GLOBAL PERSPECTIVES ON EDUCATIONAL TESTING: EXAMINING FAIRNESS, HIGH-STAKES AND POLICY REFORM

ADVANCES IN EDUCATION IN DIVERSE COMMUNITIES: RESEARCH, POLICY AND PRAXIS

Series Editor: Carol Camp Yeakey
Recent Volumes:

Volume 3: Teacher Unions and Education Policy: Retrenchment or Reform? – Edited by Ronald D. Henderson, Wayne J. Urban and Paul Wolman

Volume 4: Suffer the Little Children: National and International Dimensions of Child Poverty and Public Policy – Edited by Carol Camp Yeakey, Jeanita W. Richardson and Judith Brooks Buck

Volume 5: Higher Education in a Global Society: Achieving Diversity, Equity and Excellence – Edited by Walter R. Allen, Marguerite Bonous-Hammarth and Robert Teranishi

Volume 6: Power, Voice and the Public Good: Schooling and Education in Global Societies – Edited by Rodney K. Hopson, Carol Camp Yeakey and Francis Musa Boakari

Volume 7: As the World Turns: Implications of Global Shifts in Higher Education for Theory, Research and Practice – Edited by Walter R. Allen, Robert T. Teranishi and Marguerite Bonous-Hammarth

Volume 8: Living on the Boundaries: Urban Marginality in National and International Contexts – Edited by Carol Camp Yeakey

Volume 9: Health Disparities Among Under-Served Populations: Implications for Research, Policy and Praxis – Edited by Sheri R. Notaro

Volume 10: The Obama Administration and Educational Reform – Edited by Eboni M. Zamani-Gallaher

Volume 11: Mitigating Inequality: Higher Education Research, Policy, and Practice in an Era of Massification and Stratification – Edited by Robert T. Teranishi, Loni Bordoloi Pazich, Marcelo Knobel and Walter R. Allen

Volume 12: The Power of Resistance: Culture, Ideology and Social Reproduction in Global Contexts – Edited by Rowhea M. Elmesky, Carol Camp Yeakey and Olivia Marcucci

ADVANCES IN EDUCATION IN DIVERSE COMMUNITIES: RESEARCH, POLICY AND PRAXIS VOLUME 13

GLOBAL PERSPECTIVES ON EDUCATIONAL TESTING: EXAMINING FAIRNESS, HIGH-STAKES AND POLICY REFORM

BY

KEENA ARBUTHNOT
Louisiana State University, USA

United Kingdom – North America – Japan
India – Malaysia – China

Emerald Publishing Limited
Howard House, Wagon Lane, Bingley BD16 1WA, UK

First edition 2017

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-1-78635-434-1 (Print)
ISBN: 978-1-78635-433-4 (Online)
ISBN: 978-1-78714-896-3 (Epub)

ISSN: 1479-358X (Series)

ISOQAR certified Management System, awarded to Emerald for adherence to Environmental standard ISO 14001:2004.

Certificate Number 1985
ISO 14001

INVESTOR IN PEOPLE

This book is dedicated to William, Alfreda & Stephanie

This page intentionally left blank

CONTENTS

LIST OF TABLES

This page intentionally left blank

ACKNOWLEDGMENTS

I have many mentors and colleagues whom I must thank for their selflessness in helping me to achieve my academic and career goals, including, but not limited to, Drs. Edmund W. Gordon, M. Christopher Brown II, William Trent, Sara Lawrence-Lightfoot, Eugene Kennedy, Stafford Hood, Carol Camp Yeakey, Katherine E. Ryan, and Bridget Terry Long. Each of you have been instrumental in shaping my career, and I am truly grateful for your leadership and vision to light the way for those of us who come after you. Additionally, I must acknowledge those whom I consider my contemporaries and who work alongside me to leave our mark in our respective fields. We support and inspire one another and provide much needed shoulders on which to lean in times of trial, and are true encouragers in times of triumph. These individuals include, but are not limited to, Dawn Williams Salter, Maurice Samuels, Darrell Ray, and Sassy Wheeler. Last, I must acknowledge those scholars whom I have had the distinct honor of mentoring and advising during their academic pursuits, including but not limited to, Tireka Cobb, Jared Avery, Adam Elder, Erin Scott-Stewart, and Guadalupe Lamadrid.

Personally, I want to thank my parents William and Alfreda Arbuthnot and my sister Stephanie Arbuthnot whose love and support have always been the wind beneath my wings. I cannot begin to tell each of you how much I appreciate the unconditional love and unwavering support that you have shown me throughout the years. You continue to encourage me to meet my highest potential and to always put God first. As a child growing up, each one of you taught me in your own special way how to face adversities with dignity and fortitude. Those lessons taught me how to survive in the face of opposition and how to thrive in the midst of it. In addition to my immediate family, I have been blessed with an incredibly supportive extended family, including but not limited to, Barbara Whitener Gardner, Kelly Gardner, Kimberly Gardner, Michael Bates, the Murry family and other aunts, uncles, and cousins. I also want to thank my friends, many of whom have become family, for always making the choice to love and support me regardless of the circumstances. There are too many to name but they include Jas and Samaah Sullivan, Vashti Person, Franklin Mosley III, Kaisha Mozee, Valerie Byers,

Tyrone Perkins, LaShonda Harvey, Eric Cook, Bridget Perkins, Kimberly Cole, and the Seven Sisters. Last, to all my godchildren, mentees and former students, including but not limited to, Tierra Webb, Aja and Mariah Brooks, Dallas Hawkins, Jennifer Cook, Aisha Camphor and Marcus Jones, you are all an inspiration to me.

I want to also thank all of the friends and colleagues I have made across the world. The opportunity to meet wonderful people from various places around the globe has reshaped my research focus not only on issues plaguing testing and measurement in the United States, but also on expanding this line of research to issues pertaining to education worldwide. Improving education and understanding the way in which all students learn, regardless of race, ethnicity, or country of origin, have become central to my research agenda. My experiences abroad have shown me that all parents, educators, and policy-makers intuitively want the same thing: *to make education accessible, fair and engaging to all children* regardless of who they are or from where they come.

