RESEARCH & PRACTICE IN ASSESSMENT

VOLUME ELEVEN | SUMMER 2016
www.RPAjournal.com
ISSN # 2161-4120

A PUBLICATION OF THE VIRGINIA ASSESSMENT GROUP

RESEARCH & PRACTICE IN ASSESSMENT

Editorial Staff

Editor
Katie Busby
Tulane University

Associate Editor
Ciji A. Heiser
The University of North Carolina at Chapel Hill

Editoral Assistant
Sarah Andert

Lauren Germain
SUNY Upstate Medical University

Associate Editor
Megan Shaffer
Santa Clara University

Editorial Board

anthony lising antonio
Stanford University

Susan Bosworth
College of William & Mary

Jennifer A. Lindholm
University of California, Los Angeles

Daryl G. Smith
Claremont Graduate University

Linda Suskie
Assessment & Accreditation Consultant

John T. Willse
University of North Carolina at Greensboro

Ex–Officio Members

Virginia Assessment Group President
Timothy W. Merrill
Reynolds Community College

Virginia Assessment Group President–Elect
Lee Rakes
Viginia Military Institute

Past Editors

Robin D. Anderson
2006

Joshua T. Brown
2010-2014

Keston H. Fulcher
2007-2010

Review Board

Amee Adkins
Illinois State University

Robin D. Anderson
James Madison University

Angela Baldasare
University of Arizona

Brian Bourke
Murray State University

Chris Coleman
University of Alabama

Lindsey Jakiel Diulus
Nunez Community College

Dorothy C. Doolittle
Christopher Newport University

Seth Matthew Fishman
Villanova University

Teresa Flateby
Georgia Southern University

Brian French
Washington State University

Matthew Fuller
Sam Houston State University

Megan Moore Gardner
University of Akron

Karen Gentemann
George Mason University

Marc E. Gillespie
St. John's University

Molly Goldwasser
Duke University

Chad Gotch
Wahsington State University

Michele J. Hansen
IUPUI

Debra S. Harmening
University of Toledo

Ghazala Hashmi
J. Sargeant Reynolds Community College

S. Jeanne Horst
James Madison University

Natasha Jankowski
NILOA

Kendra Jeffcoat
San Diego State University Community College

Kimberly A. Kline
Buffalo State College

Kathryne Drezek McConnell
Association of American Colleges & Universities

Sean A. McKitrick
Middle States Commission

Deborah L. Moore
North Carolina State University

John V. Moore
Community College of Philadelphia

Ingrid Novodvorsky
University of Arizona

Loraine Phillips
University of Texas at Arlington

Suzanne L. Pieper
Northern Arizona University

William P. Skorupski
University of Kansas

Pamela Steinke
University of St. Francis

Matthew S. Swain
HumRRO

Wendy G. Troxel
Illinois State University

Catherine Wehlburg
Texas Christian University

Craig S. Wells
University of Massachusetts, Amherst

Thomas W. Zane
Salt Lake Community College

Carrie L. Zelna
North Carolina State University

TABLE OF CONTENTS

4 **FROM THE EDITOR**

Rewarded and Challenged
- Katie Busby

5 **ARTICLES**

Investigating the Dimensionality of Examinee Motivation Across Instruction Conditions in Low-Stakes Testing Contexts
- Sara J. Finney, Catherine E. Mathers & Aaron J. Myers

18 **The Use of Surveys in Teacher Education Programs to Meet Accreditation Standards: Preservice Teachers' Culturally Responsive Beliefs and Practices**
- Jason C. Immekus

29 **Higher Education Faculty Engagement in a Modified Mapmark Standard Setting**
- S. Jeanne Horst & Christine E. DeMars

