groups. One way to interpret this finding is that Black/African American and Asian students demonstrated response scale tendencies (i.e., demonstrated response styles). Nonetheless, Table 13 shows that the overall effect of the vignette correction on MG-DIF was negligible. Therefore, although correcting for group differences in response scale use with the vignettes changed DIF patterns in two items, negligible effect sizes indicated measurement scale comparability in relation to the average.

Table 12

Multiple-Group Ethnicity DIF, Before and After Vignette Adjustments

Item	Model 1:3 (Omnibus DIF)				Model 2:3 (Nonuniform DIF)				Model 1:2 (Uniform DIF)					
	UnA		A		UnA		A		UnA			A		
	p	R^2	p	R^2	p	R^2	p	R^2	p	R^2	β_1	p	R^2	β_1
1	.758	.001	.703	.001	.787	.000	.688	.000	.519	.000	0.34%	.527	.000	0.17%
2	.084	.002	.002	.003	.084	.001	.132	.001	.232	.001	0.37%	.001	.002	0.53%
3	.119	.002	.051	.002	.744	.000	.174	.001	.027	.001	0.10%	.061	.001	0.06%
4	<.001	.007	<.001	.004	.007	.002	.068	.001	<.001	.006	0.24%	<.001	.003	0.11%
5	.354	.001	.154	.001	.767	.002	.154	.001	.131	.001	0.06%	.271	.001	0.05%

Note. In the MG-DIF, the base group served as the reference group against which the focal ethnic groups were compared. UnA = unadjusted; A = vignette-adjusted; p = p-value associated with the difference in likelihood ratio χ^2 tests between nested models; β_1 = percentage change in β_1 from EM1 to EM2, with > 5% being slight to moderate DIF; R^2 = difference in McFadden's pseudo R^2 between nested models. DIF classification categories: negligible DIF if R^2 <.035, moderate DIF if .035 ≤ R^2 ≤ .070, and large DIF if R^2 > .070.

Table 13

Changes in Multiple-Group Ethnicity DIF After Vignette Adjustments

Achievement Motivation Item	Model 1:3 (Omnibus DIF)	Model 2: 3 (Nonuniform DIF)	Model 1:2 (Uniform DIF)
	ΔR^2	ΔR^2	ΔR^2
1. I want top grades in most or all of my courses.	.000	.000	.000
2. I want to be able to select from among the best opportunities available when I graduate.	+.001	.000	+.001
3. I want to be the best, whatever I do.	.000	+.001	.000
4. I see myself as an ambitious person.	-.003	-.001	-.003
5. I want to be one of the best students in my class.	.000	-.001	.000

Note. NU = Nonuniform DIF; * = item flagged for DIF before being adjusted; ‡ = item flagged for DIF after vignette-adjustment; ΔR^2 = change in McFaddden's pseudo R^2 between unadjusted and vignette-adjusted DIF; - indicates a decrease in value; + indicates an increase in value.

Table 14

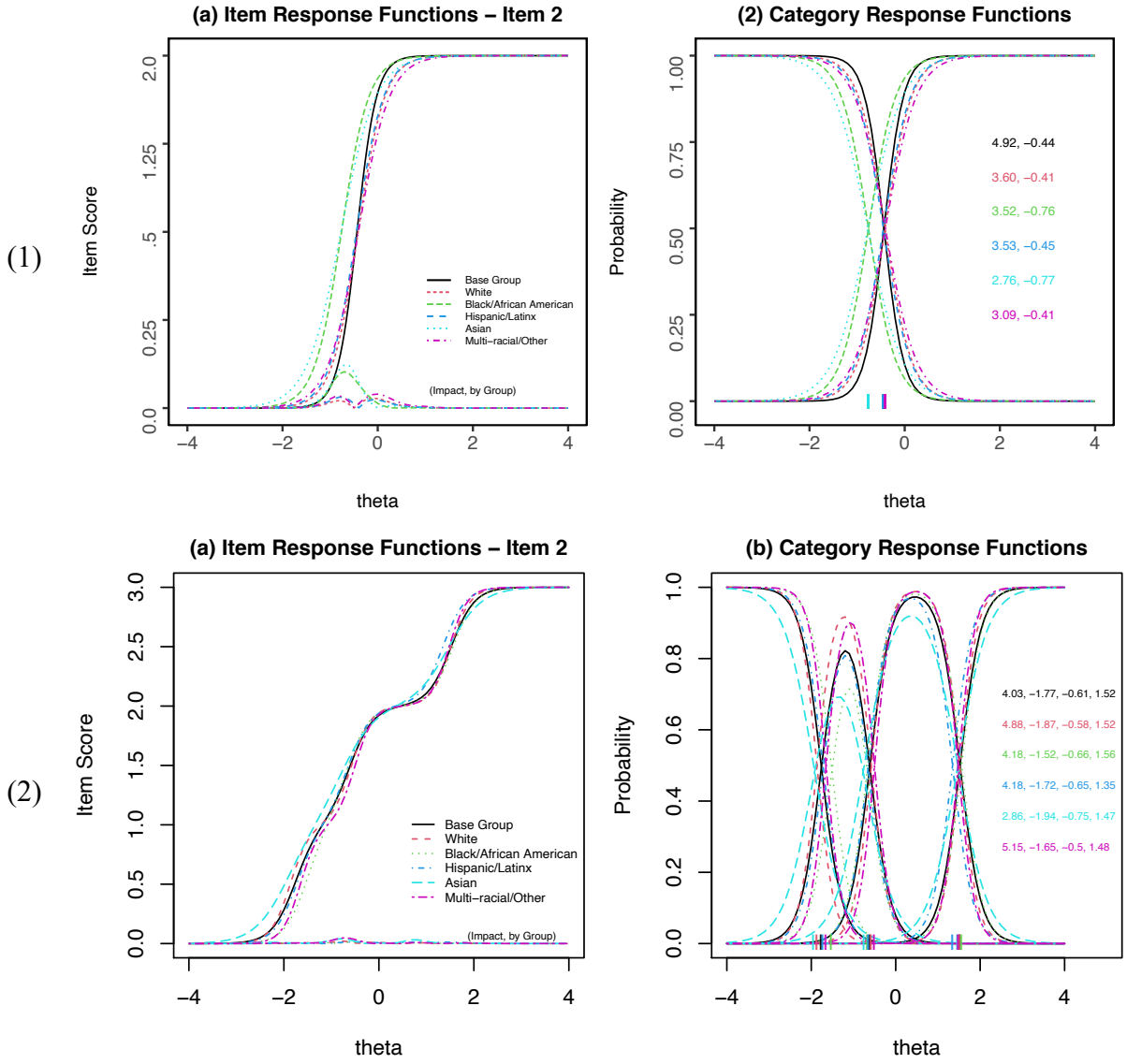
Significant Ethnicity Comparisons, Before and After Vignette Adjustments

DIF Comparison	Unadjusted						Adjusted					
		NU DIF		Uniform DIF				NU DIF		Uniform DIF		
	Item	p	R^2	p	R^2	β_1	Item	p	R^2	p	R^2	β_1
BG vs. Black/African American[+]	4			.001	.003	0.76%						
BG[+] vs. Hispanic/Latinx	4			<.001	.004	0.39%				.001	.002	0.38%
BG vs. Asian	4	.001	.004									
White vs. Black/African	3			<.001	.004	0.40%	3	.002	.002			
American[^+]	4			<.001	.004	0.78%	4	.001	.002			
White[+] vs. Hispanic/Latinx	4			<.001	.004	0.88%	4			<.001	.002	0.70%
White vs. Asian	4	<.001	.003									
Black/African American[+]	2			.004	.003	1.23%	2			.004	.002	0.74%
vs. Hispanic/Latinx[‡]	4			<.001	.011	1.05%	4			<.001	.005	0.29%
Black/African American[^+]	3			<.001	.011	0.87%						
vs. Asian	4			<.001	.014	1.94%	4			.002	.006	0.29%

Note. Group listed first in comparison was the reference group in the DIF comparison; NU = Nonuniform; BG = base group; [‡] = group favored in Item 2 uniform DIF; [^] = group favored in Item 3 uniform DIF; [+] = group favored in Item 4 uniform DIF.

Figure 13

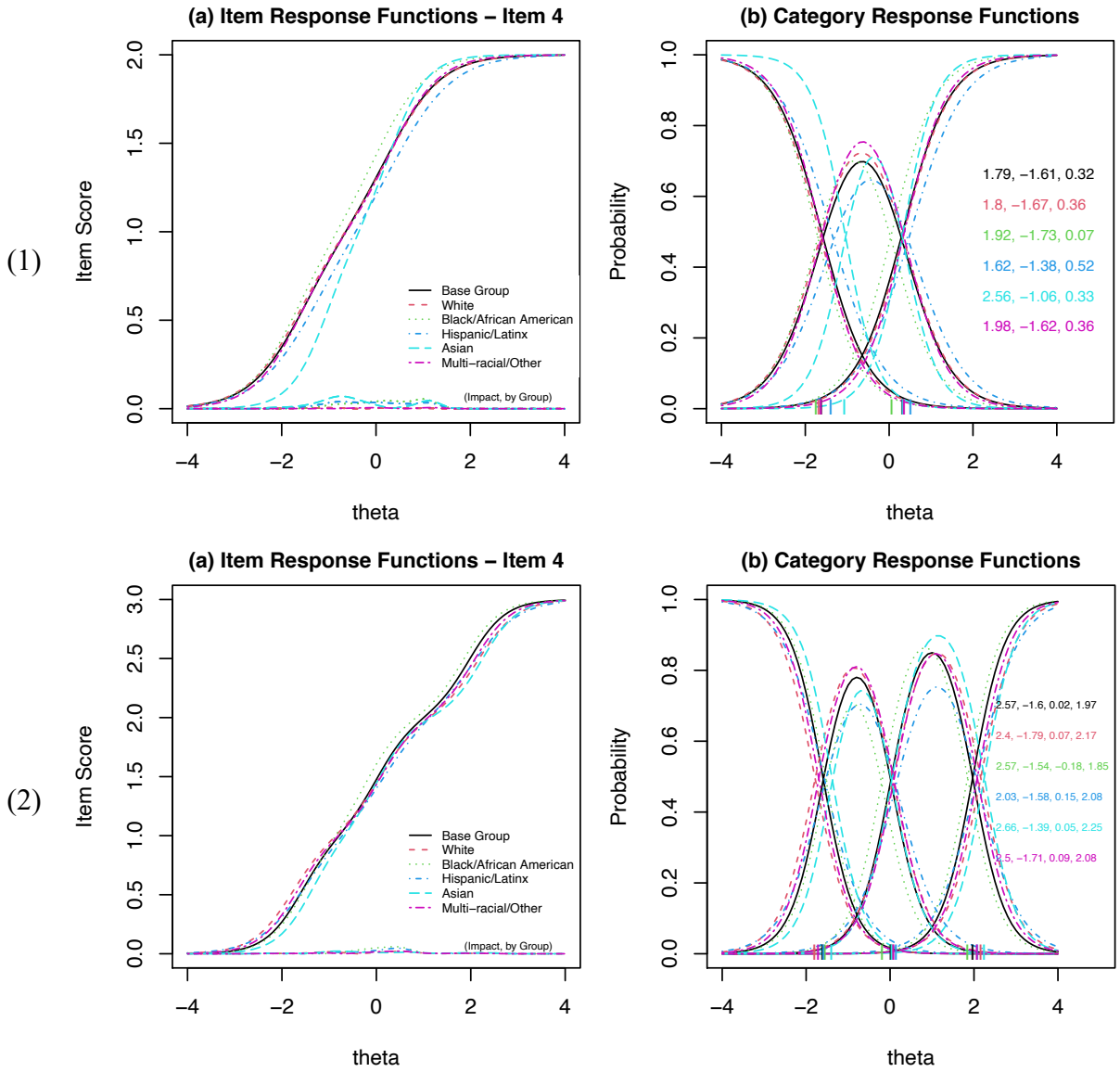
Multiple-Group Ethnicity DIF – Item 2



Note. Panel 1 displays DIF before applying the vignette adjustment; Panel 2 displays DIF after applying the vignettes. Item response functions (IRF) reflect the item true score function; impact = absolute summed difference in IRFs between the base group and focal group, weighted by focal group. The hash marks above the x-axis in category response function (CRF) plots are the category boundaries, with values for group-specific parameters displayed above ($a, b_1 \ldots b_k$).

Figure 14

Multiple-Group Ethnicity DIF – Item 4



Note. Panel 1 displays DIF before applying the vignette adjustment; Panel 2 displays DIF after applying the vignettes. Item response functions (IRF) reflect the item true score function; impact = absolute summed difference in IRFs between the base group and focal group, weighted by focal group. The hash marks above the x-axis in category response function (CRF) plots are the category boundaries, with values for group-specific parameters displayed above ($a, b_1 \ldots b_k$).

Pairwise Comparisons. Table 14 displays results from significant ethnicity DIF pairwise comparisons. Before rescoring items with the vignettes, Item 2 favored Hispanic/Latinx students in comparison to Black/African American students; however, the DIF was located at very low levels of θ (i.e., θ = -2 to θ = -4), and IRFs showed that there was no difference between groups starting at θ = -2 and higher. Item 3 favored Black/African American students in all pairwise comparisons. For Item 4, uniform DIF favored White students in the comparisons with Hispanic/Latinx students; uniform DIF in Item 4 also favored Black/African American students in comparisons with White students and Hispanic/Latinx students. On Item 4, DIF between White and Asian students was nonuniform, with the item favoring White students until approximately θ = 1, at which point there was a shift to Asian students being favored by the item.

For Black/African American students, the pervasive, constant DIF pattern seen across items indicated that the source of DIF was likely systematic at the item level. In contrast, the divergent pattern of category thresholds seen in comparisons with Asian students revealed that the source of DIF was likely local to the score level. Across the pairwise comparisons, most of the DIF was located between θ = -2 and θ = 0. Given the ease of these items, however, impact-weighted density lines showed that few students fell at those levels of motivation as effect sizes ranged from .002 to .014, and no $\Delta\beta_1$ exceeded 1%. As such, all DIF in pairwise comparisons before vignette adjustments was classified as negligible.

After using the vignettes to account for response scale use, DIF remained in the comparison between White and Hispanic/Latinx students (Table 14). In contrast, in comparisons between White and Black/African American students, a group whose response scale tendencies became evident in base group comparisons, accounting for response scale use with the vignettes changed the form of DIF from uniform to nonuniform on Items 3 and 4. Although DIF was still flagged

between Black/African American and Hispanic/Latinx students on Items 2 and 4, the vignette adjustment reduced pseudo R^2 by .006 (55%) on Item 4. Similarly, between Black/African American and Asian students, the other group that showed response scale patterns in base group comparisons, the vignette adjustment reduced pseudo R^2 by .011 such that DIF was no longer flagged on Item 3 and DIF on Item 4 was reduced by .008 (57%). Likewise, after accounting for scale use, DIF was no longer flagged on Item 4 in the comparison between White and Asian students.

These results are comparable to the effect of the vignette adjustment on pseudo R^2 in gender comparisons. That is, overall, the magnitude of the vignette adjustment on pseudo R^2 in ethnicity pairwise comparisons was comparable to the effect of the vignettes on R^2 in gender comparisons. However, no McFadden's pseudo R^2 value was above .007 in the vignette-adjusted ethnicity DIF analyses, and no β_1 was greater than 0.49%. As such, all DIF after the vignette adjustment was classified as negligible and ethnic groups appeared to be on comparable measurement scales. Moreover, based on the effect size classification criteria used in this study, the effect of the vignettes on DIF in ethnicity pairwise comparisons was negligible.

5 DISCUSSION

Motivation is a well-documented predictor of a variety of positive student outcomes (Hulleman et al., 2010; Kriegbaum et al., 2018; Robbins et al., 2004). However, researchers have also found evidence of threats to measurement comparability in achievement motivation items, including the presence of group-specific response scale use (e.g., He & Van de Vijver, 2016a). As such, this study evaluated measurement scale comparability in achievement motivation items by testing for differential item functioning (DIF) using pairwise comparisons and multiple-group DIF with a base group. The effect of group differences in response scale use was tested as the source of DIF by comparing changes to DIF outcomes and changes to the form and magnitude of DIF after adjusting self-report item responses with anchoring vignettes.