INTRODUCTION

Much of my career has focused on investigating group differences in standardized test performance and test fairness issues in the United States. My interest in this area was sparked by my experiences as a high school mathematics teacher. As a teacher, I realized the impact that high-stakes tests had on students, particularly Black students, who are considered to be the minority in the United States. Research has shown that on most standardized tests, White students outperformed their Black counterparts. I wanted to determine whether these high-stakes tests were fair to all test taker groups, and, consequently, spent years as a graduate student and professor examining fairness issues and investigating test- and item-level performance patterns. My research investigated Black and White test takers and highlighted the differences in the test-taking experiences between the two groups (Arbuthnot, 2009, 2011a, Arbuthnot & Ryan, 2005). The results undeniably showed that cultural differences between Blacks and Whites had a significant impact on their standardized test experience and performance (Arbuthnot, 2011a,-, 2015a-, 2015b; Arbuthnot & Lyons-Thomas, 2016). Moreover, the research identified different mathematics subtests and types of items that were differentially harder for one group in comparison to the other. Additionally, the research focused on how students from various groups process test items differently while taking a standardized test. The research provided a new way to understand and conceptualize the way that different race/ethnic groups, with contrasting cultural backgrounds, experienced and performed on tests. My book *Filling in the Blanks: Understanding the Black White Achievement Gap* (Arbuthnot, 2011) provided comprehensive details from over a decade of research regarding the variations between Black and White test-takers' experiences. One line of the book's research examined the similarities in the test-taking experience and performance of Black students and female students on standardized mathematics tests. The results indicated that Black and female students performed similarly on these examinations. To explain this finding, I examined the commonalities between Black students and female students in the United States. Consequently, the cultural similarities between the two groups helped to explain why the test-taking patterns of Black and female students were comparable.

Simultaneously, I dedicated my research to addressing issues related to test fairness as well. My writing focused on examining the high-stakes testing systems and providing empirical research to challenge test fairness issues. This research highlighted the roles and responsibilities of test developers in ensuring that tests were fair to all test takers, as well as challenging test users and consumers to critically examine the way in which they interpreted test results (Arbuthnot, 2011a, 2012a, 2015a, 2015b; Arbuthnot & Lyons-Thomas, 2016).

I continued my research and worked on issues related to domestic test fairness; concurrently, I was presented with opportunities to participate in consulting and grant opportunities abroad. Most of my opportunities abroad involved countries in the Middle East, mainly Qatar and the United Arab Emirates. From these experiences, I developed a better understanding of the educational challenges that Middle Eastern countries faced. I realized that some of the issues that students faced in that region were similar to the difficulties that I recognized in my research with Black and female students in the United States. With my background in standardized testing, I turned to international assessments to investigate the similarities between minority and female students in the United States and students from the Middle East, expanding my line of research on testing and fairness to a global scale.

This book highlights the basis and justification for the research and provides a detailed exploration of how to examine fairness issues on international assessments. It is my hope that audiences around the world will utilize this book in the quest for understanding and conceptualizing variations in the way in which test takers from different countries and cultures learn and perform on tests.

The purpose of this book is to investigate fairness issues on international assessments. The text begins with an overview of the current state of international assessments and reveals the ways in which many countries have utilized results from international assessment initiatives to inform educational policy and practice at the national level. The book then describes the various international assessment programs that have been implemented over the last several decades. The text then focuses on the Trends in International Mathematics and Science Study (TIMSS) assessment, one of the most popular and longstanding international assessment programs. The book utilizes TIMSS assessment data that includes the 4th and 8th grade mathematics test in conjunction with information obtained from a variety of stakeholder questionnaires (i.e., students, teachers, and administrators) from each of the participating countries. Additionally, the use of details concerning fairness issues is included as well as how research conducted on fairness issues in the

United States provides a framework for examining issues of fairness at the global level.

In order to investigate fairness issues on the global scale, the author introduces The Arbuthnot Assessment Fairness (TAAF) Framework as a means to systematically examine test- and item-level performance patterns and fairness issues. The TAAF Framework has four phases: (a) country selection, (b) test and subtest performance and test-taking patterns, (c) item-level differences and patterns, and (d) cultural impact. Completing these four phases of the TAAF Framework provides a clearer understanding and interpretation of test- and item-level differences and provides an in-depth examination of fairness issues on international assessment. The author provides a detailed example of how to use the TAAF Framework to inform multiple stakeholder groups about differences in performance patterns among countries and test fairness issues. The countries of Chinese Taipei, Finland, the United States, and Qatar were chosen for the research based on three criteria: cultural context, performance patterns, and educational reform policies. The book provides a critical examination of the educational practices, assessments, accountability measures, and cultural norms for each of these countries.

The book utilizes the empirical data provided by the mathematical portion of the TIMSS international assessments to demonstrate and elucidate how to analyze the international assessment data and use multiple data sources to examine issues of fairness on international assessments. This book ends by challenging readers to deliberate more thoughtfully and to exercise caution with respect to the ways in which test- and item-level performance data is interpreted on international assessments.

To conclude, the author encourages readers to be mindful when interpreting test performance and insists that the various stakeholder groups, the test developers, educational theorists, policymakers, and practitioners, better understand the differences in performance patterns between countries and the issues surrounding test fairness on international assessments.

This page intentionally left blank

CHAPTER 1

INTERNATIONAL ASSESSMENTS AND FAIRNESS ISSUES

Many countries strive to become world leaders in education, especially in math, technology and science. With the ever-changing world and the need to progress and innovate rapidly, countries are pressured to ensure that their citizens are among the most talented and innovative in the world. Countries strive to have exceptional educational systems that produce the next generations' world leaders and innovators, who, in turn, will support and strengthen the economic competitiveness of their respective countries.