42 **BOOK REVIEWS**

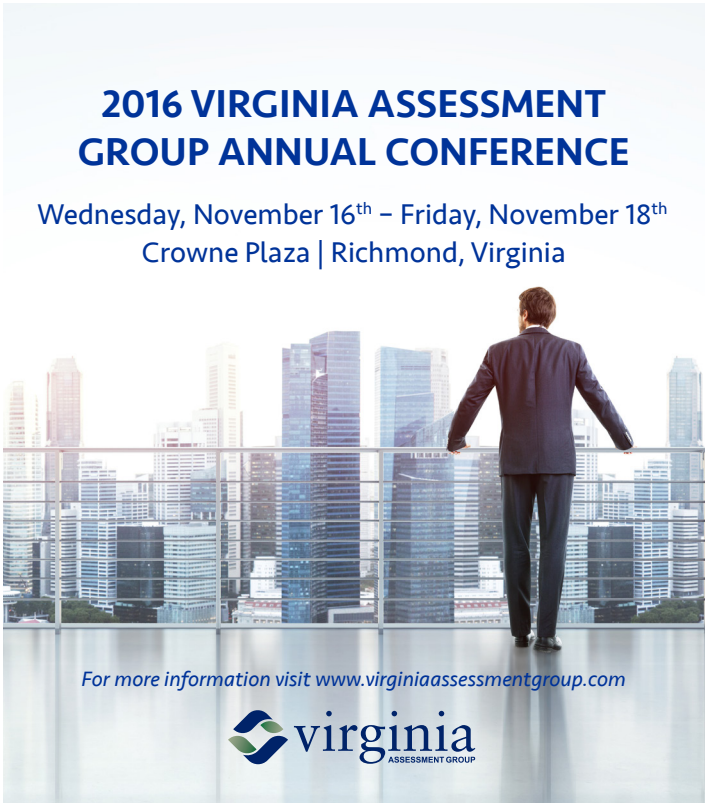
Book Review of: Pedigree: How Elite Students Get Elite Jobs
- Jamie Alea

45 **Book Review of: Service-Learning Essentials. Questions, Answers, and Lessons Learned**
- Agnieszka Nance

46 **NOTES IN BRIEF**

Assessment Practices in Higher Education and Results of the German Research Program Modeling and Measuring Competencies in Higher Education (KoKoHs)
- Olga Zlatkin-Troitschanskaia, Hans Anand Pant, Christiane Kuhn, Miriam Toepper & Corinna Lautenbach

55 **An Analysis of Programs Serving Men of Color in the Community College: An Examination of Funding Streams, Interventions, and Objectives**
- Fnann Keflezighi, Levi Sebahar & J. Luke Wood



CALL FOR PAPERS

Research & Practice in Assessment is currently soliciting articles and reviews for its Winter 2016 issue. Manuscripts submitted to RPA may be related to various higher education assessment themes, and should adopt either an assessment measurement or an assessment policy/foundations framework. Contributions are accepted at any time, but submissions received by August 1 will receive consideration for the winter issue. Manuscripts must comply with the RPA Submission Guidelines and be sent electronically to: editor@rpajournal.com

RESEARCH & PRACTICE IN ASSESSMENT

The goal of *Research & Practice in Assessment* is to serve the assessment community as an online journal focusing on higher education assessment. It is dedicated to the advancement of scholarly discussion amongst researchers and practitioners in this evolving field. The journal originated from the Board of the Virginia Assessment Group, one of the oldest continuing professional higher education assessment organizations in the United States. *Research & Practice in Assessment* is a peer-reviewed publication that uses a double-blind review process. Approximately forty percent of submissions are accepted for issues that are published twice annually. *Research & Practice in Assessment* is listed in Cabell's Directory and indexed by EBSCO, ERIC, Gale, and ProQuest.

Published by:
VIRGINIA ASSESSMENT GROUP | www.virginiaassessment.org

Publication Design by Patrice Brown | Copyright © 2016

FROM THE EDITOR

Rewarded & Challenged

Faculty, student affairs educators, and administrative leaders are rewarded and challenged as they engage in assessment activities to measure student learning, inform curriculum and program development, and respond to calls for accountability. It is rewarding to examine programs in a purposeful way and determine what impact those programs have on student learning and development, but those efforts are not without challenges. For example, researchers and practitioners must thoughtfully determine what measures of learning are needed and appropriate for a particular use and how the results of those measures will be used.

The contributions presented in this issue of *Research & Practice in Assessment* demonstrate how assessment scholars have addressed challenges faced in assessing students' skills and knowledge. Hopefully your work will benefit as a result of this collection of research and practice that takes place in the classroom and beyond.

The Summer 2016 issue includes three peer-reviewed articles that exemplify important assessment practices in higher education. Addressing the challenge of measuring examinee motivation in low-stakes testing situations, Finney, Mathers, and Myers investigate the psychometric properties of a popular measure of student motivation. Immekus examines the appropriate use of surveys to measure outcomes of a teacher preparation program and how those results are used by faculty for the purposes of program improvement and accreditation. Horst and DeMars present a standard setting procedure not often seen in higher education, and describe how this modified Mapmark procedure was used by faculty to make program and curricular decisions.