Overall, although using the anchoring vignettes to account for the effects of group differences in response scale use as a source of DIF changed DIF patterns in some items, all DIF identified before and after the vignette adjustments was negligible. Therefore, gender and ethnic groups appeared to be on comparable measurement scales, and cross-group comparisons would be psychometrically fair. Nonetheless, this study yielded some noteworthy findings that can contribute to the sparse literature base on response scale use and scale comparability in motivation items as well as the effects of anchoring vignettes and group differences in response scale use on DIF. First, though gender and response scale were relatively unrelated, Black/African American and Asian students appeared to demonstrate group-specific response scale use. Second, the vignette correction functioned differently in base group versus pairwise comparisons. Finally, using the nonparametric vignette adjustment to account for response scale use had little effect on DIF if a group did not demonstrate a noticeable response style. These points will be elaborated

on below, followed by a discussion of the implications of these findings, suggestions for future research, limitations, and conclusions.

Main Findings

Group-Specific Response Scale Use

To interpret DIF results, Hambleton (2006) suggested that looking for patterns across DIF items can be more insightful than looking at individual items. To that end, prior to self-reported motivation items being adjusted with the anchoring vignettes, the location of DIF in items and the patterns of group differences in response category thresholds indicated the presence of group-specific response scale use by Black/African American and Asian students; the presence of scale use by gender was less clear. In terms of gender, before items were adjusted with the vignettes, DIF patterns indicated the source of DIF was at the item level. However, male and female students rated the vignettes similarly. Moreover, while more females than males rated themselves as equal to the highest level vignette, more males rated themselves as lower or higher than the highest level vignette. Although these results could be taken as evidence of a modest response style for female students and extreme response style for male students, it would be premature to draw that conclusion. Therefore, the presence of response scale use by gender was unclear and inconsistent. This finding aligns with, for example, Wetzel, Carstensen, et al. (2013), who found both the presence of gender DIF favoring women and the presence of response styles in an achievement-striving facet of a personality scale among German adults; nonetheless, they concluded that gender DIF and response style are two independent influences on item responses.

In contrast, Black/African American and Asian students seemed to demonstrate group-specific response scale use. More specifically, Black/African American students were favored in

both base group comparisons as well as in all pairwise comparisons, except in Item 2, where Hispanic/Latinx students were favored at extremely low levels of motivation. Results indicated that, after being matched by level of motivation, Black/African American students consistently had an easier time shifting from the lower response categories to higher categories on items flagged for DIF. This group of students was also the only group to evidence DIF in pairwise comparisons on Items 2 and 3; the other groups were only flagged for DIF on Item 4. The pattern of response category thresholds in comparisons with Black/African American students showed constant, pervasive DIF in favor of that group, indicating a systematic source of DIF at the item level. Black students also had the highest percentage of students who rated themselves as equal to or higher than the highest level vignette. Taken together, this DIF pattern across items seems to show that Black/African American students demonstrated a group-specific overall preference for agree categories and the upper end of the response scale. This finding is consistent with meta-analytic results showing that Black/African Americans used more extreme responding than other ethnic/cultural groups (Batchelor & Miao, 2016).

Asian students also appeared to demonstrate group-specific response scale use, but their scale use followed a nonuniform pattern in base group and pairwise comparisons. Although a divergent pattern of DIF suggests a source of DIF local to the score level, *across* items, Asian students demonstrated similar reporting behaviors. At the top of the response scale, Asian students had an easier time endorsing the top response categories. This finding is somewhat unexpected as Asian students typically respond with a more modest style (Min et al., 2016). However, Hofer et al. (2010) similarly found that Chinese students demonstrated high levels of achievement motivation in a highly competitive environment. In contrast, at the bottom of the response scale, Asian students showed the opposite DIF pattern and one more consistent with the response scale

use found more often among Asian groups (e.g., Min et al., 2016). In particular, at low levels of motivation, Asian students appeared to need more motivation to shift from the lowest response category to the adjacent one. Furthermore, Asian students had the second-highest percentage of students who rated themselves as equal to the highest level vignette (behind Black/African American students). Thus, Asian students at the low end of the scale may have been demonstrating modest or middling response scale use.

One possible explanation for the different response scale use demonstrated by Black/African American and Asian students can be contextualized along the individualist/collectivist continuum. More specifically, in their meta-analysis, Coon and Kemmelmeir (2001) showed that Black/African Americans scored higher than Latino, European, and Asian Americans on measures of individualism and self-esteem. As culture and an individualist cultural orientation has been shown to be related to extreme response scale use (Chen et al., 1995; Warnecke et al., 1997), one understanding of the scale use demonstrated by Black/African American students in this sample is that it may be due to individualist tendencies. In contrast to their findings regarding Black/African American individuals, Coon and Kemmelmeir (2001) also found that Asian Americans scored higher than any other group on collectivism, which Chen et al. (1995) showed was related to modest responding and using the middle of the response scale. As such, response scale use by Asian students at lower levels of motivation in this study may reflect, for example, culturally reinforced modesty standards or cultural frowning on bragging (He & Van de Vijver, 2016a; Min et al., 2016); that is, Asian students may have viewed it as bragging to agree with a statement about how ambitious they perceive themselves to be. Conversely, Asian students at the highest levels of motivation may have been capturing the ways in which some Asian cultures emphasize effort and self-improvement and not “letting down” teachers, family, or friends as

well as how academic and professional success are valued (De Castella et al., 2013; He & Van de Vijver, 2016a; Hofer et al., 2010).

Methodological Effects of Anchoring Vignettes

Base Group Comparisons. One finding from this study was that the methodological effects of the vignettes were different in base group comparisons versus pairwise comparisons. In the base group comparisons, the effect of the vignettes was to confirm the presence of group-specific response scale use. In pairwise comparisons, the effect of the vignettes was to provide information about the source of DIF. More specifically, if the nonparametric scoring of the vignettes functioned as intended, then when vignette responses are low, adjusted self-report responses are higher; the opposite scoring pattern should occur for self-report responses when vignette ratings are high (von Davier et al., 2018). If vignette responses include extreme categories, adjusted self-report scores cover more of the middle of the scale and exclude the extreme ends. In contrast, if vignette responses are more central and exclude extreme categories, adjusted self-report scores cover more extreme categories (von Davier et al., 2018). These adjustments are hypothesized to capture and account for response scale use (von Davier et al., 2018).

In this study, the hypothesized effect of the vignettes on DIF was such that if the vignettes accounted for scale use and if group differences in response scale use were the source of DIF, then rescaling items with the vignettes would decrease DIF. Moreover, Ellis and Kimmel (1992) argued that comparing groups to an average or composite group, such as the base group constructed for these DIF analyses, reveals the presence of group-specific response scale use. Accordingly, if response scale use is accounted for and DIF is reduced in relation to a common base group, then that would provide evidence of response scale use as a source of DIF. Conversely, if DIF remained in base group comparisons after adjusting items with the vignettes, then

response scale use was not likely to be the source of DIF. Before adjusting items with the vignettes, Black/African American, Hispanic/Latinx, and Asian students evidenced DIF in base group comparisons. Specifically, Black/African American students appeared to use the upper end of the response scale, while Asian students had a nonuniform response pattern, showing a modest style at the bottom of the scale. In contrast, though Hispanic/Latinx students' responses tended to fall at the middle or lower end of the response scale, they did not demonstrate a notable response pattern.

After adjusting self-report scores with the vignettes, DIF was no longer flagged in the comparisons between the base group and Black/African American and Asian students, but DIF remained between the base group and Hispanic/Latinx students. Furthermore, after rescaling, DIF was no longer flagged in the comparison between Black/African and Asian students on Item 3 or between White students and Asian students on Item 4. As such, adjusting the self-report scores appeared to draw Black/African American and Asian students closer to the average. In contrast, DIF did not change in relation to the base group for Hispanic/Latinx students. Taken together, DIF changes in base group comparisons with Black/African American and Asian, but not Hispanic/Latinx students seem to provide evidence supporting the group-specific response scale use identified from DIF patterns before the vignette adjustments. From a methodological perspective, this finding illustrates how anchoring vignettes can be used in base group comparisons to show the presence and effect of group-specific response scale use. Furthermore, when anchoring vignettes are used to account for response scale use, such scale use appears to be a source of DIF. Finally, these findings show that ethnic groups in this sample did seem to demonstrate particular response scale tendencies when responding to self-reported achievement motivation items.

Pairwise Comparisons. The methodological effect of the vignette adjustments in pairwise comparisons was different than in base group comparisons. In pairwise comparisons, the vignettes increased item information and helped to inform potential sources of DIF. More specifically, Lu and Bolt (2015) noted that, in general, PISA attitudinal items tend to elicit agree responses. Items in this sample appeared to follow that pattern and demonstrated considerable ceiling effects. That is, item responses were clustered at the top of the scale, and therefore items provided little information about students with average or higher levels of self-reported achievement motivation. For pairwise comparisons across both gender and ethnicity, per the intended effect of the vignettes, vignette adjustments increased item discrimination. The response scale was "stretched" such that the form of DIF changed and parts of the scale that lacked discrimination before rescoring showed increased item discrimination and item information after rescoring. In turn, the stretched scaled provided more fine-grained information about the possible source of DIF.

For gender, as noted, before the vignette adjustments, DIF analyses revealed that items favored female students, suggesting a source of DIF inherent to the item level. After adjusting self-report items with the vignettes, however, group thresholds revealed a shift in who was favored by the item, from female students at low levels of motivation to male students at high levels of motivation. This pattern indicated a source of DIF local to the score level, rather than inherent to the item itself. Specifically, in this sample, boys with the highest levels of motivation had an easier time than girls with the same level of motivation endorsing higher response categories. This finding was consistent with, for example, Elmore and Oyserman (2012) who used an experimental design and found that expectations about future success and gender influence cur-

rent academic efforts, particularly for boys. The effect of the vignettes in ethnicity pairwise comparisons was similar. For example, the form of DIF shifted from uniform to nonuniform in the comparison between White and Black/African American students. Though White students were favored at the lowest end of the response scale, Black/African American students were favored starting at below average (i.e., θ = -1.5) and higher levels of motivation. Thus, in pairwise comparisons, the vignettes increased item discrimination, and as a result, provided more information about the source of DIF.

Effect of the Vignettes and Response Scale Use as a Source of DIF

Findings from this study show that, essentially, the effect of the vignettes as a methodological correction for response scale use on DIF was only useful to the extent that groups demonstrated response scale tendencies. As such, when group-level response styles were the source of DIF, the nonparametric vignette scoring did reduce DIF in some pairwise comparisons. For example, DIF was no longer flagged in the comparison between White and Asian students; thus, accounting for scale use appeared to put White and Asian students on the same measurement scale. Similarly, DIF was no longer flagged in Item 3 in the comparison between Black/African American and Asian students, both groups that demonstrated response scale tendencies. Thus, accounting for response scale use preferences seemed to reduce DIF and improve measurement scale comparability for these groups. However, if groups did not show particular scale use, the effect of the vignettes on DIF was minimal. For some items, the vignettes even had the opposite effect and actually increased DIF effect sizes. Overall, therefore, accounting for response scale use as a source of DIF with the vignettes had little effect on DIF.

Of note, the effect of the vignettes on changes to DIF in pairwise comparisons was relatively similar across gender and ethnicity. von Davier et al. (2018) showed that the nonparametric vignette adjustments always introduce dependency because self-report items are, by definition, recoded as a function of vignette responses. The nonparametric scoring homogenizes the covariance structure such that reliability and correlations always increase, even when ties or misordering are present. Therefore, the comparable effect size of the vignettes across gender and ethnicity in pairwise comparisons could indicate that changes to DIF were actually artifactual and due to spurious dependencies introduced by the nonparametric scoring. For example, the shift in gender DIF in Item 1 from uniform to nonuniform DIF or the shift in gender DIF from Item 5 to Item 3 may have simply been spurious stretching of the response scale. This has implications for using the vignettes in a DIF analysis because it implies different sources of DIF, from the item level to the score level, and could impact how DIF results are interpreted. Despite this issue with the vignettes, findings from this study show that correcting self-reported achievement motivation items with anchoring vignettes did change the form, magnitude, and outcome of DIF across gender and some ethnicity group comparisons, though the effect on DIF was negligible.

Implications

Implications for Motivation Research and Educational Measurement and Assessment

Although groups demonstrated differences in response scale use to motivation items, those tendencies ultimately had little practical effect on measurement scale comparability. Given the absence of measurement bias in the PISA achievement motivation items, it appears to be psychometrically fair to make cross-group comparisons about achievement motivation based those items. Nonetheless, group differences in response scale use were present in motivation items, and

motivation researchers should account for such differences when measuring achievement motivation to ensure that construct-irrelevant variance due to response styles does not distort measurement. Hopwood et al. (2009) further argued that it is important to understand the meaning of a group's response scale use because response scale use in and of itself can be substantively meaningful. In this study, group-specific response scale use was consistent with cultural differences across ethnic groups, and results show how complex achievement motivation is. That is, even within groups, responses to motivation items capture different cultural phenomena. As such, test developers need to be aware of how culture influences response scale use and measurement in motivation items and account for such differences, particularly when constructing scoring rules and test norms (Guo et al., 2016; Sue, 1996).

Motivation researchers should also investigate response scale use from a socio-cultural perspective to see if it can substantively inform any findings regarding group differences. For example, some response style researchers have hypothesized that the function of extreme response scale is to reduce cognitive load when items become more difficult and effortful to interpret (Krosnick, 1991). Although this kind of scale use has been connected to cognitive abilities and socioeconomic status (Batchelor & Miao, 2016; He, Buchholz, et al., 2017), a socio-cultural perspective contends that such scale use could be a reflection of individualism. For example, given that Black/African Americans were shown to score high on measures of individualism (Coon & Kemmelmeier, 2001) and that they showed the most extreme scale use of any ethnic group in the United States (Batchelor & Miao, 2016), future researchers could explore individualism or other cultural characteristics as alternate explanations for response scale use among Black/African Americans on motivation items.

In addition to group differences in response scale use, the same items were not salient across gender and ethnicity. That is, Item 1 (“top grades”) was salient to gender, but not to ethnicity. Conversely, Item 4 (“ambitious”) was salient to ethnicity, but not for gender. Item 2 (“best opportunities”) was primarily salient for gender, but did evidence DIF in the multiple-group DIF and in the Black/African American and Hispanic/Latinx comparison. These differences may tap into the identity-based motivation model, which predicts that the identity most salient to students in a certain context influences what they attend to and the choices they make (Oyserman & Destin, 2010). For example, research on academic disidentification has illustrated how Black/African American students, particularly males, have so many negative experiences at school that they stop equating self-esteem with academic achievement and they stop relating effort to academic achievement (Cokley, 2002). Thus, in this study, ambition may have been more salient for Black/African American students, but motivation to obtain top grades or to be the best student was less relevant.

That different items were salient to different groups has implications for both classroom application and test development. In the classroom, how teachers motivate students and the different interventions they implement or instructional strategies they use should be informed by relevant student identities and values. This is particularly important as Trumbull and Rothstein-Fisch (2011) noted that few members of school communities recognize how achievement motivation is influenced by culture. For example, Oyserman and Destin (2010) administered an intervention based on the identity-based motivation model and found that students acted when behaviors and goals were both meaningful and identity-congruent. Similarly, the reasons underlying why individuals endorsed performance-approach goals were more important in predicting well-being than endorsing the actual items themselves (Gillet et al., 2014). In this study, for instance,

getting top grades (Item 1) may not have tapped into relevant motivation for Black/African American students, but wanting access to the best opportunities (Item 2) may have been germane to their achievement motivation. As such, teachers may have an easier time motivating students if they are aware of what is meaningful to students and appeal to what is important and relevant to them. Students from traditionally marginalized groups also need to know why a school task is useful and how it is applicable to their future goals (Martinez & Guzman, 2013).