To gauge the differences in the educational strengths of countries worldwide and to improve teaching and learning on a global scale, several international assessment programs have been implemented. These international assessment programs administer tests and assessments that are intended to provide a comprehensive set of data that countries can use to improve teaching and learning and to gauge their educational ranking among other countries. In the last several decades, countries have competed to perform well and increase their rankings on international assessments. Research has also shown that many countries have utilized the results from these international assessments as a catalyst to reform their country's educational system (Bernhaum & Moore, 2012; Carnoy, 2015). Since then the stakes involved with participating in and excelling on these international assessments have increased. The results of the testing programs provide an assessment of their

Global Perspectives on Educational Testing: Examining Fairness, High-Stakes and Policy Reform
Advances in Education in Diverse Communities: Research, Policy and Praxis, Volume 13, 1–21

ISSN: 1479-358X/doi:10.1108/S1479-358X20160000013007

educational system as well as international bragging rights. There was an immediate response to the results from the 2015 The Trends in International Mathematics and Science Study (TIMSS) and Program for International Student Assessment (PISA) were released. The findings made the headlines of newspapers and publications worldwide, and the publications' responses were mixed based on the performance of the individual countries, as seen in the following news and headline excerpts (Hatch, 2016):

> "France students rank last in EU for maths, study finds" (*France 24 International News*).
>
> "Ireland ranks 15th in global league table for maths, science" (*Irish Times*).
>
> "Moroccan math and science education struggling, but improving: Survey" (*Moroccan World News*).
>
> "Singapore students top global achievement test in mathematics and science" (*Straits Times*).
>
> "UAE pupils improve math and science skills, global study shows" (*The National*).
>
> "US students score higher than average on international math test: Students in some Asian nations excel; US students improve" (*Wall Street Journal*).
>
> "PISA 2015 brings more bad news for Australia" (*Teacher Magazine*).
>
> "PISA 2015: Estonia's basic education best in Europe" (*The Baltic Course*).
>
> "PISA: Finland only country where girls top boys in science" (*YLE News*).
>
> "Education: Macau students' 'score high' on PISA 2015" (*Macau Daily Times*).

As the headlines indicated, the results of these international assessments have a profound impact on education systems worldwide. Some countries use the assessments as a means to evaluate or monitor their country's educational system and progress over time, while others use them as a means to compare themselves with other countries worldwide (Hatch, 2016). Because of the far-reaching impact on the interpretation and evaluation of worldwide educational systems, researchers must examine more closely the tests and assessments used in these programs. Specifically, issues related to test fairness need to be investigated to ensure that the international tests and assessments that are being used for high-stakes purposes are fair to the entire population of test takers (Arbuthnot, 2011a, 2012a, 2015a, 2015b; Arbuthnot & Lyons-Thomas, 2016). Although this may be a seemingly complex and daunting task, it is essential to address issues of fairness based on the intended uses of the results from these assessments.

Several international testing and assessment programs have been designed to assess the academic performance of groups of test takers from various

countries. All of the testing initiatives have different purposes and foci. Details about four major international assessments are presented below.

Table 1 shows how the four international testing programs differ on several factors. In general, the PISA and TIMSS international assessments have been the most popular worldwide and have a significant impact on educational systems globally. Investigating international testing initiatives is one way of gaining a more comprehensive picture of teaching and learning on a global scale. For the purposes of this book, the mathematics portion of the TIMSS assessment will be examined. Specifically, mathematics was chosen to be the focus of this research because Science, Technology, Engineering, and Mathematics (STEM) fields and education are of global concern. STEM education and innovation are increasingly important on a global scale and has been shown to have a direct impact on countries' productivity and economic development, as are performing well and being competitive in STEM fields. Details about the TIMSS assessment follow.

TIMSS ASSESSMENT

TIMSS is an international assessment in which more than 50 countries participate. Active since 1995, the purpose of the assessment program is to gather and disseminate information about students that will help improve teaching and learning in mathematics and science (Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009). The TIMSS assessment is guided by the following four major questions: What should students learn? Who provides the instruction and how is it organized? Where and when does instruction take place? and What have students learned? The TIMSS assessment "provides a comprehensive and internationally comparable data about what mathematics and science concepts, processes, and attitudes students have learned by the 4th and 8th grades" (Mullis, Martin, Ruddock, et al., 2009). Policy-wise, one of the main goals of TIMSS is to help countries make evidence-based decisions that impact educational policy including: (a) measuring the effectiveness of their educational system in a global context; (b) identifying gaps in learning resources and opportunities; (c) pinpointing any areas of weakness and stimulating curriculum reform; (d) measuring the impact of new educational initiatives; and (e) training researchers and teachers in assessment and evaluations.

As previously mentioned, TIMSS is designed to measure the achievement of 4th, 8th, and 12th grade students in the areas of mathematics and science (Mullis, Martin, Ruddock, et al., 2009). TIMSS is administered in the

Table 1. Comparison of Major International Assessments Programs.

	PISA	PIRLS	TIMSS	PASEC
Name	Program for International Student Assessment	Progress in International Reading Literacy Study	Trends in International Mathematics and Science Study	Analysis Program in Educational Systems CONFEMEN
History	Launched in 1997 with first survey taken in 2000	First survey taken in 2001	First survey taken in 1995 with five different grades, 1999 with 8th grade only, then 2003 with 4th and 8th grades	Launched in 2013 with first administration in 2014
Cycle	Every 3years	Every 5years	Every 4years	Every 4years
Sampling Strategy	Schools are randomly selected and all students in the specified age range take the assessment	Schools are randomly selected then classrooms within selected schools are randomly selected; all students in selected classes take the assessment	Schools are randomly selected then classrooms within selected schools are randomly selected; all students in selected classes take the assessment	Schools are randomly selected using stratified sampling then one classroom within selected schools are randomly selected; all students in the selected class take the assessment
Participants	15-Year-old students (between 15 years, 3 months and 16 years, 2 months)	4th grade students	4th and 8th grade students	2nd and 6th grade students
Countries	72	45	Between 40 and 50	Over 20 countries in Africa, Asia, and the Middle East
Content Domains	Science, Mathematics, Reading, Collaborative Problem Solving, and Financial Literacy	Reading	Mathematics and Science	Reading and Mathematics
Stakeholder Questionnaires	Students and school principals take mandatory questionnaires; optional questionnaire for parents	Student, parents/guardians, teachers, principals, and national research coordinator	Student, parents/guardians, teachers, principals, and national research coordinator	Students, principals and teachers, and national research coordinator

Source: TIMSS data. Arbuthnot (2017).

participating countries once every 4 years for the 4th and 8th grade students. The TIMSS sampling strategy consists of randomly sampling schools from each country, and random samples of classrooms within a school. Specifically, no less than 150 schools from each country are included, with approximately 4,000 students per grade.