In the reviews, Alea examines the myth of upward mobility gained through higher education and the impact privilege has on student achievement in the review of *Pedigree: How Elite Students Get Elite Jobs*. Skills and knowledge are often reinforced through experiential activities such as service learning. Nance reviews *Service-Learning Essentials. Questions, Answers, and Lessons Learned*, an important examination of the pedagogy of service-learning.

This issue also includes two Notes in Brief highlighting the rewarding work taking place nationally and internationally. Zlatkin-Troitschanskaia, Pant, Kuhn, and Lautenbach discuss a comprehensive research initiative to assess skills and knowledge across programs and institutions in Germany and Austria. Keflezghi, Sebahar, and Wood examine similarities of minority male initiatives in higher education and propose a framework that can be used in future studies of these programs. I hope you find the scholarship in this issue both rewarding and challenging.

Regards,

Katie Busby

Tulane University

Abstract

Research investigating methods to influence examinee motivation during low-stakes assessment of student learning outcomes has involved manipulating test session instructions. The impact of instructions is often evaluated using a popular self-report measure of test-taking motivation. However, the impact of these manipulations on the psychometric properties of the test-taking motivation measure has yet to be investigated, resulting in questions regarding the comparability of motivation scores across instruction conditions and the scoring of the measure. To address these questions, the factor structure and reliability of test-taking motivation scores were examined across instruction conditions during a low-stakes assessment session designed to address higher education accountability mandates. Incoming first-year college students were randomly assigned to one of three instruction conditions where personal consequences associated with test results were incrementally increased. Confirmatory factor analyses indicated a two-factor structure of test-taking motivation was supported across conditions. Moreover, reliability of motivation scores was adequate even in the condition with greatest personal consequence, which was reassuring given low reliability has been found in high-stakes contexts. Thus, the findings support the use of this self-report measure for the valuable research that informs motivation instruction interventions for low-stakes testing initiatives common in higher education assessment.

AUTHORS

Sara J. Finney, Ph.D.
James Madison University

Catherine E. Mathers
James Madison University

Aaron J. Myers
James Madison University

CORRESPONDENCE

Email
finneysj@jmu.edu

Investigating the Dimensionality of Examinee Motivation Across Instruction Conditions in Low-Stakes Testing Contexts

Institutional accountability mandates prompt assessment of student learning (U.S. Department of Education, 2006). Although designed to accurately assess learning, many "accountability tests" are low stakes for students, meaning there are no personal consequences associated with performance for the examinee. Nonetheless, these tests are high stakes for universities in that scores are used to inform the evaluation and modification of programs, comparisons across institutions, accreditation, and resource allocation. With the prevalence of tests that are low stakes for examinees come issues that require attention from the assessment community. One such issue is the role that examinee motivation plays in low-stakes assessment contexts and its measurement.

The Need to Report Examinee Motivation

Examinee motivation is inherently linked to the validity of assessment interpretations. The more motivated examinees are to perform well, the better the test scores reflect ability (Wise & DeMars, 2005). Effortless test performance due to low motivation complicates inferences from test scores. Thus, low-stakes testing contexts, in particular, may result in test scores that are difficult to interpret for accreditation, strategic planning, and accountability purposes.

Consequently, score interpretations should be made in accordance with the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). The *Standards* state, "In evaluation or accountability settings, test results should be used in conjunction with information from other sources when the use of the additional information contributes to the validity of the overall interpretation"

(p. 213). There is a specific call for information regarding "the degree of motivation of the test takers" in nonconsequential testing conditions as part of Standard 13.9.

Measuring Examinee Motivation via the Student Opinion Scale (SOS)

Reporting and interpreting examinee motivation requires its measurement. One particularly popular self-report measure of examinee motivation is the 10-item Student Opinion Survey (SOS; Thelk, Sundre, Horst, & Finney, 2009). The SOS has been implemented as a measure of examinee motivation in at least 9 countries, 33 universities, and 30 published studies (Sessoms & Finney, 2015). It has been used to examine the relationship between motivation and test performance (e.g., Abdelfattah, 2010; Swerdzewski, Harmes, & Finney, 2009; Wise & DeMars, 2005; Zilberberg, Finney, Marsh, & Anderson, 2014), to identify and filter out test scores from examinees with low motivation (e.g., Rios, Liu, & Bridgeman, 2014; Steedle, 2014; Swerdzewski, Harmes, & Finney, 2011), to examine personality characteristics that correlate with test-taking motivation (e.g., Barry, Horst, Finney, Brown, & Kopp, 2010; Barry & Finney, 2016; DeMars, Bashkov, & Socha, 2013; Kopp, Zinn, Finney, & Jurich, 2011), and to evaluate methods for increasing test-taking motivation (e.g., Finney, Sundre, Swain, & Williams, 2016; Hawthorne, Bol, Pribesh, & Suh, 2015; Liu, Bridgeman, & Adler, 2012; Steedle, 2010; Waskiewicz, 2011). The appropriateness of the use of the SOS for the latter is the focus of the current study.