For assessment designers, such as those who administer PISA, that different items were salient to gender and ethnicity has implications for item and test development. Test writers need to pay attention to the different ways motivation functions across groups and to be thoughtful about the construct-relevance of items to groups (e.g., social goals for individuals from Eastern societies vs. achievement goals for individuals from individualistic societies) and the cultural variability of constructs within groups (Chiesi et al., 2020).

Moreover, issues with the PISA motivation items can be used to inform future assessment development. In particular, in the multiple-group DIF and ethnicity pairwise comparisons, Item 4 was consistently flagged for DIF more than any other item. Item 4 was worded differently than the other items, asking students to evaluate a self-perception ("I see myself"). In contrast, the other four items referred to what students wanted, such as wanting to "be the best student" or to have access to the "best opportunities." Not only do these items prompt a maximalist mentality by priming students to think about what being the "best" means (Cheek & Schwartz, 2016), but also they provide a clearer reference point than Item 4, which was more abstract and more open-ended. Karabenick et al. (2007) noted that clarity of an item, such as how concrete a prompt is, can affect the cognitive validity of motivation items and motivation-related constructs. In terms of the cognitive response process to survey items described by Tourangeau (2018), Warnecke et

al. (1997) noted that when a survey item has a clear and precise response, no judgment is needed and so item responses are less likely to be influenced by factors other than the construct being measured. In contrast, in the absence of specific cues from an item when a respondent has to make a judgment about information retrieved from memory, the response process becomes more complex and is more likely to be affected by gender and ethnicity (Warnecke et al., 1997).

Results from this study illustrate that when students are queried regarding self-perceptions of ambition, ethnic group status is related to responses on abstract items. Therefore, test developers need to be aware of the effects of abstract items on measurement and to develop concrete items with clear reference points. Educational measurement specialists also need to think about how culture affects item interpretation and account for such differences across all phases of testing, including score reporting. This would seem to be particularly important in the context of a politically influential assessment such as PISA (Pepper et al., 2018), the results of which are intended to be used in cross-group comparisons and often serve as the basis for educational policies in many countries around the world (Hopfenbeck et al., 2018).

Implications for DIF Researchers

Findings from this study can be used to inform DIF research regarding sources of DIF and the effects of anchoring vignettes on DIF. This study also illustrated a methodological application of base group multiple-group DIF and how it can be used in response scale research. For researchers interested in sources of DIF, the findings from this study are consistent with others, such as Gnambs and Hanfstingl (2014) and Wetzel, Carstensen, et al. (2013), who found that correcting for group differences in response scale had a negligible effect on DIF in motivation measures. The consistencies with these studies are important for two reasons. First, while the samples evaluated by Gnambs and Hanfstingl (2014) and Wetzel, Carstensen, et al. (2013) were

relatively homogenous (i.e., German adults and teens), the sample in this study was diverse and included students with a broad range of backgrounds. Second, those researchers used latent class analysis, and this study used a hybrid DIF framework. That different samples and different methodologies yielded similar findings offers convergent evidence that group differences in response scale use appear to have little impact on DIF. That is, while accounting for response styles may yield more accurate DIF detection and more accurate construct measurement, response scale use and group differences in scale use are not necessarily a major source of DIF. However, more research would be needed to evaluate if these findings generalize to other social/emotional constructs or self-reported health or other age groups.

This study demonstrated a novel use of anchoring vignettes as a method to account for response scale use as a source of DIF in a scale comparability study and DIF analysis. For DIF researchers, the effect of the vignette adjustment on the base group DIF was that it confirmed the presence of response scale use by groups and showed scale use as a source of DIF. However, the effect of the vignettes on pairwise comparisons when a group did not show scale use was minimal. Moreover, even when groups did show scale use tendencies, the effect of the vignettes on reducing DIF was negligible. As such, given that pairwise comparisons are the basis for virtually all DIF, with multiple-group DIF methods being underutilized (Oshima et al., 2015) and the psychometric problems observed with the vignettes, it seems that the energy of DIF researchers would be better directed towards other methods to account for scale use.

From a methodological perspective, in cross-country DIF comparisons, the same problems arise as with multilevel variables like ethnicity in terms of which country should be selected to serve as the reference group; multiple-group DIF with a base group overcomes the reference

group problem. Furthermore, in the same way measurement scales across ethnic groups are confounded by response scale use, at the country-level, item responses are similarly affected by response-style variance (He & Van de Vijver, 2016a). Moors (2004) called response scale use a "threat to every measurement" regardless of context, cross-cultural or otherwise. Despite recommendations to account for scale use in cross-group comparisons (Bolt & Johnson, 2009), some methods of controlling for response styles, such as latent class analysis, may not be feasible or accessible for researchers without advanced statistical training. Moreover, some methods require researchers to interpret response styles or to statistically derive indicators of response scale use (e.g., Wetzel et al., 2016).

To that end, this study illustrated that base-group multiple-group DIF is not computationally intensive, and it is a reasonable method for researchers who are interested in comparing groups based on categorical survey data to identify the presence of group-specific response scale use. In this study, base group DIF comparisons identified group-specific response scale use by Black/African American and Asian students. How that scale use is quantified in this method also has a relatively straightforward interpretation—a group is above, equal to, or less than the average response scale. Thus, base-group DIF gives researchers a sense of the magnitude of and direction of scale use and, in turn, how scale use could affect other measurements. Furthermore, for researchers doing complicated cross-country analyses, this metric of scale use may be easier to interpret than, for example, derived latent classes (Wetzel, Böhnke, et al., 2013).

Suggestions for Further Research

In light of the above findings, some recommendations for future research can be made. First, Item 2 evidenced an interesting DIF pattern before and after the vignette adjustment that lends itself to further research regarding power in multiple-group DIF. Before being rescaled

with the vignettes, Item 2 did not evidence DIF in the multiple-group DIF, but DIF was flagged in the pairwise comparison between Black/African American and Hispanic/Latinx students. After the vignette adjustment, Item 2 was flagged for DIF in the multiple-group comparison. One explanation for this is that the vignettes simply redistributed DIF and the DIF flagged in Item 2 was artifactual or spurious, possibly even reflecting pseudo-DIF. An alternative explanation has to do with power and the number of response categories in a multiple-group DIF. That is, the minimum cell count was set to 5 for each response category (Kang & Chen, 2008; Orlando & Thissen, 2000). This meant that before the vignette adjustment, in the multiple-group DIF, only two response categories were retained (i.e., the item was dichotomized), but in the pairwise comparison between Black/African American and Hispanic/Latinx students, four categories were retained.

In contrast, after the adjustments, the multiple-group DIF for Item 2 had four categories and DIF was flagged. As the effect of the vignettes was that they simultaneously increased the number of response categories and item discrimination, it would be interesting to study the independent effects of both on multiple-group DIF. Given that multiple-group ordinal logistic regression DIF has generally not been well-studied and that less than a handful of studies have reported on the impact of the number of response categories in polytomous items on DIF (e.g., Allahyari et al., 2016; Hidalgo et al., 2016), future simulation studies could manipulate these factors as independent variables to learn more about power in ordinal logistic regression multiple-group DIF.

Second, although anchoring vignettes have been used to establish the presence of response styles (e.g., He & Van de Vijver, 2016a), this study established the presence of response styles in ethnicity by using a common base group as the reference group in DIF. However, gender DIF was not evaluated in relation to the base group. Given that the base group comparison

revealed scale use and that scale use by gender has revealed inconsistent findings, in future studies, researchers could examine gender-specific scale use in relation to a base group. Similarly, consistent with a continuous perspective of response scale use, within-group differences in response scale use should be examined in relation to a base group constructed only from members of that group. For example, by comparing Black/African American males and females to a base group comprised of Black/African Americans, the base group comparison reveals how response scales are used by Black/African American males and females. Similarly, future DIF analyses of motivation items may be more informative when DIF groupings are based on gender by ethnicity interactions (e.g., comparing White females to Asian females). As Dorans and Holland (1992) noted, "marginal DIF analysis", or the separate analysis of DIF by gender and ethnicity, ignores potential important gender by ethnicity interactions. Furthermore, given that achievement motivation is known to correlate with personality variables such as conscientiousness (Dumfart & Neubauer, 2016; Wetzel, Böhnke, et al., 2013), as well as academic achievement, future studies should look at the relationship between them as this would have implications for item development and test design.

Although the utility of the vignettes was questionable, future research on vignettes could evaluate if anchoring vignettes are better or more effective at correcting for scale use and DIF in certain kinds of constructs or to compare differences between item wording and the construct measured. Moreover, researchers have found evidence of an interaction between gender/ethnicity of the respondent and gender/ethnicity of the vignette characters (e.g., Grol-Prokopczyk, 2014; Grol-Prokopczyk et al., 2011; Knott et al., 2017). The effects of such interactions should con-

tinue to be investigated. Finally, Primi et al. (2018) found a relationship between vignette responses and academic achievement. A follow-up study to this one could further examine the relationship between group differences in response scale use and academic achievement.

Limitations

As with any study, there were limitations in this one. First, there were only five motivation items. In general, DIF with short scales is less reliable, mostly due to the quality of the matching score (Hidalgo et al., 2016; Scott et al., 2009), but is still concerning with latent trait DIF. Second, despite OECD recommendations to use sampling weights with PISA data, no sampling weights were used in this study. This is consistent with the approach to sampling weights taken by researchers in DIF analyses of PISA items that utilized latent trait modeling (e.g., Oliveri et al., 2014). Nonetheless, the use of sampling weights may have affected DIF results and the effect of sampling weights on DIF are important to investigate. Third, DIF groupings were based on gender and ethnicity. Though these are typical DIF groupings, ethnic group labels are assigned to individuals and cannot be interpreted as reflecting how they would self-identify (Oyserman & Destin, 2010). Moreover, ascribing such labels to groups may mask important and meaningful within-group differences. Fourth, based on concerns expressed by Primi et al. (2018), rescoring ties at the lowest score level could have also impacted results by drawing results closer to the average than they would have been had a different approach to ties been used.

Finally, Item 4 was the most problematic item for ethnicity, and its construct relevance to an achievement motivation scale is questionable. Item 4 asked students to rate their self-perceptions of ambition, which Judge and Kammeyer-Mueller (2012) argue is separate from achievement motivation. In contrast to Item 4, the other four items asked students about their desire to be the best and to obtain top grades and to access to the best opportunities. There was little DIF

in these items for ethnic groups, with the exception of Item 2. The absence of DIF in Items 1, 2, 3, and 5, but the presence of DIF in Item 4 draws attention to concerns that these items are measuring different constructs or they tapped into different response processes. For example, the pattern of DIF in the PISA items across ethnicity seems to align with the distinction drawn between aspiration and ambition (Judge & Kammeyer-Mueller, 2012). Such limitations, however, do not negate item-specific scale comparability findings.

Conclusions

This study revealed the presence of negligible DIF in the PISA 2015 achievement motivation items. Overall, although using the anchoring vignettes to account for the effect of group differences in response scale use as a source of DIF changed DIF patterns in some items, all DIF identified before and after the vignette adjustments was negligible. Therefore, gender and ethnic groups appeared to be on comparable measurement scales and cross-group comparisons would be psychometrically fair. This study added to the research literature by showing evidence supporting the presence of group-specific response scale use in motivation items, by showing the presence of group-specific response scale use in multiple-group DIF with a base group, and by showing that the nonparametric anchoring vignette adjustment to account for response scale use as a source of DIF had little effect on DIF if a group did not demonstrate a noticeable response style.

REFERENCES

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

Agresti, A. (2002). *Categorical data analysis* (2nd ed). Wiley-Interscience.

Allahyari, E., Jafari, P., & Bagheri, Z. (2016). A simulation study to assess the effect of the number of response categories on the power of ordinal logistic regression for differential item functioning analysis in rating scales. *Computational and Mathematical Methods in Medicine*, *2016*, 1–8. https://doi.org/10.1155/2016/5080826

Ames, A. J., & Penfield, R. D. (2015). An NCME instructional module on item-fit statistics for item response theory models. *Educational Measurement: Issues and Practice*, *34*(3), 39–48. https://doi.org/10.1111/emip.12067

Ames, C. A. (1990). Motivation: What teachers need to know. *Teachers College Record*, *91*(3), 409–421.

Au, N., & Lorgelly, P. K. (2014). Anchoring vignettes for health comparisons: An analysis of response consistency. *Quality of Life Research*, *23*(6), 1721–1731. https://doi.org/10.1007/s11136-013-0615-2

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences*, *40*(6), 1235–1245. https://doi.org/10.1016/j.paid.2005.10.018

Bago d'Uva, T., Van Doorslaer, E., Lindeboom, M., & O'Donnell, O. (2008). Does reporting heterogeneity bias the measurement of health disparities? *Health Economics*, *17*(3), 351–375. https://doi.org/10.1002/hec.1269

Balluerka, N., Plewis, I., Gorostiaga, A., & Padilla, J.-L. (2014). Examining sources of DIF in psychological and educational assessment using multilevel logistic regression. *Methodology*, *10*(2), 71–79. https://doi.org/10.1027/1614-2241/a000076

Baranik, L. E., Stanley, L. J., Bynum, B. H., & Lance, C. E. (2010). Examining the construct validity of mastery-avoidance achievement goals: A meta-analysis. *Human Performance*, *23*(3), 265–282. https://doi.org/10.1080/08959285.2010.488463

Batchelor, J. H., & Miao, C. (2016). Extreme response style: A meta-analysis. *Journal of Organizational Psychology*, *16*(2), 51–62.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 159–181. https://doi.org/10.1111/bmsp.12086

Boer, D., Hanke, K., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology*, *49*(5), 713–734. https://doi.org/10.1177/0022022117749042

Bolt, D. M., & Gierl, M. J. (2006). Testing features of graphical DIF: Application of a regression correction to three nonparametric statistical tests. *Journal of Educational Measurement*, *43*(4), 313–333. https://doi.org/10.1111/j.1745-3984.2006.00019.x

Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement*, *33*(5), 335–352. https://doi.org/10.1177/0146621608329891

Bolt, D. M., Lu, Y., & Kim, J.-S. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods*, *19*(4), 528–541. https://doi.org/10.1037/met0000016

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, *44*(11 Suppl 3), S176–S181. https://doi.org/10.1097/01.mlr.0000245143.08679.cc

Butler, R., & Hasenfratz, L. (2017). Gender and competence motivation. In A. J. Elliot, C. S. Dweck, & D. S. Yeager (Eds.), *Handbook of competence and motivation: Theory and application* (2nd ed., pp. 489–511). The Guildford Press.

Bzostek, S., Sastry, N., Goldman, N., Pebley, A., & Duffy, D. (2016). Using vignettes to rethink Latino-white disparities in self-rated health. *Social Science & Medicine*, *149*, 46–65. https://doi.org/10.1016/j.socscimed.2015.11.031

Cai, T. (2013). Investigation of ways to handle sampling weights for multilevel model analyses. *Sociological Methodology*, *43*(1), 178–219. https://doi.org/10.1177/0081175012460221

Caldwell, T., & Obasi, E. M. (2010). Academic performance in African American undergraduates: Effects of cultural mistrust, educational value, and achievement motivation. *Journal of Career Development*, *36*(4), 348–369. https://doi.org/10.1177/0894845309349357

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

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

Campbell, H. L., Barry, C. L., Joe, J. N., & Finney, S. J. (2008). Configural, metric, and scalar invariance of the modified achievement goal questionnaire across African American and White university students. *Educational and Psychological Measurement*, *68*(6), 988–1007. https://doi.org/10.1177/0013164408315269

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

Cheek, N. N., & Schwartz, B. (2016). On the meaning and measurement of maximization. *Judgment and Decision Making*, *11*(2), 126–146.