The mathematics and science tests consist of the following content domains:

4th Grade Assessment

- Math: numbers, geometric shapes and measures, and data display
- Science: life science, physical science, and Earth science

8th Grade Assessment

- Math: numbers, algebra, geometry, and data and chance
- Science: biology, chemistry, physics, and Earth science

12th Grade Assessment

- Math: algebra, calculus, and geometry
- Science: advanced physics (mechanics and thermodynamics, electricity and magnetism, and wave phenomena and atomic/nuclear physics)

Each of these tests includes both multiple choice and constructed response test items. The TIMSS examination is a timed test that students have 90 minutes to complete. To calculate the total score, multiple choice items are worth 1 point and constructed response items are worth 1 or 2 points and allow test takers to receive partial credit. Upon completion of the test, students complete a questionnaire that asks a battery of questions about their background, home life, family, teachers, and school. In addition, multiple stakeholders, including students, parents, teachers, and national curriculum coordinators complete questionnaires. The TIMSS assessment program has allowed the public to access comprehensive datasets that include information about TIMSS participants' test- and item-level performance, disaggregated by country, and the questionnaire responses from the multiple stakeholder groups. All of this information will be utilized in the remaining chapters in this text.

When making reference to the TIMSS assessment in this book, assessment includes several sources of data such as the mathematics test and questionnaire responses from multiple stakeholders; consequently, the TIMSS mathematics test is but one component of the TIMSS assessment. Specifically, the Standards for Educational and Psychological Testing (1999) defined

assessment as "any systematic method of obtaining information from tests and other sources, used to draw inferences about characteristics of people, objects, or programs" (p. 172). Additionally, a test is defined as an evaluative tool or device that is scored and analyzed using a standardized process (AERA, APA, & NCME, 1999). Examining these multiple data sources can be instrumental in understanding the nuances and reasons for such vast differences in educational achievement globally, which in turn, could have a profound impact on educational theory, policy, and practice.

INTERNATIONAL ASSESSMENTS AND POLICY

The results from several of these international testing initiatives have stimulated global dialogue around issues of educational reform and improvement. In response, many countries have developed educational reform programs to improve test performance amongst their students (Bernhaum & Moore, 2012; Supovitz, 2009). Breakspear (2012) examined the impact the international assessment PISA had on the educational reform policies and practices across various countries. For this study, a survey was distributed to representatives from the 65 participating countries. Slightly more than 50% ($n = 37$) of the representatives took the survey. Over 50% of the respondents indicated that the overall international ranking had an impact on their educational policies. Additionally, the results showed that several countries had utilized the PISA results and data as a mechanism to change their educational reform movements. For example, in the early 2000s, Germany initiated an educational reform effort focused on improving student learning and test performance. This educational reform focused on creating national standards and providing resources and support for disadvantaged students (Ertl, 2006). Yet another example of a response to the PISA results was Denmark, which afforded more financial resources to their educational reform system. This new system emphasized more assessments and evaluation in addition to providing resources for underserved students (Egelund, 2008). The term "PISA shock" was developed to describe the immediate response of countries to the results of PISA. Breakspear (2012) stated, "PISA plays an important function for policy makers as they seek to evaluate and improve system performance in response to the demands of the global knowledge economy" (p. 28). Over one-fourth of the countries involved in this research reported that PISA results would be utilized as a means to monitor and evaluate the effectiveness of their educational reform policies. Breakspear (2012) reported that "policy makers across nearly all PISA-participating countries/economies see PISA as

an important indicator of system performance, and there is evidence that the PISA evaluation has the potential to 'define' the policy problems and set the agenda for policy debate at the national and state levels" (p. 27).

In addition to affecting educational reform policies, Breakspear's (2012) study also revealed how some countries specifically utilized the PISA results. The findings showed that over 70% of respondents reported that the PISA results were "extremely" or "very" important as an indicator of a school's effectiveness. In addition, nearly one-third of the respondents indicated that the PISA results directly influenced their national assessment strategy. A respondent from the Slovak Republic wrote, "Under the influence of PISA, we implemented new national measurements of reading and mathematics as a direct consequence of poor results of our country in PISA 2003 and 2006 cycles" (Breakspear, 2012, p. 19). Similarly, a respondent from Japan wrote, "It had been decided to introduce a national assessment of student performance after the release of the PISA 2003 results in 2004, and actually implemented since 2007. PISA-type assessment items are being used in the national assessments" (Breakspear, 2012, p. 19). Lastly, over 40% of the respondents reported that national standards in their countries were modified or changed in response to the PISA results.

This research illustrates and establishes the far-reaching and critical worldwide impact of international assessments on various aspects of educational systems. As described, respondents from participating countries reported that the PISA reports had been used in various ways, including (a) monitoring school effectiveness, (b) changing national assessment standards, and (c) monitoring educational reform systems. Since the results from international assessments have a significant impact on educational policies as well as practice, and, thus, these assessments can be regarded as high-stakes. The following sections highlight the impact that the results of the international assessments have had on certain countries as exemplars in education and how other countries emulate them to improve their own educational systems.

BENCHMARKING TRENDS

To improve test scores, many countries have looked to other countries' educational system and reform efforts to establish benchmarks. Benchmarking occurs when a country tries to replicate or emulate another countries' educational policy or practices in an effort to improve educational outcomes. Consequently, many countries that are experiencing issues or failure on

international assessments are, in a sense, copying the educational policies from other countries. Carnoy (2015) wrote,

> Rankings of countries educational systems based on international test scores and policy lessons drawn from high scoring countries educational systems have taken on a life of their own, turning some national educational systems into superstars to be admired and flooded with educational tourists, and other national systems, such as the United States, into sad sacks, criticized, and mocked for being stagnant and failing. (p. 15)

This benchmarking process has become common amongst many countries; however, there are many concerns about this practice, concerns that are justified based on the notion that critical differences exist in context and culture across countries, making the adoption or "standardization" of educational reform across borders problematic. Consequently, copying or emulating what has been successful in one country may not work or be as effective in another because of contextual and cultural issues. Countries including Finland, Korea, Shanghai, Chinese Taipei, Singapore, Canada, Australia, and New Zealand have been benchmarked and/or emulated because of their relative success on international assessments. For example, Breakspear (2012) reported that a Chilean respondent wrote the following about their educational reform efforts:

> The experiences from several countries have been considered when developing different educational policies. For example, Finland is quoted in regard to equity, high performance, teacher training, and absence of a high-stakes assessment system. The United Kingdom was reviewed for curriculum, institutional organization, and, recently, for assessment consequences. And some US states … like Massachusetts, for assessment and accountability. (p. 17)

Although this quote captures the overarching theme and perspective of many countries that benchmarking efforts are advantageous, there is still growing concern about limitations of benchmarking certain countries due to culture and contextual variations and differences.