The SOS was developed using expectancy-value (EV) theory (Wigfield & Eccles, 2000; Wolf & Smith, 1995). To determine the level of expended effort on a task, an individual considers (a) how well they *expect* to perform, and (b) the perceived *value* the task provides. EV theory applied to the context of test taking assumes an examinee's expended effort on the test is a function of their expected test performance and perceived value of the test. Assessing task value is essential in low-stakes testing contexts: examinees completing a test with no personal consequences for performance will likely put forth less effort because doing well has no attainment, intrinsic, or utility value. Thus, the resulting test scores may not be accurate representations of student ability (Wise & DeMars, 2005). This indirect effect of perceived test value on test performance (via test-taking effort) has been empirically supported in low-stakes contexts (Cole, Bergin, & Whittaker, 2008; Zilberberg et al., 2014).

The SOS was designed to operationalize the expended effort and test value components of test-taking motivation. Effort is defined as the level of effort expended toward test completion (e.g., "I engaged in good effort throughout this test"). Test value is defined as how important doing well is to the examinee (e.g., "Doing well on this test was important to me"). Again, theoretically, perceived importance influences expended effort. That is, importance and effort are considered theoretically distinct constructs. Empirical study of the dimensionality of the SOS scores has supported a two-factor structure over a one-factor structure of motivation in low-stakes testing contexts (e.g., Thelk et al., 2009). Invariance of the two-factor structure was found across age groups, gender, test modality (Thelk et al., 2009), and time (Sessoms & Finney, 2015) in low-stakes testing contexts.

Considering previous research examining the factor structure of noncognitive measures suggests dimensionality can differ across testing contexts (e.g., Barry & Finney, 2009; De Leeuw, Mellenbergh, & Hox, 1996), it is curious there have been no empirical studies assessing if the factor structure of the SOS is affected as the stakes or consequences of the test change. A difference in factor structure could impact the scoring of the SOS and, more important, could suggest test-taking motivation is conceptualized differently in different testing contexts. This issue becomes particularly important given the use of the SOS to evaluate the impact of increasing test consequences via test instructions. As called for in a recent issue of *Research & Practice in Assessment*, "Research on instruments that examine test-taker motivation on low-stakes tests is growing, but more is needed to fill the existing gap in the literature regarding examinee reactions to tests and the test conditions that affect performance and motivation" (Hawthorne et al., 2015, p. 36). We addressed this call by investigating the potential change in the psychometric properties of the SOS as the personal relevance and consequences for examinees were increased across test instruction conditions, as detailed below.

Evaluating Motivation Instruction Interventions Using the SOS

Given the relationship between examinee motivation and test scores in low-stakes testing contexts, assessment practitioners have investigated ways to increase motivation. One obvious solution is to increase the stakes for examinees (e.g., test scores impact grades). There are considerable complexities associated with a high-stakes testing program, which include the need to guard against and monitor cheating, the need for larger item pools for re-testing after remediation, the influence of test anxiety on test scores, and resistance from faculty (Wise & DeMars, 2005). Another option is to provide monetary compensation for performance (e.g., O'Neil, Sugrue, & Baker, 1995), which necessitates immense financial resources. Moreover, monetary incentives have not been proven consistently effective in improving test performance (O'Neil, Abedi, Miyoshi, & Mastergeorge, 2005).

Effortless test performance due to low motivation complicates inferences from test scores. Thus, low-stakes testing contexts, in particular, may result in test scores that are difficult to interpret for accreditation, strategic planning, and accountability purposes.

Another option presently receiving attention is motivation instruction interventions (e.g., Finney et al., 2016; Hawthorne et al., 2015; Kornhauser, Minahan, Siedlecki, & Steedle, 2014; Liu et al., 2012; Liu, Rios, & Borden, 2015; Waskiewicz, 2011). Motivation instruction studies involve evaluating the impact of test session instructions on examinee motivation and test performance. Of note, the test remains low stakes for examinees in that scores do not inform grades, graduation, or other academic outcomes. Instead, the instructions manipulate the message conveyed to examinees with the goal of making the test more personally relevant (see Appendix for a representative set of instructions). This active area of research may uncover an approach to influence examinee motivation in low-stakes contexts that requires no financial or human resources.