Chen, C., Lee, S., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science*, *6*(3), 170–175. https://doi.org/10.1111/j.1467-9280.1995.tb00327.x

Chen, H.-F., Jin, K.-Y., & Wang, W.-C. (2017). Modified logistic regression approaches to eliminating the impact of response styles on DIF detection in Likert-type scales. *Frontiers in Psychology*, *8*, 1143. https://doi.org/10.3389/fpsyg.2017.01143

Chen, Y., Thissen, D., Anand, D., Chen, L. H., Liang, H., & Daughters, S. B. (2019). Evaluating differential item functioning (DIF) of the Chinese version of the Behavioral Activation for Depression Scale (C-BADS). *European Journal of Psychological Assessment*, *36*(2), 303–323. https://doi.org/10.1027/1015-5759/a000525

Cheung, K-C., Mak, S-K., & Sit, P-S. (2018). Resolving the attitude–achievement paradox based on anchoring vignettes: Evidences from the PISA 2012 mathematics study. *Asia Pacific Education Review*, *19*(3), 389–399. https://doi.org/10.1007/s12564-018-9526-9

Chiesi, F., Lau, C., Marunic, G., Sanchez-Ruiz, M.-J., Plouffe, R. A., Topa, G., Yan, G., & Saklofske, D. H. (2020). Emotional intelligence in young women from five cultures: A TEIQue-SF invariance study using the omnicultural composite approach inside the IRT framework. *Personality and Individual Differences*, *164*, 110128. https://doi.org/10.1016/j.paid.2020.110128

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An *R* package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software*, *39*(8), 1–30. https://doi.org/10.18637/jss.v039.i08

Clauser, B. E., & Mazor, K. M. (2005). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31–44. https://doi.org/10.1111/j.1745-3992.1998.tb00619.x

Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement*, *17*(4), 335–350. https://doi.org/10.1177/014662169301700402

Cokley, K. O. (2002). Ethnicity, gender, and academic self-concept: A preliminary examination of academic disidentification and implications for psychologists. *Cultural Diversity and Ethnic Minority Psychology*, *8*(4), 378–388. https://doi.org/10.1037//1099-9809.8.4.378

Cokley, K. O., Komarraju, M., King, A., Cunningham, D., & Muhammad, G. (2003). Ethnic differences in the measurement of academic self-concept in a sample of African American and European American college students. *Educational and Psychological Measurement*, *63*(4), 707–722. https://doi.org/10.1177/0013164402251055

Conley, A. M. (2012). Patterns of motivation beliefs: Combining achievement goal and expectancy-value perspectives. *Journal of Educational Psychology*, *104*(1), 32–47. https://doi.org/10.1037/a0026042

Conroy, D. E. (2017). Achievement motives. In A. J. Elliot, C. S. Dweck, & D. S. Yeager (Eds.), *Handbook of competence and motivation: Theory and application* (2nd ed., pp. 25–42). The Guildford Press.

Coon, H. M., & Kemmelmeier, M. (2001). Cultural orientations in the United States: (Re)examining differences among ethnic groups. *Journal of Cross-Cultural Psychology*, *32*(3), 348–364. https://doi.org/10.1177/0022022101032003006

Covington, M. V. (2000). Goal theory, motivation, and school achievement: An integrative review. *Annual Review of Psychology*, *51*(1), 171–200. https://doi.org/10.1146/annurev.psych.51.1.171

Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. *Medical Care*, *44*(Suppl 3), S115–S123. https://doi.org/10.1097/01.mlr.0000245183.28384.ed

Crane, P. K., Gibbons, L. E., Ocepek-Welikson, K., Cook, K., Cella, D., Narasimhalu, K., Hays, R. D., & Teresi, J. A. (2007). A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Quality of Life Research*, *16*(S1), 69–84. https://doi.org/10.1007/s11136-007-9185-5

Cury, F., Elliot, A. J., Da Fonseca, D., & Moller, A. C. (2006). The social-cognitive model of achievement motivation and the 2 × 2 achievement goal framework. *Journal of Personality and Social Psychology*, *90*(4), 666–679. https://doi.org/10.1037/0022-3514.90.4.666

Day, E. A., Radosevich, D. J., & Chasteen, C. S. (2003). Construct- and criterion-related validity of four commonly used goal orientation instruments. *Contemporary Educational Psychology*, *28*(4), 434–464. https://doi.org/10.1016/S0361-476X(02)00043-7

De Castella, K., Byrne, D., & Covington, M. (2013). Unmotivated or motivated to fail? A cross-cultural study of achievement motivation, fear of failure, and student disengagement. *Journal of Educational Psychology*, *105*(3), 861–880. https://doi.org/10.1037/a0032464

DeMars, C. E. (2010). Type I error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement*, *70*(6), 961–972. https://doi.org/10.1177/0013164410366691

DeVellis, R. F. (2006). Classical test theory. *Medical Care*, *44*(11 Suppl 3), S50–S59.

Dever, B. V., & Kim, S. Y. (2016). Measurement equivalence of the PALS academic self-efficacy scale. *European Journal of Psychological Assessment*, *32*(1), 61–67. https://doi.org/10.1027/1015-5759/a000331

D'Lima, G. M., Winsler, A., & Kitsantas, A. (2014). Ethnic and gender differences in first-year college students' goal orientation, self-efficacy, and extrinsic and intrinsic motivation. *The Journal of Educational Research*, *107*(5), 341–356. https://doi.org/10.1080/00220671.2013.823366

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization. *ETS Research Report Series*, *1*, i–40. https://doi.org/10.1002/j.2333-8504.1992.tb01440.x

Dumfart, B., & Neubauer, A. C. (2016). Conscientiousness is the most powerful noncognitive predictor of school achievement in adolescents. *Journal of Individual Differences*, *37*(1), 8–15.

Dupeyrat, C., Escribe, C., Huet, N., & Régner, I. (2011). Positive biases in self-assessment of mathematics competence, achievement goals, and mathematics performance. *International Journal of Educational Research*, *50*(4), 241–250. https://doi.org/10.1016/j.ijer.2011.08.005

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Edman, J. L., & Brazil, B. (2009). Perceptions of campus climate, academic efficacy and academic success among community college students: An ethnic comparison. *Social Psychology of Education*, *12*(3), 371–383. https://doi.org/10.1007/s11218-008-9082-y

Elder, C. (1997). What does test bias have to do with fairness? *Language Testing*, *14*(3), 261–277. https://doi.org/10.1177/026553229701400304

Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist*, *34*(3), 169–189.

Elliot, A. J. (2006). The hierarchical model of approach-avoidance motivation. *Motivation and Emotion*, *30*(2), 111–116. https://doi.org/10.1007/s11031-006-9028-7

Elliot, A. J., Aldhobaiban, N., Kobeisy, A., Murayama, K., Gocłowska, M. A., Lichtenfeld, S., & Khayat, A. (2016). Linking social interdependence preferences to achievement goal adoption. *Learning and Individual Differences*, *50*, 291–295. https://doi.org/10.1016/j.lindif.2016.08.020

Elliot, A. J., Aldhobaiban, N., Murayama, K., Kobeisy, A., Gocłowska, M. A., & Khyat, A. (2018). Impression management and achievement motivation: Investigating substantive links. *International Journal of Psychology*, *53*(1), 16–22. https://doi.org/10.1002/ijop.12252

Elliot, A. J., & Church, M. A. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology*, *72*(1), 218–232.

Elliot, A. J., & Dweck, C. S. (2005). Competence and motivation: Competence as the core of achievement motivation. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 3–14). Guilford Press.

Elliot, A. J., Dweck, C. S., & Yeager, D. S. (2017). Competence and motivation: Theory and application. In A. J. Elliot, C. S. Dweck, & D. S. Yeager (Eds.), *Handbook of competence and motivation: Theory and application* (2nd ed., pp. 3–5). The Guildford Press.

Elliot, A. J., Murayama, K., & Pekrun, R. (2011). A 3 × 2 achievement goal model. *Journal of Educational Psychology*, *103*(3), 632–648. https://doi.org/10.1037/a0023952

Elliot, A. J., & Thrash, T. M. (2001). Achievement goals and the hierarchical model of achievement motivation. *Educational Psychology Review*, *13*(2), 139–156.

Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology*, *77*(2), 177–184.

Elmore, K. C., & Oyserman, D. (2012). If 'we' can succeed, 'I' can too: Identity-based motivation and gender in the classroom. *Contemporary Educational Psychology*, *37*(3), 176–185. https://doi.org/10.1016/j.cedpsych.2011.05.003

Ferrando, P. J. (2014). A factor-analytic model for assessing individual differences in response scale usage. *Multivariate Behavioral Research*, *49*(4), 390–405. https://doi.org/10.1080/00273171.2014.911074

Feuerherd, M., Knuth, D., Muehlan, H., & Schmidt, S. (2014). Differential item functioning (DIF) analyses of the Impact of Event Scale-Revised (IES-R): Results from a large European study on people with disaster experiences. *Traumatology*, *20*(4), 313–320. https://doi.org/10.1037/h0099858

Finch, W. H. (2011). The use of multiple imputation for missing data in uniform DIF analysis: Power and type I error rates. *Applied Measurement in Education*, *24*(4), 281–301. https://doi.org/10.1080/08957347.2011.607054

Finch, W. H. (2016). Detection of differential item functioning for more than two groups: A Monte Carlo comparison of methods. *Applied Measurement in Education*, *29*(1), 30–45. https://doi.org/10.1080/08957347.2015.1102916

Finch, W. H. (2020). Using fit statistic differences to determine the optimal number of factors to retain in an exploratory factor analysis. *Educational and Psychological Measurement*, *80*(2), 217–241. https://doi.org/10.1177/0013164419865769

Finch, W. H., Hernández Finch, M. E., & French, B. F. (2016). Recursive partitioning to identify potential causes of differential item functioning in cross-national data. *International Journal of Testing*, *16*(1), 21–53. https://doi.org/10.1080/15305058.2015.1039644

Forrest, C. B., Bevans, K. B., Pratiwadi, R., Moon, J., Teneralli, R. E., Minton, J. M., & Tucker, C. A. (2014). Development of the PROMIS® pediatric global health (PGH-7) measure. *Quality of Life Research*, *23*(4), 1221–1231. https://doi.org/10.1007/s11136-013-0581-8

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, *33*(3), 315–332. https://doi.org/10.1111/j.1745-3984.1996.tb00495.x

Fryer, J. W., & Elliot, A. J. (2007). Stability and change in achievement goals. *Journal of Educational Psychology*, *99*(4), 700–714. https://doi.org/10.1037/0022-0663.99.4.700

Fulmer, S. M., & Frijters, J. C. (2009). A review of self-report and alternative approaches in the measurement of student motivation. *Educational Psychology Review*, *21*(3), 219–246. https://doi.org/10.1007/s10648-009-9107-x

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, *22*(2), 153–164. https://doi.org/10.1214/088342306000000691

Gierl, M. J., Bisanz, J., Bisanz, G. L., & Boughton, K. A. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis paradigm. *Journal of Educational Measurement*, *40*(4), 281–306.

Gillet, N., Lafrenière, M.-A. K., Vallerand, R. J., Huart, I., & Fouquereau, E. (2014). The effects of autonomous and controlled regulation of performance-approach goals on well-being: A process model. *British Journal of Social Psychology*, *53*(1), 154–174. https://doi.org/10.1111/bjso.12018

Gnambs, T., & Hanfstingl, B. (2014). A differential item functioning analysis of the German Academic Self-Regulation Questionnaire for adolescents. *European Journal of Psychological Assessment*, *30*(4), 251–260. https://doi.org/10.1027/1015-5759/a000185

Gray, D. L., Hope, E. C., & Matthews, J. S. (2018). Black and belonging at school: A case for interpersonal, instructional, and institutional opportunity structures. *Educational Psychologist*, *53*(2), 97–113. https://doi.org/10.1080/00461520.2017.1421466

Grimm, S. D., & Church, A. T. (1999). A cross-cultural study of response biases in personality measures. *Journal of Research in Personality*, *33*(4), 415–441. https://doi.org/10.1006/jrpe.1999.2256

Grol-Prokopczyk, H. (2014). Age and sex effects in anchoring vignette studies: Methodological and empirical contributions. *Survey Research Methods*, *8*(1), 1–17.

Grol-Prokopczyk, H., Freese, J., & Hauser, R. M. (2011). Using anchoring vignettes to assess group differences in general self-rated health. *Journal of Health and Social Behavior*, *52*(2), 246–261. https://doi.org/10.1177/0022146510396713

Grol-Prokopczyk, H., Verdes-Tennant, E., McEniry, M., & Ispány, M. (2015). Promises and pitfalls of anchoring vignettes in health survey research. *Demography*, *52*(5), 1703–1728. https://doi.org/10.1007/s13524-015-0422-1

Guo, H., Zu, J., Kyllonen, P., & Schmitt, N. (2016). Evaluation of different scoring rules for a noncognitive test in development. *ETS Research Report Series*, *2016*(1), 1–13. https://doi.org/10.1002/ets2.12089

Hamamura, T., Meijer, Z., Heine, S. J., Kamaya, K., & Hori, I. (2009). Approach—avoidance motivation and information processing: A cross-cultural analysis. *Personality and Social Psychology Bulletin*, *35*(4), 454–462. https://doi.org/10.1177/0146167208329512

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care*, *44*(11 Suppl 3), S182–S188. https://doi.org/10.1097/01.mlr.0000245443.86671.c4

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. SAGE Publications, Inc.

Harzing, A.-W. (2006). Response styles in cross-national survey research: A 26-country study. *International Journal of Cross Cultural Management*, *6*(2), 243–266. https://doi.org/10.1177/1470595806066332

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*(2), 139–164. https://doi.org/10.1177/014662168500900204

He, J., Buchholz, J., & Klieme, E. (2017). Effects of anchoring vignettes on comparability and predictive validity of student self-reports in 64 cultures. *Journal of Cross-Cultural Psychology*, *48*(3), 319–334. https://doi.org/10.1177/0022022116687395

He, J., & Van de Vijver, F. J. R. (2016a). The motivation-achievement paradox in international educational achievement tests: Toward a better understanding. In R. B. King & A. B. I. Bernardo (Eds.), *The Psychology of Asian Learners* (pp. 253–268). Springer Singapore. https://doi.org/10.1007/978-981-287-576-1_16

He, J., & Van de Vijver, F. J. R. (2016b). Response styles in factual items: Personal, contextual and cultural correlates. *International Journal of Psychology*, *51*(6), 445–452. https://doi.org/10.1002/ijop.12263

He, J., Van de Vijver, F. J. R., Fetvadjiev, V. H., de Carmen Dominguez Espinosa, A., Adams, B., Alonso-Arbiol, I., Aydinli-Karakulak, A., Buzea, C., Dimitrova, R., Fortin, A., Hapunda, G., Ma, S., Sargautyte, R., Sim, S., Schachner, M. K., Suryani, A., Zeinoun, P., & Zhang, R. (2017). On enhancing the cross-cultural comparability of Likert-scale personality and value measures: A comparison of common procedures. *European Journal of Personality*, *31*(6), 642–657. https://doi.org/10.1002/per.2132

Helms, J. E. (2006). Fairness is not validity or cultural bias in racial-group assessment: A quantitative perspective. *American Psychologist*, *61*(8), 845–859. https://doi.org/10.1037/0003-066X.61.8.845

Hidalgo, M. D., López-Martínez, M. D., Gómez-Benito, J., & Guilera, G. (2016). A comparison of discriminant logistic regression and Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning (IRTLRDIF) in polytomous short tests. *Psicothema*, *28*(1), 83–88. https://doi.org/10.7334/psicothema2015.142

Hidalgo, M. D., & López-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement*, *64*(6), 903–915. https://doi.org/10.1177/0013164403261769

Hill, N. E., & Torres, K. (2010). Negotiating the American dream: The paradox of aspirations and achievement among Latino students and engagement between their families and schools. *Journal of Social Issues*, *66*(1), 95–112. https://doi.org/10.1111/j.1540-4560.2009.01635.x

Hofer, J., Busch, H., Bender, M., Ming, L., & Hagemeyer, B. (2010). Arousal of achievement motivation among student samples in three different cultural contexts: Self and social standards of evaluation. *Journal of Cross-Cultural Psychology*, *41*(5–6), 758–775. https://doi.org/10.1177/0022022110375160

Hong, W., Bernacki, M. L., & Perera, H. N. (2020). A latent profile analysis of undergraduates' achievement motivations and metacognitive behaviors, and their relations to achievement in science. *Journal of Educational Psychology*, *112*(7), 1409–1430. https://doi.org/10.1037/edu0000445

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J.-A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research*, *62*(3), 333–353. https://doi.org/10.1080/00313831.2016.1258726

Hopkins, D. J., & King, G. (2010). Improving anchoring vignettes: Designing surveys to correct interpersonal incomparability. *Public Opinion Quarterly*, *74*(2), 201–222. https://doi.org/10.1093/poq/nfq011

Hopwood, C. J., Flato, C. G., Ambwani, S., Garland, B. H., & Morey, L. C. (2009). A comparison of Latino and Anglo socially desirable responding. *Journal of Clinical Psychology*, *65*(7), 769–780. https://doi.org/10.1002/jclp.20584

Huang, C. (2012). Discriminant and criterion-related validity of achievement goals in predicting academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*(1), 48–73. https://doi.org/10.1037/a0026223

Huang, C. (2013). Gender differences in academic self-efficacy: A meta-analysis. *European Journal of Psychology of Education*, *28*(1), 1–35. https://doi.org/10.1007/s10212-011-0097-y

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, *20*(3), 296–309. https://doi.org/10.1177/0022022189203004

Hulleman, C. S., Schrager, S. M., Bodmann, S. M., & Harackiewicz, J. M. (2010). A meta-analytic review of achievement goal measures: Different labels for the same constructs or different constructs with similar labels? *Psychological Bulletin*, *136*(3), 422–449. https://doi.org/10.1037/a0018947

Hyde, J. S., & Durik, A. M. (2005). Gender, competence, and motivation. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 375–391). Guilford Press.