CONCERNS AND SHORTCOMING OF UTILIZING INTERNATIONAL ASSESSMENTS

The impact that international assessments have had on education globally have caused an outpouring of concerns by many groups. A growing body of literature has highlighted the negative, unintended consequences of these assessment programs and their relative uses (Carnoy, 2015; Plank & Condliffe, 2013).

First, many argue that international assessments promote countries to teach to the test. The notion of teaching to the test is commonly argued in the United States as a result of the test-based accountability measures that were implemented in the early 2000s (Carnoy, 2015). Teaching to the test has been shown to narrow the curriculum and can limit certain positive aspects of student learning, like higher-order thinking skills. Other criticisms emphasize how the roles of the teacher, student, and parent can be impacted when intense focus is placed on international test or assessment results (Paris & Urdan, 2000). Finally, many have critiqued issues directly related to the calculation and reporting of test scores and performance. Some have argued that the scores should take into account socioeconomic differences when they are being analyzed and reported. Additionally, it has been argued that the results should be disaggregated by race/ethnicity to provide a much clearer picture of test performance patterns (Carnoy, 2015). Others contend that various levels of data from these international assessments should be available to the public in a timely fashion, allowing external sources to confirm the results from test administrations (Carnoy, 2015). These concerns shed light on the issues that need to be addressed on international assessments. Given the high-stakes involved, it is imperative that attention is paid to the unintended consequences of these types of international testing programs. Specifically, exploring the issues relative to test fairness is necessary to ensure that these international tests are fair to their diverse population of test takers (Arbuthnot, 2011a, 2012a, 2015a, 2015b; Arbuthnot & Lyons-Thomas, 2016).

TEST FAIRNESS

International assessments can be considered high stakes in nature, because they have been shown to have a significant impact on education worldwide. Since international assessments are high stakes, investigating the test-taking experiences of test takers from different countries and cultural groups ensures that the tests are fair to all groups. Consequently, given the global high-stakes testing environment, it is common to hear researchers and the public debate issues of test fairness (Arbuthnot, 2011a, 2012a, 2015a, 2015b; Arbuthnot & Lyons-Thomas, 2016). Are tests fair to all groups of test takers? Is there a way to ensure the fairness of tests? How do we assess test fairness? Although there is intense interest in test fairness issues, there is no single technical definition of test fairness; however, there is a categorization of the varying views of fairness and potential threats to fairness. These differing views of fairness and threats to fairness are outlined in the Standards for Educational and

Psychological Testing (2014), referenced hereafter as the Standards (2014). Details follow about the Standards and how fairness issues are conceptualized and understood.

STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING

In the 1960s, the Standards for Educational and Psychological Testing were created in response to ongoing critiques of fairness and equity issues related to testing and measurement and are recognized internationally as the "gold" standard for those developing and evaluating tests. Several organizations were involved in the development of the Standards for Educational and Psychological Testing, including the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. The Standards (2014) provided a comprehensive set of guidelines to follow when developing and evaluating tests. The purpose of the Standards (2014) was:

> To provide criteria for the development and evaluation of and testing practices and to provide guidelines for assessing the validity of interpretations of test scores for the intended test uses. Although such evaluations should depend heavily on professional judgment, the Standards provide a frame of reference to ensure that relevant issues are addressed. All professional test developers, sponsors, publishers, and users should make reasonable efforts to satisfy and follow the Standards and should encourage others to do so. (p. 1)

Several versions of the Standards for Educational and Psychological Testing have been produced since their inception. Each revision provided a more contemporary view of the guidelines for test development and evaluation, as well as addressing issues related to test fairness. An examination of the most recent version of the Standards (2014) is discussed in this book, and references to test fairness issues as they relate to the development and evaluation of tests are also highlighted.

The Standards (2014) highlights four general views of test fairness: fairness in treatment during the testing process, fairness as lack of measurement bias, fairness as validity of individual test score interpretations for the intended uses, and fairness as access to the construct as measured. In general, the Standards (2014) state:

> A test that is fair within the meaning of the Standards reflects the same constructs for all test takers, and scores from it have the same meaning for all individuals in the intended population; a fair test does not advantage or disadvantage some individuals because of characteristics of all individuals in the intended test population, including those associated with race, ethnicity, gender, age, socioeconomic status, or linguistic or cultural

> background, must be considered throughout all stages of development, administration, scoring, interpretation, and use so that barriers to fair assessment can be reduced. At the same time, test scores must yield valid interpretations for intended uses, and different approaches to fairness. (p. 50)

As stated above, the Standards (2014) highlight that fair tests do not advantage or disadvantage certain groups of test takers based on characteristics unrelated to the construct being measured. Details of each of the four different views of fairness follow as well as a discussion on how these varying views can be interpreted on international assessments.

Fairness in Treatment During the Testing Process

The first categorization of fairness refers to equitable treatment in the testing process. This view of fairness is concerned with ensuring that all test takers are being tested given the appropriate testing conditions and that test takers have an equal opportunity to become familiarized with the test format. Additionally, careful consideration regarding standardization of test procedures, administration, and scoring is highlighted. However, the Standards (2014) insist that sometimes situations arise where certain groups of test takers need flexibility for them to sufficiently engage with the construct that is being measured. The Standards (2014) state, "challenges may arise due to an examinee's disability, cultural background, linguistic background, race, ethnicity, socioeconomic status, limitations that may come with aging or some combination of these, or other factors." Consequently, this view of fairness emphasizes equitable treatment and potential standardization of test procedures, administration, and scoring; however, it also acknowledges the challenges that certain groups of test takers may face based on their personal characteristics (i.e., disability, culture, race/ethnicity) (p. 51). This interpretation of fairness is quite applicable on international assessments, given the challenges that test developers face when trying to standardize the testing process for such a widely diverse population of test takers.

Fairness as Lack of Measurement Bias

The second categorization of fairness refers to a test or assessment as lacking measurement bias. The Standards (2014) define bias in two ways: item bias and predictive bias. These definitions of bias are related to item-level differences and the relationship between test performance and other external

factors. The person(s) who asserts this definition of fairness has the belief that if a test does not display item or predictive bias, then it is a fair test.