The effectiveness of motivation instructions is often evaluated using the SOS. That is, researchers examine if SOS scores differ, on average, across instruction conditions, with the goal of identifying instructions that increase motivation while preserving the low-stakes nature of the test. There is a considerable implicit assumption to this approach—one assumes that different instructions will potentially result in different average levels of motivation, yet other properties of the scores, such as the factor structure or reliability, will not be impacted. Of note, mean differences provide no insight into the stability of the factor structure and, hence, the scoring of the SOS; however, nonambiguous interpretation of mean differences necessitates no difference in factor structure across conditions. Surprisingly, there has been no empirical study evaluating if the factor structure of the SOS remains consistent across instruction conditions.

Is it reasonable to believe the factor structure of the SOS may change as instructions increase the personal relevance of the test for students? In a high-stakes context, the level of test importance should be high for examinees *and* they should put forth a great deal of effort. It is reasonable to presume that in a high-stakes testing environment, effort and importance may become indistinguishable (i.e., motivation becomes unidimensional). If this is the case, importance and effort items are interchangeable; an item from either subscale provides the same information regarding motivation. In turn, computing two subscales would no longer be appropriate. The factor structure of the SOS scores has not been examined in high-stakes contexts. Instead, the two-factor structure found in low-stakes contexts is simply assumed to generalize to high-stakes contexts, as reflected in the computation of the two subscales of effort and importance in high-stakes contexts. Importantly, there is evidence that the reliability of SOS scores differs across high- and low-stakes settings (Thelk et al., 2009). In high-stakes contexts, the SOS was sensitive to a ceiling effect, which decreased score variability, and in turn dramatically decreased estimates of internal consistency reliability. Given these results in high-stakes contexts, the reliability and dimensionality of SOS scores may differ across instruction conditions in low-stakes contexts.

Furthermore, this possibility of differing psychometric properties across instruction conditions is coupled with the perplexing practice by some researchers of scoring the SOS as a total motivation score (e.g., Kornhauser et al., 2014; Liu et al., 2012; Liu et al., 2015; Steedle, 2010). It is unclear if the authors of these studies uncovered a unidimensional solution when implementing motivation instructions and, hence, adapted the scoring of the SOS to align with this new conceptualization (i.e., a total SOS motivation score). If instruction

condition did impact the factor structure, this would imply a strong effect of instructions—the conceptualization of motivation differs depending on the instructions the examinees receive. Obviously, a difference in factor structure across instruction conditions makes comparisons of average motivation level across conditions obsolete.

Purpose of the Study

Using an operational low-stakes institutional accountability testing program, we examined the effects of gradually increasing test consequences on the psychometric properties of a popular measure of examinee motivation. Specifically, our purpose was to assess if the dimensionality and reliability of SOS scores differed across testing sessions that employed three different motivation instructions. Using confirmatory factor analysis (CFA) and SOS data from the three instruction conditions, we assessed the fit of the two-factor structure previously supported in low-stakes testing contexts and a one-factor structure implied by the creation of one total motivation score.

Assessing task value is essential in low-stakes testing contexts: examinees completing a test with no personal consequences for performance will likely put forth less effort because doing well has no attainment, intrinsic, or utility value.

Methods

Participants & Procedures

All students at a mid-sized university in the mid-Atlantic United States are required to participate in a three-hour large-scale testing session twice during their academic careers, once as incoming first-year students and again when they have accumulated 45-70 credit hours. Given students complete the same exams at both time points, this data collection scheme affords the computation of value-added scores associated with general education coursework. During the testing session, all students complete a battery of cognitive and noncognitive measures tied to general education and student affairs program objectives. Testing rooms differ in the exact measures administered and the size of the room (25 to 130 seats). Students are randomly assigned to testing room and, therefore, test configuration to ensure the desired sample size for each test. Although some students complete the tests via computer, the vast majority complete the tests via paper and pencil. All students selected for our study completed the tests via pencil and paper. Proctors in each room distribute and collect materials, read instructions, and encourage students to give their best effort. Test scores have no impact on students' academic record or graduation but do provide data for institutional accountability purposes.