IBM Corp. (2017). *IBM SPSS Statistics for Windows* (25.0) [Computer software]. IBM Corp.

Jin, K.-Y., & Chen, H.-F. (2019). MIMIC approach to assessing differential item functioning with control of extreme response style. *Behavior Research Methods*, *52*, 23–35. https://doi.org/10.3758/s13428-019-01198-1

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329–349. https://doi.org/10.1207/S15324818AME1404_2

Johanson, G. A. (1997). Differential item functioning in attitude assessment. *Evaluation Practice*, *18*(2), 127–135.

Johnson, M. K., Crosnoe, R., & Elder, G. H., Jr. (2001). Students' attachment and academic engagement: The role of race and ethnicity. *Sociology of Education*, *74*(4), 318–340. https://doi.org/10.2307/2673138

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology*, *36*(2), 264–277. https://doi.org/10.1177/0022022104272905

Jonas, K. G., & Markon, K. E. (2019). Modeling response style using vignettes and person-specific item response theory. *Applied Psychological Measurement*, *43*(1), 3–17. https://doi.org/10.1177/0146621618798663

Judge, T. A., & Kammeyer-Mueller, J. D. (2012). On the value of aiming high: The causes and consequences of ambition. *Journal of Applied Psychology*, *97*(4), 758–775. https://doi.org/10.1037/a0028084

Kane, M. T. (2010). Validity and fairness. *Language Testing*, *27*(2), 177–182. https://doi.org/10.1177/0265532209349467

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kane, M. T., & Bridgeman, B. (2017). Research on validity theory and practice at ETS. In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 489–552). Springer Open. https://doi.org/10.1007/978-3-319-58689-2_16

Kang, T., & Chen, T. T. (2008). Performance of the generalized *S-X* ^{2} item fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*(4), 391–406. https://doi.org/10.1111/j.1745-3984.2008.00071.x

Kaplan, A., & Maehr, M. L. (2007). The contributions and prospects of goal orientation theory. *Educational Psychology Review*, *19*(2), 141–184. https://doi.org/10.1007/s10648-006-9012-5

Kaplan, D., & Kuger, S. (2016). The methodology of PISA: Past, present, and future. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing Contexts of Learning* (pp. 53–73). Springer International Publishing. https://doi.org/10.1007/978-3-319-45357-6_3

Kapteyn, A., Smith, J. P., van Soest, A. H. O., & Vonkova, H. (2011). *Anchoring vignettes and response consistency* (WR-840). RAND Labor and Population.

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., de Groot, E., Gilbert, M. C., Musu, L., Kempler, T. M., & Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist*, *42*(3), 139–151. https://doi.org/10.1080/00461520701416231

Kemmelmeier, M. (2016). Cultural differences in survey responding: Issues and insights in the study of response biases. *International Journal of Psychology*, *51*(6), 439–444. https://doi.org/10.1002/ijop.12386

Kim, S.-H., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement*, *44*(2), 93–116. https://doi.org/10.1111/j.1745-3984.2007.00029.x

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, *98*(01), 191–207. https://doi.org/10.1017/S000305540400108X

King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis*, *15*(01), 46–66. https://doi.org/10.1093/pan/mpl011

King, R. B., McInerney, D. M., & Nasser, R. (2017). Different goals for different folks: A cross-cultural study of achievement goals across nine cultures. *Social Psychology of Education*, *20*(3), 619–642. https://doi.org/10.1007/s11218-017-9381-2

Kitayama, S. (2002). Culture and basic psychological processes--Toward a system view of culture: Comment on Oyserman et al. (2002). *Psychological Bulletin*, *128*(1), 89–96. https://doi.org/10.1037/0033-2909.128.1.89

Knott, R. J., Lorgelly, P. K., Black, N., & Hollingsworth, B. (2017). Differential item functioning in quality of life measurement: An analysis using anchoring vignettes. *Social Science & Medicine*, *190*, 247–255. https://doi.org/10.1016/j.socscimed.2017.08.033

Kriegbaum, K., Becker, N., & Spinath, B. (2018). The relative importance of intelligence and motivation as predictors of school achievement: A meta-analysis. *Educational Research Review*, *25*, 120–148. https://doi.org/10.1016/j.edurev.2018.10.001

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.

Kyllonen, P. C., & Bertling, J. P. (2013). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-sale assessment: Background, technical issues, and methods of data analysis* (pp. 278–286). Chapman and Hall/CRC.

Lai, J.-S., Teresi, J., & Gershon, R. (2005). Procedures for the analysis of differential item functioning (DIF) for small sample sizes. *Evaluation & the Health Professions*, *28*(3), 283–294. https://doi.org/10.1177/0163278705278276

Lalwani, A. K., Shrum, L. J., & Chiu, C. (2009). Motivated response styles: The role of cultural values, regulatory focus, and self-consciousness in socially desirable responding. *Journal of Personality and Social Psychology*, *96*(4), 870–882. https://doi.org/10.1037/a0014622

Lambert, M. C., Garcia, A. G., Epstein, M. H., & Cullinan, D. (2018). Differential item functioning of the Emotional and Behavioral Screener for Caucasian and African American elementary school students. *Journal of Applied School Psychology*, *34*(3), 201–214. https://doi.org/10.1080/15377903.2017.1345815

Lambert, M. C., Garcia, A. G., January, S.-A. A., & Epstein, M. H. (2018). The impact of English language learner status on screening for emotional and behavioral disorders: A differential item functioning (DIF) study. *Psychology in the Schools*, *55*(3), 229–239. https://doi.org/10.1002/pits.22103

Lee, C. S., Hayes, K. N., Seitz, J., DiStefano, R., & O'Connor, D. (2016). Understanding motivational structures that differentially predict engagement and achievement in middle school science. *International Journal of Science Education*, *38*(2), 192–215. https://doi.org/10.1080/09500693.2015.1136452

Liem, G. A. D., & Elliot, A. J. (2018). Sociocultural influences on achievement goal adoption and regulation: A goal complex perspective. In G. A. D. Liem & D. M. McInerney (Eds.), *Big theories revisited 2* (Vol. 12, pp. 41–68). Information Age Publishing.

Liem, G. A. D., Martin, A. J., Porter, A. L., & Colmar, S. (2012). Sociocultural antecedents of academic motivation and achievement: Role of values and achievement motives in achievement goals and academic performance. *Asian Journal of Social Psychology*, *15*(1), 1–13. https://doi.org/10.1111/j.1467-839X.2011.01351.x

Linnenbrink-Garcia, L., Patall, E. A., & Pekrun, R. (2016). Adaptive motivation and emotion in education: Research and principles for instructional design. *Policy Insights from the Behavioral and Brain Sciences*, *3*(2), 228–236. https://doi.org/10.1177/2372732216644450

Lu, Y., & Bolt, D. M. (2015). Examining the attitude-achievement paradox in PISA using a multilevel multidimensional IRT model for extreme response style. *Large-Scale Assessments in Education*, *3*(1), 2. https://doi.org/10.1186/s40536-015-0012-0

Maehr, M. L., & Meyer, H. A. (1997). Understanding motivation and schooling: Where we've been, where we are, and where we need to go. *Educational Psychology Review*, *9*, 371–409.

Magis, D., Raîche, G., Béland, S., & Gérard, P. (2011). A generalized logistic regression procedure to detect differential item functioning among multiple groups. *International Journal of Testing*, *11*(4), 365–386. https://doi.org/10.1080/15305058.2011.602810

Marksteiner, T., Kuger, S., & Klieme, E. (2019). The potential of anchoring vignettes to increase intercultural comparability of non-cognitive factors. *Assessment in Education: Principles, Policy & Practice*, *26*(4), 516–536. https://doi.org/10.1080/0969594X.2018.1514367

Markus, H. R. (2008). Pride, prejudice, and ambivalence: Toward a unified theory of race and ethnicity. *American Psychologist*, *63*(8), 651–670. https://doi.org/10.1037/0003-066X.63.8.651

Marsh, H. W., Craven, R. G., Hinkley, J. W., & Debus, R. L. (2003). Evaluation of the big-two-factor theory of academic motivation orientations: An evaluation of jingle-jangle fallacies. *Multivariate Behavioral Research*, *38*(2), 189–224. https://doi.org/10.1207/S15327906MBR3802_3

Marsh, H. W., Hau, K.-T., Artelt, C., Baumert, J., & Peschar, J. L. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing*, *6*(4), 311–360. https://doi.org/10.1207/s15327574ijt0604_1

Martinez, S., & Guzman, S. (2013). Gender and racial/ethnic differences in self-reported levels of engagement in high school math and science courses. *Hispanic Journal of Behavioral Sciences*, *35*(3), 407–427. https://doi.org/10.1177/0739986313495495

Martinková, P., Drabinová, A., Liaw, Y.-L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE—Life Sciences Education*, *16*(2), rm2. https://doi.org/10.1187/cbe.16-10-0307

McDaniel, M. A., Psotka, J., Legree, P. J., Yost, A. P., & Weekley, J. A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology*, *96*(2), 327–336. https://doi.org/10.1037/a0021983

McNamara, T., & Roever, C. (2006). Psychometric approaches to fairness: Bias and DIF. *Language Learning*, *56*(S2), 81–128.

Meece, J. L., Glienke, B. B., & Burg, S. (2006). Gender and motivation. *Journal of School Psychology*, *44*(5), 351–373. https://doi.org/10.1016/j.jsp.2006.04.004

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, *54*(1), 17–24.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749.

Michou, A., Vansteenkiste, M., Mouratidis, A., & Lens, W. (2014). Enriching the hierarchical model of achievement motivation: Autonomous and controlling reasons underlying achievement goals. *British Journal of Educational Psychology*, *84*(4), 650–666. https://doi.org/10.1111/bjep.12055

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement*, *30*(2), 107–122. https://doi.org/10.1111/j.1745-3984.1993.tb01069.x

Miller-Cotto, D., & Byrnes, J. P. (2016). Ethnic/racial identity and academic achievement: A meta-analytic review. *Developmental Review*, *41*, 51–70. https://doi.org/10.1016/j.dr.2016.06.003

Millsap, R. E. (2006). Comments on methods for the investigation of measurement bias in the Mini-Mental State Examination: *Medical Care*, *44*(11 Suppl 3), S171–S175. https://doi.org/10.1097/01.mlr.0000245441.76388.ff

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*(4), 297–334. https://doi.org/10.1177/014662169301700401

Min, I., Cortina, K. S., & Miller, K. F. (2016). Modesty bias and the attitude-achievement paradox across nations: A reanalysis of TIMSS. *Learning and Individual Differences*, *51*, 359–366. https://doi.org/10.1016/j.lindif.2016.09.008

Mojtabai, R. (2016). Depressed mood in middle-aged and older adults in Europe and the United States: A comparative study using anchoring vignettes. *Journal of Aging and Health*, *28*(1), 95–117. https://doi.org/10.1177/0898264315585506

Moors, G. (2004). Facts and artefacts in the comparison of attitudes among ethnic minorities: A multigroup latent class structure model with adjustment for response style behavior. *European Sociological Review*, *20*(4), 303–320.

Morren, M., Gelissen, J., & Vermunt, J. (2011). The impact of controlling for extreme responding on measurement equivalence in cross-cultural research. *Methodology*, *8*(4), 159–170.

Mõttus, R., Allik, J., Realo, A., Rossier, J., Zecca, G., Ah-Kion, J., Amoussou-Yéyé, D., Bäckström, M., Barkauskiene, R., Barry, O., Bhowon, U., Björklund, F., Bochaver, A., Bochaver, K., de Bruin, G., Cabrera, H. F., Chen, S. X., Church, A. T., Cissé, D. D., … Johnson, W. (2012). The effect of response style on self-reported conscientiousness across 20 countries. *Personality and Social Psychology Bulletin*, *38*(11), 1423–1436. https://doi.org/10.1177/0146167212451275

Murayama, K., & Elliot, A. J. (2012). The competition–performance relation: A meta-analytic review and test of the opposing processes model of competition and performance. *Psychological Bulletin*, *138*(6), 1035–1070. https://doi.org/10.1037/a0028324

Murdock, T. B. (2009). Achievement motivation in racial and ethnic context. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 433–462). Routledge.

Muthén, L. K., & Muthén, B. O. (2017). *M Plus* (Version 8) [Computer software]. Muthén & Muthén.

Nisbet, I., & Shaw, S. D. (2019). Fair assessment viewed through the lenses of measurement theory. *Assessment in Education: Principles, Policy & Practice*, *26*(5), 612–629. https://doi.org/10.1080/0969594X.2019.1586643

Nolen, S. B., Horn, I. S., & Ward, C. J. (2015). Situating motivation. *Educational Psychologist*, *50*(3), 234–247. https://doi.org/10.1080/00461520.2015.1075399

OECD. (2013). *PISA 2012 results: What makes schools successful?* (No. 4; Annex A6: Anchoring Vignettes in PISA 2012 Student Questionnaire). OECD Publishing.

OECD. (2014). *PISA 2012 technical report*. OECD Publishing.

OECD. (2017). *PISA 2015 technical report*. OECD Publishing.

OECD. (2019). *PISA 2018 assessment and analytical framework*. OECD Publishing.

Oliveri, M. E., Ercikan, K., Lyons-Thomas, J., & Holtzman, S. (2016). Analyzing fairness among linguistic minority populations using a latent class differential item functioning approach. *Applied Measurement in Education*, *29*(1), 17–29. https://doi.org/10.1080/08957347.2015.1102913

Oliveri, M. E., Ercikan, K., Zumbo, B. D., & Lawless, R. (2014). Uncovering substantive patterns in student responses in international large-scale assessments—Comparing a latent

class to a manifest DIF approach. *International Journal of Testing*, *14*(3), 265–287. https://doi.org/10.1080/15305058.2014.891223

Oliveri, M. E., Olson, B. F., Ercikan, K., & Zumbo, B. D. (2012). Methodologies for investigating item- and test-level measurement equivalence in international large-scale assessments. *International Journal of Testing*, *12*(3), 203–223. https://doi.org/10.1080/15305058.2011.617475

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64. https://doi.org/10.1177/01466216000241003

Oshima, T. C., & Morris, S. B. (2008). Raju's differential functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice*, *27*(3), 43–50. https://doi.org/10.1111/j.1745-3992.2008.00127.x

Oshima, T. C., Wright, K., & White, N. (2015). Multiple-group noncompensatory differential item functioning in Raju's differential functioning of items and tests. *International Journal of Testing*, *15*(3), 254–273. https://doi.org/10.1080/15305058.2015.1009980

Osterlind, S. J., & Everson, H. T. (2010). *Differential item functioning* (2nd ed.). Sage Publications.