Item Bias

As stated above both predictive bias and item bias are aspects of measurement bias. While predictive bias focuses on the relationship between the test and some established external criteria, item bias is concerned with item-level differences. This book will focus on the latter. To investigate issues related to item bias, researchers utilize different types of statistical analyses to identify problematic items (Arbuthnot, 2009, 2011a; Arbuthnot & Lyons-Thomas, 2016; Arbuthnot & Ryan, 2005; Dorans & Holland, 1993).

Differential Item Functioning. One way to examine fairness issues is to identify items that seem to be biased toward one group in comparison with another. Traditionally, experts conduct differential item function (DIF) analyses to examine item bias. Dorans and Holland (1993) defined DIF as:

> [DIF is] a psychometric difference in how an item functions for two groups. DIF refers to a difference in item performance between two comparable groups of examinees, that is, groups that are matched with respect to the construct being measured by the test. The comparison of matched or comparable groups is critical because it is important to distinguish between differences in item functioning from differences between groups. (p. 35)

Similarly, Millsap and Everson (1993) explained that an item without DIF is defined as one in which, given a person's ability and group membership, the probability of getting an item right is equal to the probability of getting an item right given a person's ability. This definition illustrates that a person's ability level or group membership should not affect the probability of getting an item correct. Consequently, if group membership does make a difference in a test taker getting an item correct, then this item could potentially be considered biased. Test developers have utilized DIF analyses to examine item-level bias issues on standardized tests. Historically, a multitude of DIF research has been conducted in the United States, focusing primarily on examining item bias comparing Black versus White test takers and examining differences between male and female groups as well. Following are results of the DIF research focused on these differences. This research can be instrumental in disentangling fairness issues in international contexts as well.

Several studies have examined DIF items comparing Black and White test takers on standardized math tests. This body of literature has identified the types of math items where Blacks outperform their White counterparts and the areas

where Whites outperform Blacks (Arbuthnot, 2009, 2011a; Arbuthnot & Ryan, 2005; Gallagher et al., 1999; Scheuneman & Grima, 1997). Blacks or African Americans are one of the largest minority groups in the United States, while Whites are considered the majority group. It has been well documented that Whites have outperformed Black test takers on many standardized tests and assessments. DIF studies have been conducted to examine whether the differences in performance patterns between Black and White test takers is a result of item bias (Arbuthnot, 2009, 2011a; Arbuthnot & Ryan, 2005). Arbuthnot (2011a) provided a comprehensive list of the different areas that research in mathematics has shown DIF between Black and White test takers. Specifically, several areas were identified where one group of test takers performed differentially better than the other group. With reference to content domains, Black test takers performed differentially better than White test takers in Algebra. Conversely, White students performed differentially better than Black students in the Geometry and Measurement content domain.

The research also shows specific topic areas and types of mathematics test items that differentially favor Black or White test takers. The results from this research shows that Black students performed differentially better than White students on items that involved (a) coordinate planes, (b) simple math sentences, and (c) money, while White students performed differentially better in the data interpretation topic area and items that involved (a) a graph or a table, (b) unconventional items (i.e., estimation), (c) multiple steps, (d) a visual solution, (e) figures, (f) word problems, and (g) a real world setting. The above information highlights the relative areas of strengths and weaknesses for Black and White test takers and can be instrumental in understanding better differences across other groups as well.

In addition to the studies that examine item bias between Black and White test takers, a dearth of studies have examined differences in mathematics test performance between males and females in the United States as well (Arbuthnot, 2011a; Albano & Rodriguez, 2013; Ferguson & Arbuthnot, 2005; Gallagher, 1998; Garner & Engelhard, 1999; Harris & Carlton, 1993; Ryan & Arbuthnot, 2002; Ryan & Chiu, 2001; Ryan & Fan, 1996; Willingham & Cole, 1997). In general, research in United States has shown that males have historically outperformed females on mathematics standardized tests. DIF studies comparing male and female mathematics' performance have examined these differences and highlighted potential bias at the item level. The findings from these studies showed that female test takers performed differentially better on items from the Algebra content domain, while males perform differentially better on items in the Geometry and Measurement content domain (Albano & Rodriguez, 2013; Ryan & Chiu, 2001; Ryan & Fan, 1996). Additionally, DIF

research has highlighted topic areas and item types that differentially favor males or females (Garner & Engelhard, 1999; Harris & Carlton, 1993). In general, research indicates that items with any or all of the following components favor male test takers: (a) have multiple solution pathways to obtain the correct solution, (b) spatial items, (c) unconventional items (i.e., estimation), and (d) data analysis items. On the other hand, the following components favor female students: (a) pure algebraic, (b) a formulaic solution to solve, (c) multistep items that require accuracy, and (d) real world problems (Gallagher, 1998).

As indicated, several research studies have identified items on standardized tests that display DIF and are advantageous to one group over another; these studies have provided information about content domains, topic areas, and types of items that seem to favor one group over another. Although identifying or flagging items that show DIF is fairly straightforward, research has shown that it is often times difficult to pinpoint the sources or explanations for these marked differences at the item level (Arbuthnot, 2009, 2011a; Arbuthnot & Lyons-Thomas, 2016; Arbuthnot & Ryan, 2005; Sireci, 2004). Consequently, many of the explanations and sources of DIF are hypotheses based on prior research and the expertise of researchers. Although it is difficult to precisely identify the source or explanation for DIF items, this information is important and useful to challenge issues related to test fairness and to improve the overall test development process (Arbuthnot, 2011a, 2012a, 2015a, 2015b; Arbuthnot & Lyons-Thomas, 2016).

The previous data focused on items that have been shown to exhibit DIF. However, since conducting DIF analyses is not plausible in certain situations, the Standards (2014) also discuss the utility of identifying item bias based solely on group differences in performance. In examining bias issues, the Standards (2014) consistently stated that group differences in performance patterns alone do not necessarily mean that an item or test is biased. However, the Standards (2014) also stated that:

> Examination of group differences also may be important in generating new hypotheses about bias, fair treatment, and the accessibility of the construct as measured. … In many cases, it is not clear whether the differences are due to real differences between groups in the construct being measured or to some source of bias (i.e., construct-irrelevant variance or construct underrepresentation). In most cases, it may be a combination of real differences and bias (p. 54).