Using data collected from this operational low-stakes testing program offered the unique and convenient opportunity to evaluate the impact of instructions on the psychometric properties of the SOS in an authentic testing environment. More specifically, the data analyzed in the current study were collected from incoming first-year students engaged in this testing program. To investigate the effects of test instructions on the factor structure of the SOS, students were randomly assigned to one of three conditions that incrementally increased the "dose" of the personal relevance of the test via the test instructions (see Appendix). In the first condition, students were instructed that their scores would be used in aggregate form at the institutional level (Institutional Condition). The Institutional instructions are the standard instructions all students have received over the past two decades of accountability testing at the institution. In the second condition, students were told that their scores would be used at the institutional level and their personal score would be available for individual feedback (Feedback Condition). In the final condition, students were told that their scores would be used at the institutional level, their personal score would be available to them for feedback, *and* their personal score would be released to faculty (Personal Condition). We purposefully selected these instructions as they are realistic in low-stakes testing contexts. Previous study of the SOS was conducted only in the Institutional condition; thus, it was unclear if the SOS would function adequately if institutions employed instructions similar to the Feedback and Personal condition instructions.

The assigned test instructions were read aloud to students and projected on the screen in front of the room. Proctors can positively affect effort (Lau, Swerdzewski, Jones, Anderson, & Markle, 2009); thus, all proctors received standardized training regarding administering

test instructions. Proctors were trained to draw students' attention to the test instructions to ensure the experimental conditions were understood. Furthermore, instructions for all conditions were presented with colored text (black for Institutional, blue for Feedback, & red for Personal) to draw attention to the conditions. Moreover, proctor gender, ethnicity, and age were held constant across instruction conditions to minimize any potential effect on motivation.

The study utilized one test configuration that was standardized across the instruction conditions. This configuration contained an arduous measure of scientific reasoning, which was administered first in the testing session, immediately followed by the SOS. Thus, student responses to the SOS represented students' perceived importance and expended effort for the scientific reasoning test just completed.

Of the 3,976 incoming first-year students engaged in the testing program, 1,287 were randomly assigned to one of the three test instruction conditions. A small proportion of students did not answer all SOS items, thus the effective sample size was reduced to N = 1,245. Of these students, 61.37% were female and the average age was 18.44 years. Students could self-identify in more than one ethnicity category, which resulted in 88.92% identifying as White; 5.06% as Black; 4.90% as Hispanic; 5.62% as Asian; 1.85% as American Indian; 0.96% as Pacific Islander; and 1.85% did not specify an ethnicity. These sample demographics align with the university demographics. At the university, 60% of students are female; 77.78% identify as White; 4.43% as Black; 5.75% as Hispanic; 4.35% as Asian; 0.18% as American Indian; 0.13% as Pacific Islander; and 3.48% unspecified. Of the 1,245 students, 385 received Institutional instructions, 385 received Feedback instructions, and 475 received Personal instructions. More examinees received the Personal instructions than the Institutional and Feedback instructions because this was the first administration of the Personal instructions, whereas Institutional and Feedback instructions had been administered in previous years.

Measures

To evaluate the dimensionality of the SOS across the three testing conditions, examinees completed a cognitive test of scientific reasoning and then immediately indicated their motivation with respect to that scientific reasoning test.

Scientific Reasoning Test. Scientific reasoning was assessed using the Natural World Test, Version 9 (SR; Sundre & Thelk, 2010; Sundre, Thelk, & Wigtil, 2008), a 66-item cognitive test designed to measure students' scientific reasoning skills. This test has been in use since its creation in 1996. It was designed to assess the scientific reasoning student learning objectives upon which a 10-12 credit hour curriculum has been designed. Faculty who teach this curriculum wrote every test item. Thus, the learning objectives and curriculum have been aligned. This cognitively demanding test typically takes an hour to complete. This test was the first test completed in the testing session, followed immediately by the SOS.

The [Student Opinion Scale] SOS was designed to operationalize the expended effort and test value components of test-taking motivation.

SOS. The Student Opinion Scale (Thelk et al., 2009) is a 10-item, self-report measure of test-taking motivation consisting of five effort items and five importance items (see Table 1). The SOS instructions referred to the scientific reasoning test and the SOS was completed directly after the scientific reasoning test. The Effort subscale consists of five items that measure the degree to which examinees put forth effort on a given test (e.g., "I gave my best effort on this test"). The Importance subscale consists of five items that measure the degree to which examinees view a given test as important (e.g., "Doing well on this test was important to me"). Examinees responded to the items using a scale of 1 (Strongly Disagree) to 5 (Strongly Agree).