Oyserman, D., & Destin, M. (2010). Identity-based motivation: Implications for intervention. *The Counseling Psychologist*, *38*(7), 1001–1043. https://doi.org/10.1177/0011000010374775

Paccagnella, O. (2013). A new tool for measuring customer satisfaction: The anchoring vignette approach. *Statistica Applicata*, *23*(3), 1–18.

Paulhus, D. L. (1991). Measurement and control of response style. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes*. Academic Press.

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement*, *44*(3), 187–210. https://doi.org/10.1111/j.1745-3984.2007.00034.x

Penfield, R. D., Alvarez, K., & Lee, O. (2008). Using a taxonomy of differential step functioning to improve the interpretation of DIF in polytomous items: An illustration. *Applied Measurement in Education*, *22*(1), 61–78. https://doi.org/10.1080/08957340802558367

Penfield, R. D., Gattamorta, K., & Childs, R. A. (2009). An NCME instructional module on using differential step functioning to refine the analysis of DIF in polytomous items. *Educational Measurement: Issues and Practice*, *28*(1), 38–49. https://doi.org/10.1111/j.1745-3992.2009.01135.x

Pepper, D., Hodgen, J., Lamesoo, K., Kõiv, P., & Tolboom, J. (2018). Think aloud: Using cognitive interviewing to validate the PISA assessment of student self-efficacy in mathematics. *International Journal of Research & Method in Education*, *41*(1), 3–16. https://doi.org/10.1080/1743727X.2016.1238891

Peracchi, F., & Rossetti, C. (2012). Heterogeneity in health responses and anchoring vignettes. *Empirical Economics*, *42*(2), 513–538. https://doi.org/10.1007/s00181-011-0530-8

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research*, *21*(2), 381–391. https://doi.org/10.1086/209405

Poortinga, Y. H. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology*, *24*(6), 737–756.

Primi, R., Santos, D., John, O. P., De Fruyt, F., & Hauck-Filho, N. (2018). Dealing with person differential item functioning in social-emotional skill assessment using anchoring vignettes. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative Psychology* (Vol. 233, pp. 275–286). Springer International Publishing. https://doi.org/10.1007/978-3-319-77249-3_23

Primi, R., Zanon, C., Santos, D., De Fruyt, F., & John, O. P. (2016). Anchoring vignettes: Can they make adolescent self-reports of social-emotional skills more reliable, discriminant, and criterion-valid? *European Journal of Psychological Assessment*, *32*(1), 39–51. https://doi.org/10.1027/1015-5759/a000336

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*(3), 517–529. https://doi.org/10.1037/0021-9010.87.3.517

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Gershon, R., Reise, S. P., Lai, J.-S., & Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, *45*(5 Suppl 1), S22–S31.

Reise, S. P. (2014). Item response theory. In R. L. Cautin & S. O. Lilienfeld (Eds.), *The encyclopedia of clinical psychology* (pp. 1–10). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118625392.wbecp357

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552–566.

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, *26.1*, 108–116. https://doi.org/10.7334/psicothema2013.260

Rizopoulos, D. (2006). ltm: An *R* package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5). https://doi.org/10.18637/jss.v017.i05

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes?: A meta-analysis. *Psychological Bulletin*, *130*(2), 261–288. https://doi.org/10.1037/0033-2909.130.2.261

Robitzsch, A., & Rupp, A. A. (2009). Impact of missing data on the detection of differential item functioning: The case of Mantel-Haenszel and logistic regression analysis. *Educational and Psychological Measurement*, *69*(1), 18–34. https://doi.org/10.1177/0013164408318756

Rutkowski, L., Gonzalez, E., von Davier, M., & Zhou, Y. (2013). Assessment design for international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 75–96). Chapman and Hall/CRC.

Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher*, *45*(4), 252–257. https://doi.org/10.3102/0013189X16649961

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*(1), 31–57. https://doi.org/10.1177/0013164413498257

Samejima, F. (2010). The general graded response model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 77–108). Routledge.

Sandilands, D., Oliveri, M. E., Zumbo, B. D., & Ercikan, K. (2013). Investigating sources of differential item functioning in international large-scale assessments using a confirmatory approach. *International Journal of Testing*, *13*(2), 152–174. https://doi.org/10.1080/15305058.2012.690140

Sari, H. I., & Huggins, A. C. (2015). Differential item functioning detection across two methods of defining group comparisons: Pairwise and composite group comparisons. *Educational and Psychological Measurement*, *75*(4), 648–676. https://doi.org/10.1177/0013164414549764

Schunk, D. H., Meece, J. R., & Pintrich, P. R. (2014). Motivation: Introduction and historical foundations. In D. H. Schunk, J. R. Meece, & P. R. Pintrich (Eds.), *Motivation in education: Theory, research, and applications* (4th ed., pp. 1–50). Pearson.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., Gundy, C., Koller, M., Petersen, M. A., & Sprangers, M. A. G. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology*, *62*(3), 288–295. https://doi.org/10.1016/j.jclinepi.2008.06.003

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., Gundy, C., Koller, M., Petersen, M. A., Sprangers, M. A. G., & the EORTC Quality of Life Group and the Quality of Life Cross-Cultural Meta-Analysis Group. (2010). Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. *Health and Quality of Life Outcomes*, *8*, 81. https://doi.org/10.1186/1477-7525-8-81

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., Koller, M., Petersen, M. A., & Sprangers, M. A. G. (2007). The use of differential item functioning analyses to identify cultural differences in responses to the EORTC QLQ-C30. *Quality of Life Research*, *16*(1), 115–129. https://doi.org/10.1007/s11136-006-9120-1

Segeritz, M., & Pant, H. A. (2013). Do they feel the same way about math?: Testing measurement invariance of the PISA "Students' Approaches to Learning" instrument across immigrant groups within Germany. *Educational and Psychological Measurement*, *73*(4), 601–630. https://doi.org/10.1177/0013164413481802

Senko, C., & Dawson, B. (2017). Performance-approach goal effects depend on how they are defined: Meta-analytic evidence from multiple educational outcomes. *Journal of Educational Psychology*, *109*(4), 574–598. https://doi.org/10.1037/edu0000160

Senko, C., & Hulleman, C. S. (2013). The role of goal attainment expectancies in achievement goal pursuit. *Journal of Educational Psychology*, *105*(2), 504–521. https://doi.org/10.1037/a0031136

Senko, C., Hulleman, C. S., & Harackiewicz, J. M. (2011). Achievement goal theory at the crossroads: Old controversies, current challenges, and new directions. *Educational Psychologist*, *46*(1), 26–47. https://doi.org/10.1080/00461520.2011.538646

Sharkness, J. (2014). Item response theory: Overview, applications, and promise for institutional research. *New Directions for Institutional Research*, *2014*(161), 41–58. https://doi.org/10.1002/ir.20066

Shernoff, D. J., & Schmidt, J. A. (2008). Further evidence of an engagement–achievement paradox among U.S. high school students. *Journal of Youth and Adolescence*, *37*(5), 564–580. https://doi.org/10.1007/s10964-007-9241-z

Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, *19*(2–3), 170–187. https://doi.org/10.1080/13803611.2013.767621

Smith, J. L. (2004). Understanding the process of stereotype threat: A review of mediational variables and new performance goal directions. *Educational Psychology Review*, *16*(3), 177–206. https://doi.org/10.1023/B:EDPR.0000034020.20317.89

Solon, G., Haider, S. J., & Woolridge, J. M. (2015). What are we weighting for? *The Journal of Human Resources*, *50*(2), 301–316.

Song, H., Cai, H., Brown, J. D., & Grimm, K. J. (2011). Differential item functioning of the Rosenberg Self-Esteem Scale in the US and China: Measurement bias matters: Cross-cultural equivalence of self-esteem. *Asian Journal of Social Psychology*, *14*(3), 176–188. https://doi.org/10.1111/j.1467-839X.2011.01347.x

Stankov, L., Lee, J., & von Davier, M. (2018). A note on construct validity of the anchoring method in PISA 2012. *Journal of Psychoeducational Assessment*, *36*(7), 709–724. https://doi.org/10.1177/0734282917702270

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology*, *89*(3), 497–508. https://doi.org/10.1037/0021-9010.89.3.497

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods*, *11*(4), 402–415. https://doi.org/10.1037/1082-989X.11.4.402

Steinmayr, R., & Spinath, B. (2008). Sex differences in school achievement: What are the roles of personality and achievement motivation? *European Journal of Personality*, *22*(3), 185–209. https://doi.org/10.1002/per.676

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*(2), 293–325.

Su, Y.-H., & Wang, W.-C. (2005). Efficiency of the Mantel, generalized Mantel–Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education*, *18*(4), 313–350. https://doi.org/10.1207/s15324818ame1804_1

Sue, S. (1996). Measurement, testing, and ethnic bias: Can solutions be found? In G. R. Sodowsky & J. C. Impara (Eds.), *Multicultural assessment in counseling and clinical psychology* (pp. 7–36). Buros Institute of Mental Measurements.

Suh, Y. (2016). Effect size measures for differential item functioning in a multidimensional IRT model: DIF effect sizes for MIRT. *Journal of Educational Measurement*, *53*(4), 403–430. https://doi.org/10.1111/jedm.12123

Svetina, D., & Rutkowski, L. (2017). Multidimensional measurement invariance in an international context: Fit measure performance with many groups. *Journal of Cross-Cultural Psychology*, *48*(7), 991–1008. https://doi.org/10.1177/0022022117717028

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

Tan, J. A., & Hall, R. J. (2005). The effects of social desirability bias on applied measures of goal orientation. *Personality and Individual Differences*, *38*(8), 1891–1902. https://doi.org/10.1016/j.paid.2004.11.015

Teresi, J. A., & Jones, R. N. (2016). Methodological issues in examining measurement equivalence in patient reported outcomes measures: Methods overview to the two-part series, "Measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) short forms." *Psychological Test and Assessment Modeling*, *58*(1), 37–78.

Teresi, J. A., Ramirez, M., Jones, R. N., Choi, S., & Crane, P. K. (2012). Modifying measures based on differential item functioning (DIF) impact analyses. *Journal of Aging and Health*, *24*(6), 1044–1076. https://doi.org/10.1177/0898264312436877

Thoman, D. B., Smith, J. L., Brown, E. R., Chase, J., & Lee, J. Y. K. (2013). Beyond performance: A motivational experiences model of stereotype threat. *Educational Psychology Review*, *25*(2), 211–243. https://doi.org/10.1007/s10648-013-9219-1

Tourangeau, R. (2018). The survey response process from a cognitive viewpoint. *Quality Assurance in Education*, *26*(2), 169–181. https://doi.org/10.1108/QAE-06-2017-0034

Trumbull, E., & Rothstein-Fisch, C. (2011). The intersection of culture and achievement motivation. *The School Community Journal*, *21*(2), 25–53.

Turner, R. C., & Keiffer, E. A. (2019). Impact of unbalanced DIF item proportions on group-specific DIF identification. *Communications in Statistics - Theory and Methods*, *48*(15), 3746–3760. https://doi.org/10.1080/03610926.2018.1481968

Urdan, T. C., & Bruchmann, K. (2018). Examining the academic motivation of a diverse student population: A consideration of methodology. *Educational Psychologist*, *53*(2), 114–130. https://doi.org/10.1080/00461520.2018.1440234

van der Sluis, S., Vinkhuyzen, A. A. E., Boomsma, D. I., & Posthuma, D. (2010). Sex differences in adults' motivation to achieve. *Intelligence*, *38*(4), 433–446. https://doi.org/10.1016/j.intell.2010.04.004

van Soest, A., & Vonkova, H. (2014). Testing the specification of parametric models by using anchoring vignettes: *Testing the Specification of Parametric Models*. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *177*(1), 115–133. https://doi.org/10.1111/j.1467-985X.2012.12000.x

Vantieghem, W., & Van Houtte, M. (2018). Differences in study motivation within and between genders: An examination by gender typicality among early adolescents. *Youth & Society*, *50*(3), 377–404. https://doi.org/10.1177/0044118X15602268

Vaske, J. J., Beaman, J., & Sponarski, C. C. (2017). Rethinking internal consistency in Cronbach's alpha. *Leisure Sciences*, *39*(2), 163–173. https://doi.org/10.1080/01490400.2015.1127189

von Davier, M., Shin, H.-J., Khorramdel, L., & Stankov, L. (2018). The effects of vignette scoring on reliability and validity of self-reports. *Applied Psychological Measurement*, *42*(4), 291–306. https://doi.org/10.1177/0146621617730389

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, *140*(4), 1174–1204. https://doi.org/10.1037/a0036620

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment*, *29*(4), 364–376. https://doi.org/10.1177/0734282911406666

Wand, J., King, G., & Lau, O. (2011). **anchors**: Software for anchoring vignette data. *Journal of Statistical Software*, *42*(3). https://doi.org/10.18637/jss.v042.i03

Warnecke, R. B., Johnson, T. P., Chávez, N., Sudman, S., O'Rourke, D. P., Lacey, L., & Horm, J. (1997). Improving question wording in surveys of culturally diverse populations. *Annals of Epidemiology*, *7*(5), 334–342. https://doi.org/10.1016/S1047-2797(97)00030-6

Weiss, S., & Roberts, R. D. (2018). Using anchoring vignettes to adjust self-reported personality: A comparison between countries. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00325

Westbrook, L., & Saperstein, A. (2015). New categories are not enough: Rethinking the measurement of sex and gender in social surveys. *Gender & Society*, *29*(4), 534–560. https://doi.org/10.1177/0891243215584758

Wetzel, E., Böhnke, J. R., Carstensen, C. H., Ziegler, M., & Ostendorf, F. (2013). Do individual response styles matter?: Assessing differential item functioning for men and women in the NEO-PI-R. *Journal of Individual Differences*, *34*(2), 69–81. https://doi.org/10.1027/1614-0001/a000102

Wetzel, E., & Carstensen, C. H. (2017). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment*, *33*(5), 352–364. https://doi.org/10.1027/1015-5759/a000291

Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality*, *47*(2), 178–189. https://doi.org/10.1016/j.jrp.2012.10.010

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2016). The stability of extreme response style and acquiescence over 8 years. *Assessment*, *23*(3), 279–291. https://doi.org/10.1177/1073191115583714

Wigfield, A., Eccles, J. S., Roeser, R. W., & Schiefele, U. (2008). Development of achievement motivation. In W. Damon & R. M. Lerner (Eds.), *Child and adolescent development: An advanced course* (pp. 406–434). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470147658.chpsy0315

Wilson, T. M., Zheng, C., Lemoine, K. A., Martin, C. P., & Tang, Y. (2016). Achievement goals during middle childhood: Individual differences in motivation and social adjustment. *The Journal of Experimental Education*, *84*(4), 723–743. https://doi.org/10.1080/00220973.2015.1094648

Wood, D., & Graham, S. (2010). Why race matters: Social context and achievement motivation in African American youth. In T. C. Urdan & S. A. Karabenick (Eds.), *The decade ahead: Applications and contexts of motivation and achievement* (1. ed, pp. 175–210). Emerald Publishing.

Wormington, S. V., & Linnenbrink-Garcia, L. (2017). A new look at multiple goal pursuit: The promise of a person-centered approach. *Educational Psychology Review*, *29*(3), 407–445. https://doi.org/10.1007/s10648-016-9358-2

Wu, A. D., & Ercikan, K. (2006). Using multiple-variable matching to identify cultural sources of differential item functioning. *International Journal of Testing*, *6*(3), 287–300. https://doi.org/10.1207/s15327574ijt0603_5

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–170. https://doi.org/10.1177/0265532209349465

Zhong, Q., Gelaye, B., Fann, J. R., Sanchez, S. E., & Williams, M. A. (2014). Cross-cultural validity of the Spanish version of PHQ-9 among pregnant Peruvian women: A Rasch item response theory analysis. *Journal of Affective Disorders*, *158*, 148–153. https://doi.org/10.1016/j.jad.2014.02.012

Ziegler, M. (2015). "F*** you, I won't do what you told me!" – response biases as threats to psychological assessment. *European Journal of Psychological Assessment*, *31*(3), 153–158. https://doi.org/10.1027/1015-5759/a000292

Ziegler, M., & Brunner, M. (2016). Test standards and psychometric modeling. In A. A. Lipnevich, F. Preckel, & R. D. Roberts (Eds.), *Psychosocial skills and school systems in the 21st century: Theory, research, and practice* (pp. 29–55). Springer.