The recognition of differences between groups of test takers is not always attributed to bias; however, investigating these group differences more closely can produce new hypotheses about types of items or groups of items that might be biased. Consequently, it can be inferred that DIF analysis alone is not the only method to identify item bias; examining differences between

groups of test takers can be instrumental in hypothesizing about bias and fairness issues. Specifically, examining those test items that have abnormally large differences between groups can help to identify items that may be potentially biased. This is notable because access to the level of data needed to conduct DIF analyses is often limited on large-scale international assessments. Consequently, without item-level data, DIF analyses cannot be performed; however, careful examination of group differences can potentially provide vital information about bias and fairness at the test and item levels.

Fairness in Access to the Construct as Measured

The third categorization of fairness focuses on fairness in access to the construct as measured. This general view of fairness refers to the availability of the construct to the entire test taker population. Accessibility is described as testing that enables all test takers to show their status on the construct regardless of their individual characteristics (i.e., disability, age, race/ethnicity, cultural background, gender, language). This view of fairness emphasizes that one's individual characteristics should not be an advantage or disadvantage to a test taker in showing their status on the construct being measured.

The Standards (2014) stated:

> For some test takers … factors related to individual characteristics, such as age, race, ethnicity, socioeconomic status, cultural background, disability, and/or [English] language proficiency may restrict accessibility, and thus interfere with the measurement of the construct of interest. (p. 52)

This view of fairness encompasses several factors that may limit test takers when taking the test or assessment. The Standards (2014) suggested that test developers and consumers be aware of characteristics of test takers that may impede their access to the construct being measured. The Standards (2014) provided several examples of how factors such as visual impairments, language learners, and culture can potentially limit test takers access to the construct being measured.

Fairness as Validity of Individual Test Score Interpretations for the Intended Uses

The fourth categorization is fairness as validity of individual test score interpretations for the intended uses. This perception of fairness highlights the interpretation of fairness as an extension of validity. Specifically, one

may consider a test fair if there is evidence that proves the validity of the interpretation of the test score for the intended purposes. The Standards (2014) states:

> It is particularly important, when drawing inferences about an examinee's skills or abilities, to into account the individual characteristics of the test taker and how these characteristics may interact with the contextual features of the testing situation. (p. 53)

As discussed, these assessments have a remarkable impact on the educational policies and reform efforts of all countries involved. However, some have argued about validity issues in relation to whether these assessments can actually measure students' knowledge (Carnoy, 2015). When addressing validity issues, two major concerns must be addressed: construct underrepresentation and construct irrelevant variance (Messick, 1989). While construct underrepresentation is concerned with the material or content that is covered on the assessment or test, construct irrelevant variance addresses issues related to those factors that are outside of the realm of what an assessment is intended to measure. Consequently, threats to validity can undermine the interpretation of test scores. Specifically, addressing and identifying threats to the validity of the interpretation of test scores will provide valuable information regarding the ways in which educational governing bodies and policy makers can make informed decisions based on a clear picture of the assessment results (Mullis, Martin, Kennedy, Trong, & Sainsbury, 2009; Mullis, Martin, Ruddock, et al., 2009; OECD, 2013).

Whenever validity issues are in question it is important to first assess the purpose of an assessment or what it is intended to measure. Once that has been established it is critical to continually refer to the purpose of the assessment to understand the validity of the interpretation of the test scores. Problems arise when tests or assessments are used for purposes for which they were never intended to be used. Even more severe is utilizing a single measure to make major changes or modifications to educational policy and practice (AERA, APA, & NCME, 1999). Although test developers must outline appropriate uses of tests, it is the consumers who actually make choices about the way they are used, sometimes without the instruction or guidance of test developers. This shift or change in the way in which tests are used constitutes a major concern about the validity of the interpretation of test scores (Carnoy, 2015).

This view of fairness is based on the validity of the interpretation of test scores. Consequently, if the validity of the interpretation of test scores is in

question, one might assume that a test is not fair to the entire population of test takers. In addition to outlining the four views of test fairness, the Standards (2014) provided a list of potential threats to fairness and the valid interpretation of test scores.

Threats to Fair and Valid Interpretation of Test Scores

In addition to presenting the four general views of fairness, the Standards (2014) also included information about the potential threats to fairness and the valid interpretation of test scores including (a) test content, (b) test context, (c) test response, and (d) opportunity to learn. The Standards (2014) provided detailed descriptions of each of the four threats to fairness that is presented below.

Test Content

The first threat to fairness is related to the test content. Specifically, fairness issues may arise when the content of the test that is not included in the construct being measured provides an advantage or disadvantage to certain groups of test takers. The Standards (2014) stated:

> A test intended to measure critical reading, for example, should not include words and expressions especially associated with particular occupations, disciplines, cultural backgrounds, socioeconomic status, racial/ethnic groups or geographical locations, so as to maximize the measurement of the construct (the ability to read critically) and to minimize confounding of the measurement with prior knowledge and experience that are likely to advantage, or disadvantage, test takers from particular subgroups. (p. 54)

Differences in prior knowledge and experience should not impact or provide an advantage or disadvantage to a group of test takers. Consequently, a threat to fairness is evident when test content includes information that is advantageous for one group over another.

Test Context

The second threat to fairness refers to the test context that is in reference to the test and the testing environment. Researchers have found that test takers experience the testing environment in different ways. For example, stereotype threat has a significant impact on the test-taking experiences and performance for certain groups of test takers. Stereotype threat is defined

as a social–psychological threat that arises when one is in a situation or doing something for which a negative stereotype about his or her group applies. This predicament threatens an individual with being negatively stereotyped, with being judged or treated stereotypically, or with the prospect of conforming to the stereotype (Steele, 1997). When a negative stereotype about one's group becomes personally relevant, stereotype threat is the resulting sense that one's behavior or experience can be judged in terms of the stereotype. Consequently, the result is diminished performance in the domain of interest (Steele, Spencer, & Aronson, 2002). Several studies have investigated this phenomenon with regard to standardized test taking. Steele and Aronson (1995) examined how stereotype threat affected Black students' performance on a verbal section of the Graduate Record Examination (GRE). The findings showed that Black students did worse on the GRE verbal test than White participants in the high-stereotype threat condition. This study was the first of many to examine the negative impact that stereotype threat had on a certain group of test takers (i.e., Blacks, females). Additionally, research has shown that a high-stakes test environment affects test-takers' test-wiseness skills and strategy formation (Arbuthnot, 2009, 2011). The Standards (2014) raised other issues related to test context, including test instructions, complexity of tasks, and the interpersonal relationship of the examinee and the test taker.