Results

Data Screening

Prior to formally testing the fit of the one-factor and two-factor models to the SOS data from the three instruction conditions, the item-level data were examined (see Table 1). Inter-item correlations foreshadowed the dimensionality. In general, correlations among effort items were stronger than correlations among effort and importance items. Likewise, correlations

We addressed this call by investigating the potential change in the psychometric properties of the SOS as the personal relevance and consequences for examinees were increased across test instruction conditions...

Table 1
Correlations and Descriptive Statistics for the SOS by Test Instruction Condition

Institutional Condition (n = 385)										
	Item									
Item	1	3	4	5	8	2	6	7	9	10
1. Doing well on this test was important to me. [I]	1									
3. I am not curious about how I did on this test relative to others. [I]*	.252	1								
4. I am not concerned about the score I receive on this test. [I]*	.469	.495	1							
5. This was an important test to me. [I]	.585	.260	.483	1						
8. I would like to know how well I did on this test. [I]	.466	.530	.484	.393	1					
2. I engaged in good effort throughout this test. [E]	.423	.205	.234	.280	.440	1				
6. I gave my best effort on this test. [E]	.451	.180	.221	.341	.408	.601	1			
7. While taking this test, I could have worked harder on it. [E]*	.262	.167	.229	.273	.229	.436	.586	1		
9. I did not give this test my full attention while completing it. [E]*	.384	.179	.257	.281	.422	.519	.606	.590	1	
10. While taking this test I was able to persist to completion of the task. [E]	.265	.066	.113	.197	.327	.490	.387	.310	.421	1
Mean	3.633	3.321	3.272	2.974	3.672	4.121	4.015	3.214	3.861	3.964
SD	0.896	1.155	1.073	0.927	1.032	0.759	0.855	1.103	0.941	0.823
Skew	-0.438	-0.407	-0.384	0.071	-0.631	-1.131	-0.924	-0.211	-0.955	-1.082
Kurtosis	0.321	-0.637	-0.480	-0.183	0.010	2.582	1.246	-0.800	0.823	2.196

Feedback Condition (n = 385)										
1. Doing well on this test was important to me. [I]	1									
3. I am not curious about how I did on this test relative to others. [I]*	.317	1								
4. I am not concerned about the score I receive on this test. [I]*	.356	.329	1							
5. This was an important test to me. [I]	.497	.243	.484	1						
8. I would like to know how well I did on this test. [I]	.390	.484	.332	.313	1					
2. I engaged in good effort throughout this test. [E]	.500	.263	.329	.372	.338	1				
6. I gave my best effort on this test. [E]	.486	.304	.229	.253	.350	.626	1			
7. While taking this test, I could have worked harder on it. [E]*	.386	.193	.252	.277	.196	.540	.576	1		
9. I did not give this test my full attention while completing it. [E]*	.373	.198	.257	.331	.264	.613	.564	.654	1	
10. While taking this test I was able to persist to completion of the task. [E]	.284	.231	.089	.126	.252	.448	.439	.355	.421	1
Mean	3.722	3.479	3.405	3.003	3.868	4.129	4.018	3.163	3.835	3.987
SD	0.781	1.025	0.958	0.880	0.855	0.736	0.865	1.116	1.002	0.818
Skew	-0.451	-0.544	-0.496	0.155	-0.818	-0.755	-0.855	-0.134	-0.813	-0.722
Kurtosis	0.465	-0.260	-0.051	0.022	1.109	0.975	0.885	-0.837	0.190	0.574

Feedback Condition (n = 385)										
1. Doing well on this test was important to me. [I]	1									
3. I am not curious about how I did on this test relative to others. [I]*	.317	1								
4. I am not concerned about the score I receive on this test. [I]*	.356	.329	1							
5. This was an important test to me. [I]	.497	.243	.484	1						
8. I would like to know how well I did on this test. [I]	.390	.484	.332	.313	1					
2. I engaged in good effort throughout this test. [E]	.500	.263	.329	.372	.338	1				
6. I gave my best effort on this test. [E]	.486	.304	.229	.253	.350	.626	1			
7. While taking this test, I could have worked harder on it. [E]*	.386	.193	.252	.277	.196	.540	.576	1		
9. I did not give this test my full attention while completing it. [E]*	.373	.198	.257	.331	.264	.613	.564	.654	1	
10. While taking this test I was able to persist to completion of the task. [E]	.284	.231	.089	.126	.252	.448	.439	.355	.421	1
Mean	3.722	3.479	3.405	3.003	3.868	4.129	4.018	3.163	3.835	3.987
SD	0.781	1.025	0.958	0.880	0.855	0.736	0.865	1.116	1.002	0.818
Skew	-0.451	-0.544	-0.496	0.155	-0.818	-0.755	-0.855	-0.134	-0.813	-0.722
Kurtosis	0.465	-0.260	-0.051	0.022	1.109	0.975	0.885	-0.837	0.190	0.574

among importance items were generally stronger than correlations among importance and effort items. This pattern suggests better fit for a two-factor than a one-factor model across all three instruction conditions.