Zopluoglu, C., & Davenport, E. C., Jr. (2017). A note on using eigenvalues in dimensionality assessment. *Practical Assessment, Research & Evaluation*, *22*(7), 11.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Directorate of Human Resources Research and Evaluation, Department of National Defense.

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

Zusho, A., Pintrich, P. R., & Cortina, K. S. (2005). Motives, goals, and adaptive patterns of performance in Asian American and Anglo American students. *Learning and Individual Differences*, *15*(2), 141–158. https://doi.org/10.1016/j.lindif.2004.11.003

Zwick, R. (2019). Fairness in measurement and selection: Statistical, philosophical, and public perspectives. *Educational Measurement: Issues and Practice*, *38*(4), 34–41. https://doi.org/10.1111/emip.12299

APPENDICES

APPENDIX A

Distribution of Nonparametric Responses to Anchoring Vignettes

Y_i	Observed Order	Rating Pattern
1	$Y_i < Z_1 < Z_2 < Z_3$	Ordered
1	$Y_i < Z_2 < Z_1 < Z_3$	Misordered
1	$Y_i < Z_3 < Z_1 < Z_2$	Misordered
1	$Y_i < Z_1 = Z_2 < Z_3$	Tied
1	$Y_i < Z_1 = Z_3 < Z_2$	Tied, Misordered
1	$Y_i < Z_1 < Z_2 = Z_3$	Tied
1	$Y_i < Z_1 = Z_2 = Z_3$	Tied
1	$Y_i < Z_1 < Z_3 < Z_2$	Tied, Misordered
1	$Y_i < Z_2 < Z_3 < Z_1$	Misordered
1	$Y_i < Z_3 < Z_2 < Z_1$	Misordered
1	$Y_i < Z_3 < Z_1 = Z_2$	Tied, Misordered
1	$Y_i < Z_2 < Z_1 = Z_3$	Tied, Misordered
1	$Y_i < Z_2 = Z_3 < Z_1$	Tied, Misordered
2	$Y_i = Z_1 < Z_2 < Z_3$	Ordered
2	$Y_i = Z_1 < Z_2 = Z_3$	Tied
2	$Y_i = Z_1 < Z_3 < Z_2$	Misordered
3	$Z_1 < Y_i < Z_2 < Z_3$	Ordered
3	$Z_1 < Y_i < Z_2 = Z_3$	Tied
3	$Z_1 < Y_i < Z_3 < Z_2$	Tied, Misordered
4	$Z_1 < Y_i = Z_2 < Z_3$	Ordered
5	$Z_1 < Z_2 < Y_i < Z_3$	Ordered
5	$Z_2 < Z_1 < Y_i < Z_3$	Misordered
5	$Z_1 = Z_2 < Y_i < Z_3$	Tied
6	$Z_1 < Z_2 < Y_i = Z_3$	Ordered
6	$Z_2 < Z_1 < Y_i = Z_3$	Tied, Misordered
6	$Z_1 = Z_2 < Y_i = Z_3$	Tied
7	$Z_1 < Z_2 < Z_3 < Y_i$	Ordered
7	$Z_2 < Z_1 < Z_3 < Y_i$	Misordered
7	$Z_3 < Z_1 < Z_2 < Y_i$	Misordered
7	$Z_1 = Z_2 < Z_3 < Y_i$	Tied
7	$Z_1 = Z_3 < Z_2 < Y_i$	Tied, Misordered
7	$Z_1 < Z_2 = Z_3 < Y_i$	Tied
7	$Z_1 = Z_2 = Z_3 < Y_i$	Tied
7	$Z_1 < Z_3 < Z_2 < Y_i$	Misordered
7	$Z_2 < Z_3 < Z_1 < Y_i$	Misordered
7	$Z_3 < Z_2 < Z_1 < Y_i$	Misordered
7	$Z_3 < Z_1 = Z_2 < Y_i$	Misordered
7	$Z_2 < Z_1 = Z_3 < Y_i$	Tied, Misordered
7	$Z_2 = Z_3 < Z_1 < Y_i$	Tied, Misordered

Y_i	Observed Order	Rating Pattern
1, 2, 3, 4	$Y_i = Z_2 < Z_1 < Z_3$	Misordered
1, 2, 3, 4	$Y_i = Z_2 < Z_3 < Z_1$	Misordered
1, 2, 3, 4	$Y_i = Z_2 < Z_1 = Z_3$	Tied, Misordered
1, ,2, 3, 4, 5	$Z_2 < Y_i < Z_1 < Z_3$	Misordered
1, 2, 3, 4, 5	$Z_2 < Y_i < Z_3 < Z_1$	Misordered
1, 2, 3, 4, 5	$Z_2 < Y_i < Z_1 = Z_3$	Tied, Misordered
1, 2, 3, 4, 5, 6	$Y_i = Z_3 < Z_1 < Z_2$	Misordered
1, 2, 3, 4, 5, 6	$Z_2 < Y_i = Z_3 < Z_1$	Misordered
1, 2, 3, 4, 5, 6	$Y_i = Z_3 < Z_2 < Z_1$	Misordered
1, 2, 3, 4, 5, 6	$Y_i = Z_3 < Z_1 = Z_2$	Tied, Misordered
1, 2, 3, 4, 5, 6	$Y_i = Z_2 = Z_3 < Z_1$	Tied, Misordered
1, 2, 3, 4, 5, 6, 7	$Z_3 < Y_i < Z_1 < Z_2$	Misordered
1, 2, 3, 4, 5, 6, 7	$Z_1 < Z_3 < Y_i < Z_1$	Misordered
1, 2, 3, 4, 5, 6, 7	$Z_3 < Y_i < Z_2 < Z_1$	Misordered
1, 2, 3, 4, 5, 6, 7	$Z_3 < Y_i = Z_2 < Z_1$	Misordered
1, 2, 3, 4, 5, 6, 7	$Z_3 < Z_2 < Y_i < Z_1$	Misordered
1, 2, 3, 4, 5, 6, 7	$Z_3 < Y_i < Z_1 = Z_2$	Tied, Misordered
1, 2, 3, 4, 5, 6, 7	$Z_2 = Z_3 < Y_i < Z_1$	Tied, Misordered
2, 3, 4	$Y_i = Z_1 = Z_2 < Z_3$	Tied
2, 3, 4, 5	$Z_2 < Y_i = Z_1 < Z_3$	Tied, Misordered
2, 3, 4, 5, 6	$Y_i = Z_1 = Z_3 < Z_2$	Tied, Misordered
2, 3, 4, 5, 6	$Z_2 < Y_i = Z_1 = Z_3$	Tied, Misordered
2, 3, 4, 5, 6	$Y_i = Z_1 = Z_2 = Z_3$	Tied
2, 3, 4, 5, 6, 7	$Z_3 < Y_i = Z_1 < Z_2$	Misordered
2, 3, 4, 5, 6, 7	$Z_2 < Z_3 < Y_i = Z_1$	Misordered
2, 3, 4, 5, 6, 7	$Z_3 < Z_2 < Y_i = Z_1$	Misordered
2, 3, 4, 5, 6, 7	$Z_3 < Y_i = Z_1 = Z_2$	Misordered
2, 3, 4, 5, 6, 7	$Z_2 = Z_3 < Y_i = Z_1$	Tied, Misordered
3, 4, 5, 6	$Z_1 < Y_i = Z_3 < Z_2$	Misordered
3, 4, 5, 6, 7	$Z_3 < Z_1 < Y_i < Z_2$	Misordered
3, 4, 5, 6, 7	$Z_1 = Z_3 < Y_i < Z_2$	Tied, Misordered
3, 4, 5, 6, 7	$Z_1 < Z_3 < Y_i < Z_2$	Misordered
4, 5, 6	$Z_1 < Y_i = Z_2 = Z_3$	Tied, Misordered
4, 5, 6, 7	$Z_3 < Z_1 < Y_i = Z_2$	Misordered
4, 5, 6, 7	$Z_1 = Z_3 < Y_i = Z_2$	Tied, Misordered
4, 5, 6, 7	$Z_1 < Z_3 < Y_i = Z_2$	Misordered

Note. Adapted from Soest and Vonkova (2014).

APPENDIX B

Description of *Lordif* Software DIF Algorithm

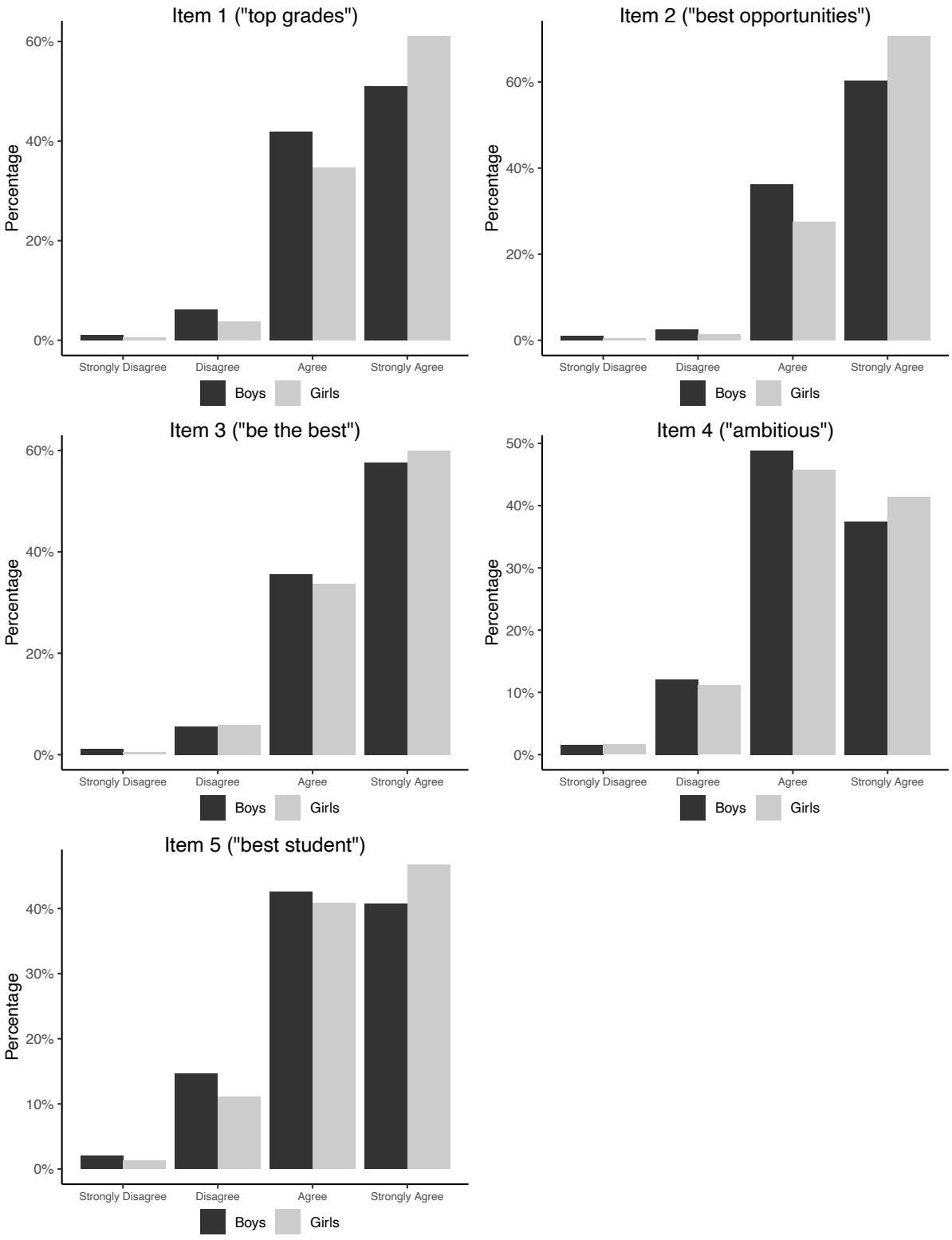
All DIF analyses were completed in the *lordif* package (Choi et al., 2011), which utilizes an iterative hybrid item response theory/logistic regression framework. The package can estimate uniform, nonuniform, and total effect DIF and offers multiple visual approaches to detecting DIF (e.g., item characteristic functions, response category functions, etc.). DIF is detected through an iterative process consisting of the following steps. First, item and person parameters are generated using a GRM. The *lordif* package generates the GRM estimate through the *ltm* package (Rizopoulos, 2006), which utilizes marginal maximum likelihood estimation and assumes data to be missing-at-random. *Lordif* reports the fit of the GRM estimates as an $S - \chi^2$ statistic; Ames and Penfield (2015) suggested that when $S - \chi^2$ is not significant (i.e., $p > .05$), item fit is better. *Lordif* treats omitted responses as not present. Second, ordinal logistic regression models are fit predicting item responses as a function of person ability (i.e., θ) and group membership (i.e., gender or ethnicity). Logistic regressions in *lordif* are fit using the *Design* package, which implements item-wise regressions (Harrell Jr. 2009). *Design* manages grouping variables with more than two levels (e.g., White, Hispanic/Latinx, Asian) by entering the grouping variable into a model as a set of dummy variables.

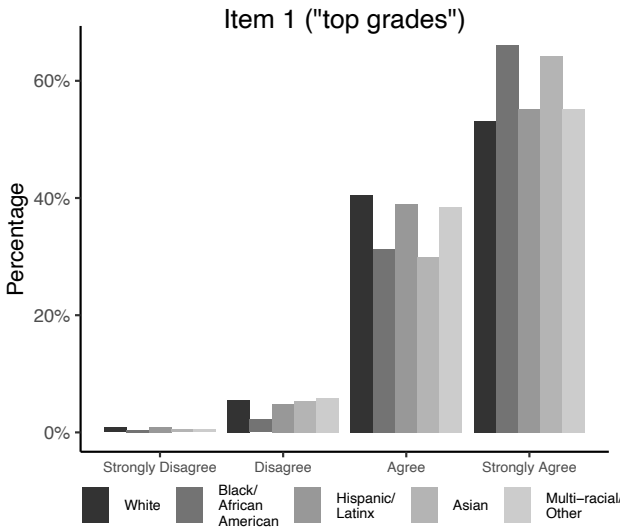
Third, θ estimates are re-calibrated (i.e., purified) to adjust for the effect of DIF on GRM trait estimates (i.e., possibly leading to possible false-positives or false-negatives). To do this, *lordif* implements an iterative process that involves re-calibrating θ estimates using group-specific IRT item parameter estimates for items that were initially detected as having DIF. First, a sparse response matrix is used to split response vectors into sparse vectors containing responses for only each group. Second, the GRM is re-fit to obtain item parameter estimates for non-DIF

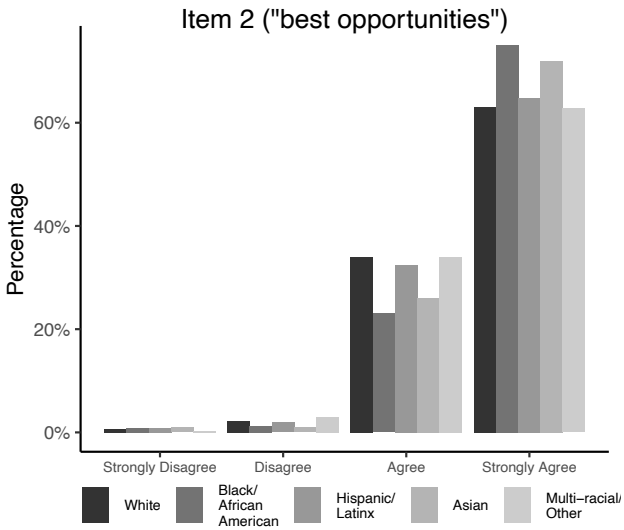
items and group-specific item parameter estimates for DIF items. Third, in order to identify the DIF impact, the sparse matrix estimates are equated with the original GRM estimates using the Stocking and Lord procedure. Fourth, ability estimates are updated and then used in a new set of logistic regressions. If items not initially flagged for DIF are detected as having DIF with the updated estimates, the process of identifying and re-calibrating items (using DIF-free estimates) is repeated until an item has been detected as having DIF on two consecutive model runs. Finally, in the last step of the DIF analysis, the impact of DIF (on items flagged as demonstrating DIF) is evaluated. Monte Carlo simulations are used to identify empirically-derived thresholds, as defined by being cut off at the most extreme (i.e., alpha) end of its cumulative distribution, for each DIF statistic and effect size. Then, the impact of DIF is identified by comparing the adjusted (i.e., purified) and unadjusted (i.e., naïve) θ estimates.