One of the most important areas to consider regarding test context is how the language of the test impacts performance and poses threats to fairness. Specifically, the Standards (2014) highlighted the difficulties faced by bilingual and multilingual test takers. The Standards (2014) stated the following:

> Testing individuals that are bilingual or multilingual poses special challenges. An individual who knows two or more languages may not test well in one or more languages … Thus, in some settings, an understanding of an individual's type and degree of bilingualism or multilingualism important for testing the individual properly. (p. 55)

As stated, bilingual and multilingual test takers' experiences in the test-taking environment present additional obstacles for certain test takers that can be regarded as a threat to test fairness.

Test Response

The third threat to fairness is the variations in responses to certain test items. Specifically, items that can be correctly solved in multiple ways elicit this

threat. The Standards (2014) provided an example based on a constructed response item:

> For example, a scoring rubric for a constructed response item might reserve the highest score level for test takers who provide more information or elaboration than was actually requested. In this situation, test takers who simply follow instructions, or test takers who value succinctness in responses, will earn lower scores. (p. 56)

This example shows that a test may be assessing characteristics that are unrelated or irrelevant to the construct being measured. However, in this example, a student who answered the item correctly and provided a succinct answer would be penalized. Variations in test responses that are not part of the construct being measured could be regarded as a threat to fairness.

Opportunity to Learn

The last threat to fairness is whether all test takers have had the same opportunity to learn or to be instructed on the material that is covered on a standardized test or assessment. The Standards (2014) highlighted how differences in opportunity to learn have an impact on the interpretation of test scores. The Standards (2014) stated:

> Opportunity to learn – the extent to which individuals have had exposure to instruction or knowledge that affords them the opportunity to learn the content and skills targeted by the test – has several implications for the fair and valid interpretation of test scores for their intended uses. Individuals' prior opportunity to learn can be an important contextual factor to consider in interpreting and drawing inferences from test scores.

Investigating issues related to opportunity to learn helps to garner a better understanding of the test results and further ensure the validity of the interpretation of the test scores. Consequently, curricular differences confound the way in which we can understand and interpret test scores, and, more importantly, examine differences in test performance. The Standards (2014) stated:

> To the extent that inequity exists, the validity of inferences about student ability drawn from achievement test scores may be compromised. Not taking into account prior opportunity to learn could lead to misdiagnosis, inappropriate placement, and/or inappropriate assignment of services, which could have significant consequences for an individual.

As stated, examining students' opportunities to learn the tested material provides needed information when interpreting and reporting test performance and scores. Differences in the opportunity to learn among groups of

test takers can be interpreted as a potential threat to fairness, especially pertaining to understanding and interpreting test scores.

In sum, the Standards (2014) outlined varying views of fairness and potential threats to fairness. The Standards (2014) also provided one overarching standard regarding fairness:

> all steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the intended uses for all examinees in the intended population. (p. 63)

The Standards (2014) were designed to provide guidelines and principles for developing, administering, using, and evaluating tests. The Standards (2014) emphasized the identification and removal of any barriers, or sources of construct irrelevant variance, that would impede the ability to make comparable and valid interpretations of scores for all test takers. Consequently, the Standards (2014) should be utilized when evaluating fairness issues relative to international tests and assessments. Information regarding fairness issues on international assessments is presented in the next section.

Again, the Standards (2014) provided an overview of how to understand and interpret issues of test fairness. Each of the views address how differences in factors, such as race, ethnicity, disability, cultural background, and language can have an impact on test performance and the interpretation of the test scores from different groups. Policymakers in some countries have investigated test fairness issues related to standardized testing in their countries. For instance, in the United States concerns were voiced about issues related to the test fairness of high-stakes testing. In the early 2000s, United States Senator Paul Wellstone and Representative Bobby Scott proposed a bill that required the examination of fairness and accuracy in high-stakes standardized tests that students were required to participate in. This proposed bill stated:

> The serious and often adverse consequences resulting from the sole reliance on large-scale tests have increasingly resulted in questions and significant concerns by students, parents, teachers, and school administrators about how to ensure that such tests are used appropriately and in a manner, that is, fair. (H.R. 4333, 2000; S. 2348, 2000)

The senators recognized that multiple stakeholder groups needed to receive information about standardized tests and that tests needed to be deemed fair. The bill emphasized transparency from test developers as to how the tests were developed and how issues related to test fairness were handled. Additionally, the bill addressed issues of misuse. Although the bill did not pass, the proposal underscored the gravity of issues of test

fairness, particularly when the stakes are considered high. Issues of fairness are not promulgated to just one country, since international assessments are considered high stakes and involve many diverse groups of test takers; thus, it is imperative that fairness issues be addressed on these assessments as well.

Many would agree on the importance of investigating fairness issues on international assessments; such a task has many challenges. Diversity of the test takers on international assessments is wide, as international assessments include test takers from different countries, who have different languages, cultural norms, and so forth. Identifying and examining the experiences of all of the different groups of test takers would be a cumbersome task, especially since the availability of item-level test performance data on most international assessments is limited. Many international assessment programs do not release student level data for secondary data analysis. Consequently, the released data are reported in aggregate form, which presents several challenges when addressing issues related to measurement bias and opportunity to learn. Despite the challenges of investigating fairness issues, these issues with international assessments must be examined. Specifically, a close examination of the differences in test takers' performances and experiences across countries and cultural groups can provide information to evaluate and address issues of fairness.

CONCLUSION

This chapter began with highlighting the high stakes nature of international assessment programs and their critical impact on global educational theory, policy, and practice. As noted, research has shown that many countries found pride and honor in attaining a high ranking on these assessments, and those countries who find themselves struggling to achieve an acceptable ranking tend to respond by enacting educational policy reform to increase test performance. The research supports the far-reaching impact that international educational assessments have on global educational policy and practice choices. While international assessments have a tremendous impact on educational policy worldwide, there is still concern about the issues related to test fairness. The author implores that examining test fairness issues on international assessments is paramount to a clear understanding of differences in performance patterns between and among countries. The next chapter provides a systematic way to examine test fairness issues on international assessments.