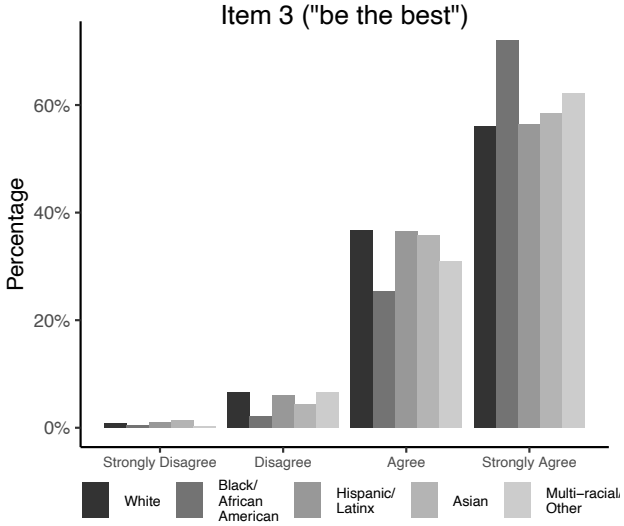
APPENDIX C

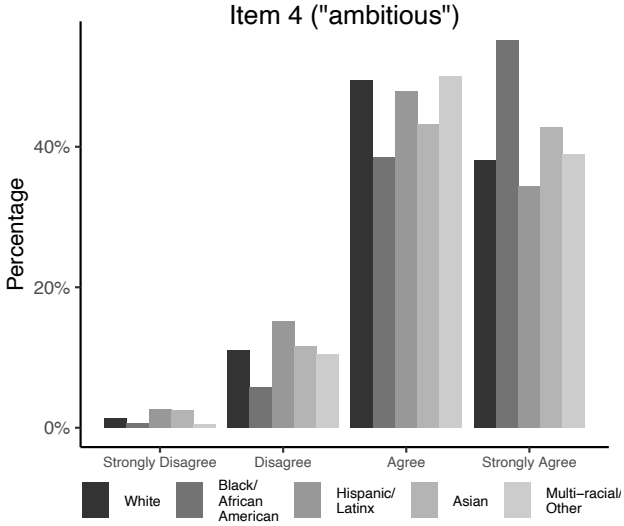
Response Distribution for Achievement Motivation Items, by Group

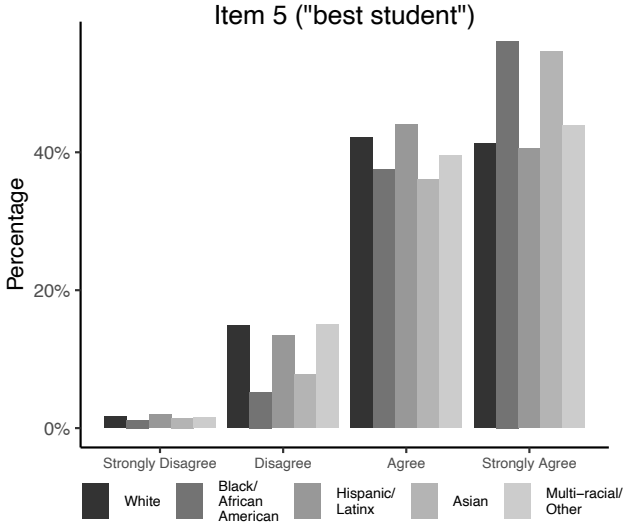



Item 1 ("top grades")
Percentage
60%
40%
20%
0%
Strongly Disagree
Disagree
Agree
Strongly Agree
White
Black/ African American
Hispanic/ Latinx
Asian
Multi−racial/ Other


Item 2 ("best opportunities")
Percentage
60%
40%
20%
0%
Strongly Disagree
Disagree
Agree
Strongly Agree
White
Black/ African American
Hispanic/ Latinx
Asian
Multi−racial/ Other


Item 3 ("be the best")
Percentage
60%
40%
20%
0%
Strongly Disagree
Disagree
Agree
Strongly Agree
White
Black/ African American
Hispanic/ Latinx
Asian
Multi−racial/ Other


Item 4 ("ambitious")
Percentage
40%
20%
0%
Strongly Disagree
Disagree
Agree
Strongly Agree
White
Black/ African American
Hispanic/ Latinx
Asian
Multi−racial/ Other


Item 5 ("best student")
Percentage
40%
20%
0%
Strongly Disagree
Disagree
Agree
Strongly Agree
White
Black/ African American
Hispanic/ Latinx
Asian
Multi−racial/ Other

APPENDIX D

Group-Specific Item Parameters, Before and After Vignette Adjustments

Gender Item Parameters

Item	Groups	$a,$	b_1	b_2	b_3	b_4
Item 1						
	Unadjusted*					
	Boys	3.39	-2.61	-1.62	-0.09	
	Girls	3.29	-2.96	-1.88	-0.31	
	Adjusted*					
	Boys	3.91	-2.62	-1.61	-0.29	1.56
	Girls	3.56	-3.03	-1.76	-0.44	1.63
Item 2						
	Unadjusted*					
	Boys	3.94	-2.54	-1.91	-0.35	
	Girls	3.47	-3.04	-2.25	-0.60	
	Adjusted*					
	Boys	4.62	-2.75	-1.68	-0.52	1.44
	Girls	3.87	-3.25	-1.89	-0.70	1.53
Item 3						
	Unadjusted					
	Boys	2.60	-2.76	-1.76	-.26	
	Girls	2.85	-3.22	-1.75	-.32	
	Adjusted*					
	Boys	3.16	-2.94	-1.70	-0.44	1.50
	Girls	3.34	-3.07	-1.75	-0.40	1.69
Item 5						
	Unadjusted*					
	Boys	2.82	-2.40	-1.17	0.22	
	Girls	3.03	-2.54	-1.30	0.10	
	Adjusted					
	Boys	3.04	-2.57	-1.39	0.00	1.78
	Girls	3.19	-2.79	-1.58	-0.11	1.94

* $p < .01$.

Multiple-Group DIF Item Parameters

Item	Group	a,	b_1	b_2	b_3
Item 2					
	Unadjusted				
	Base Group	4.92	-0.44		
	White	3.60	-0.41		
	Black/African American	3.52	-0.76		
	Hispanic/Latinx	3.53	-0.45		
	Asian	2.76	-0.77		
	Multi-racial/Other	3.09	-0.41		
	Adjusted*				
	Base Group	4.03	-1.77	-0.61	1.52
	White	4.88	-1.87	-0.58	1.52
	Black/African American	4.18	-1.52	-0.66	1.56
	Hispanic/Latinx	4.18	-1.72	-0.65	1.35
	Asian	2.86	-1.94	-0.75	1.47
	Multi-racial/Other	5.18	-1.65	-0.50	1.48
Item 4					
	Unadjusted*				
	Base Group	1.79	-1.61	0.32	
	White	1.80	-1.67	0.36	
	Black/African American	1.92	-1.73	0.07	
	Hispanic/Latinx	1.62	-1.38	0.52	
	Asian	2.56	-1.06	0.33	
	Multi-racial/Other	1.98	-1.62	0.36	
	Adjusted*				
	Base Group	2.57	-1.60	0.02	1.97
	White	2.40	-1.79	0.07	2.17
	Black/African American	2.57	-1.54	-0.18	1.85
	Hispanic/Latinx	2.03	-1.58	0.15	2.08
	Asian	2.66	-1.39	0.05	2.25
	Multi-racial/Other	2.50	-1.71	0.09	2.08

* $p < .01$.

Ethnicity Pairwise Comparisons Group-Specific Item Parameters (Organized by Item)

Item 2

Item 2	a,	b_1	b_2	b_3
Unadjusted*				
Black/African American	3.68	-2.32	-1.96	-0.51
Hispanic/Latinx	3.91	-2.82	-2.13	-0.54
Adjusted*				
Black/African American	4.56	-1.48	-0.70	1.39
Hispanic/Latinx	4.45	-1.68	-0.69	1.18

* $p < .01$.

Item 3

Item 3	a,	b_1	b_2	b_3
Unadjusted*				
White	2.49	-1.81	-0.26	
Black/African American	3.09	-2.01	-0.44	
Adjusted*				
White	2.97	-1.87	-0.39	1.84
Black/African American	3.99	-1.58	-0.56	1.56

* $p < .01$.

Item 3	a,	b_1	b_2	b_3	b_4
Unadjusted*					
Black/African American	3.06	-2.26	-0.68		
Asian	2.70	-2.03	-0.38		
Adjusted					
Black/African American	4.46	-3.09	-1.68	-0.78	1.24
Asian	2.77	-2.95	-1.80	-0.49	1.70

* $p < .01$.

Item 4

Item 4	a,	b_1	b_2	b_3	b_4
Unadjusted*					
Base Group	1.78	-1.70	0.24		
Black/African American	1.87	-1.86	-0.02		
Adjusted					
Base Group	2.63	-2.69	-1.57	0.36	1.97
Black/African American	2.69	-2.71	-1.72	-0.44	1.57

* $p < .01$.

Item 4	a,	b_1	b_2	b_3	b_4
Unadjusted*					
Base Group	1.79	-3.15	-1.57	0.35	
Hispanic/Latinx	1.61	-2.90	-1.35	0.55	
Adjusted*					
Base Group	2.74	-2.61	-1.52	0.02	1.86
Hispanic/Latinx	2.18	-2.64	-1.47	0.15	1.99

* $p < .01$.

Item 4	a,	b_1	b_2	b_3	b_4
Unadjusted*					
Base Group	1.80	-1.61	0.30		
Asian	2.57	-1.06	0.31		
Adjusted					
Base Group	2.63	-2.70	-1.57	0.04	1.97
Asian	2.77	-3.44	-1.45	-0.06	2.16

* $p < .01$.

Item 4	a,	b_1	b_2	b_3	b_4
Unadjusted*					
White	1.79	-1.68	0.35		
Black/African American	1.89	-1.80	0.02		
Adjusted*					
White	2.31	-3.18	-1.85	0.09	2.27
Black/African American	2.48	-2.73	-1.60	-0.20	1.90

* $p < .01$.

Item 4	a,	b_1	b_2	b_3	b_4
Unadjusted*					
White	1.83	-3.14	-1.59	0.41	
Hispanic/Latinx	1.62	-2.87	-1.32	0.57	
Adjusted*					
White	2.42	-3.06	-1.74	0.12	2.20
Hispanic/Latinx	2.03	-2.75	-1.52	0.22	2.16

* $p < .01$.

Item 4	a,	b_1	b_2	b_3	b_4
Unadjusted*					
White	1.82	-1.62	0.39		
Asian	2.59	-1.01	0.37		
Adjusted					
White	2.21	-3.24	-1.84	0.18	2.42
Asian	2.77	-3.44	-1.44	-0.06	2.17

* $p < .01$.

Item 4	a,	b_1	b_2	b_3	b_4
Unadjusted*					
Black/African American	1.89	-1.81	0.03		
Hispanic/Latinx	1.60	-1.45	0.47		
Adjusted*					
Black/African American	2.77	-2.52	-1.51	-0.25	1.66
Hispanic/Latinx	2.26	-2.63	-1.51	0.06	1.85

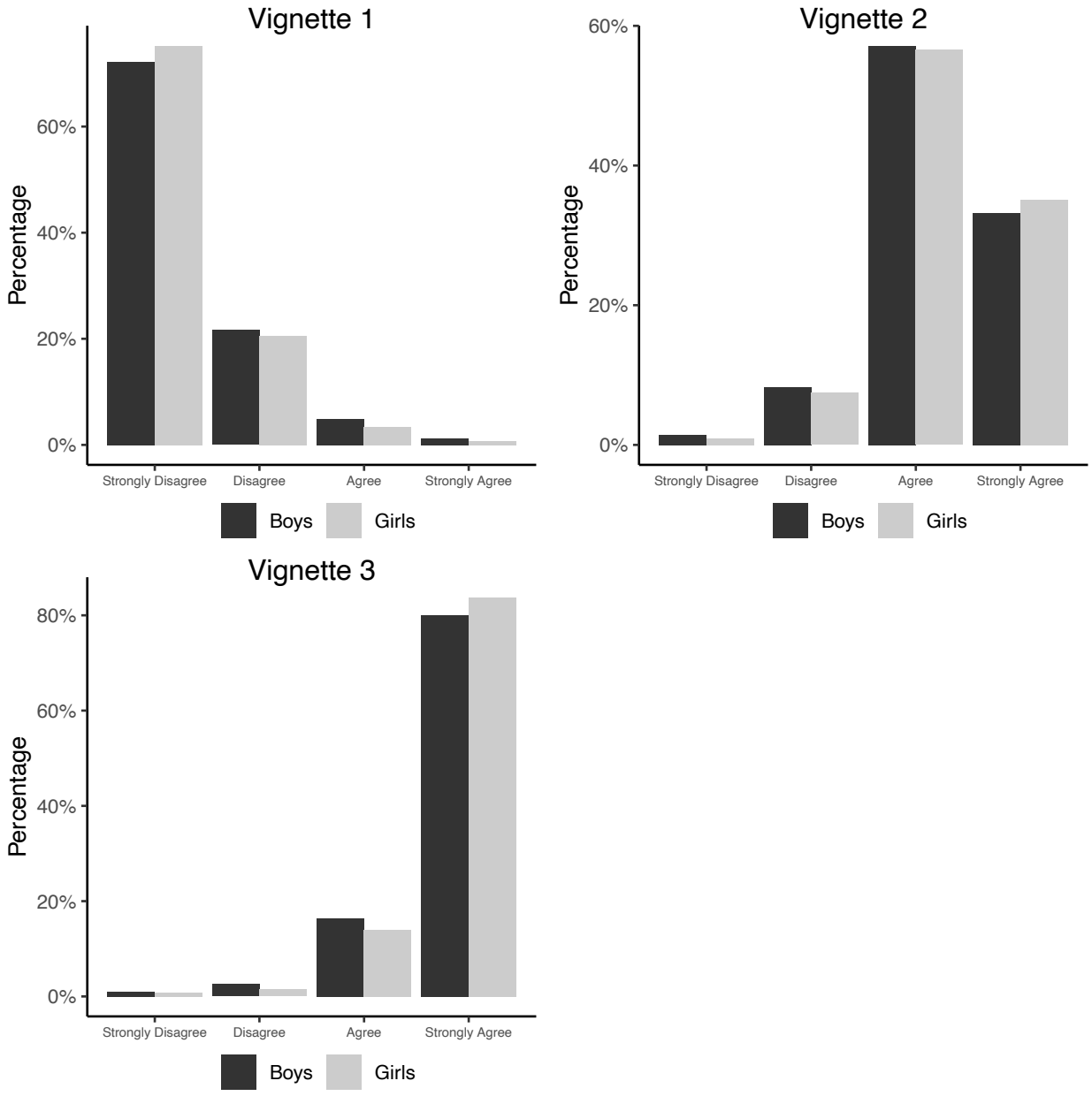
* $p < .01$.

Item 4	a,	b_1	b_2	b_3
Unadjusted*				
Black/African American	1.85	-2.07	-0.19	
Asian	2.49	-1.31	0.14	
Adjusted*				
Black/African American	2.68	-1.70	-0.40	1.60
Asian	2.81	-1.52	-0.17	2.00

* $p < .01$.

APPENDIX E

Response Distribution for Anchoring Vignettes, by Gender



APPENDIX F

Vignette Ordering Pattern, by Gender