Moreover, data were screened to assess univariate and multivariate normality as this impacts the choice of estimation method when formally estimating the models. Across conditions, all items were univariately normal with skew values less than |1.2| and kurtosis values less than |2.2|. Mardia's multivariate kurtosis coefficients ranged from 150.01 to 160.05 across the three conditions. Given multivariate non-normality, we chose an estimation method that accounts for the multivariate kurtosis of the data (Finney & DiStefano, 2013). That is, the CFA models were estimated using maximum likelihood estimation and the Satorra-Bentler adjustment was used to adjust the fit indices and the standard errors (Satorra & Bentler, 1994).

The factor structure of the SOS scores has not been examined in high-stakes contexts. Instead, the two-factor structure found in low-stakes contexts is simply assumed to generalize to high-stakes contexts, as reflected in the computation of the two subscales of effort and importance in high-stakes contexts.

Model-Data Fit

Two models were fit to the data in each condition: a two-factor model that aligns with the development of the SOS and a one-factor model that aligns with the (questionable) use of a total score. The robust root mean square error of approximation (RMSEA), robust comparative fit index (CFI), and standardized root mean square residual (SRMR) were used to assess global model-data fit. When using the Satorra-Bentler correction for multivariate non-normality, the following cutoffs have been suggested as indicators of good model fit: robust RMSEA $\leq .05$, robust CFI $\geq .95$, and SRMR $\leq .07$ (Yu & Muthén, 2002). However, because the cutoffs are based on one study and are considered overly sensitive (i.e., result in rejecting adequate models), suggested cutoffs should be used as guidelines rather than strict criteria (Marsh, Hau, & Wen, 2004). Moreover, global fit indices simply summarize the overall model-data fit, whereas correlation residuals indicate local misfit of a model (under- or overestimated relationships between items). Correlation residuals greater than |.15| were flagged for inspection.

Global fit indices and the number of correlation residuals greater than |.15| are located in Table 2. Not surprisingly, the χ^2_{SB} values were significant for both models; χ^2 tests are influenced by sample size, thus slight model misfit will be statistically significant with large samples. All global fit indices for the one-factor model were unsatisfactory within each condition. The large number of correlation residuals (ranging from |.16| to |.30| across conditions) reiterates the lack of fit of the one-factor model. In short, the one-factor model does not represent the data well.

The SRMR and most robust CFI values were satisfactory for the two-factor model across all conditions. The robust CFI within the Institutional condition was only slightly below the cutoff. RMSEA values did not meet the suggested cutoff; however, the correlation residuals indicated acceptable fit of the two-factor model, aligning with the use of the fit index cutoff as an imprecise guideline (Marsh et al., 2004). Only one residual was greater than |.15| in the Feedback condition. The correlation residual between items 3 and 8 was .18, indicating the relationship between these items was underestimated by the two-factor model. Although both items represent the Importance subscale, the relatively larger observed correlation between these items compared to the other importance items may be due to a wording effect. Given satisfactory fit of the two-factor model across conditions, the latent factor correlation and reliability estimates were examined to further investigate the effects of test instructions.

Using confirmatory factor analysis (CFA) and SOS data from the three instruction conditions, we assessed the fit of the two-factor structure previously supported in low-stakes testing contexts and a one-factor structure implied by the creation of one total motivation score.

Factor Correlation and Reliability

The correlation between the Effort and Importance factors was .61, .68, and .69 for the Institutional, Feedback, and Personal conditions, respectively. Notice the correlation increased negligibly as test consequences increased.[1] Moreover, the highest factor correlation indicated the two factors were related but not redundant. Importantly, internal consistency reliability of the Effort subscale scores (α = .83, .84, .84) and Importance subscale scores (α = .79, .74, .81) were adequate across the Institutional, Feedback, and Personal conditions, thus supporting their use. The magnitude and similarity of the reliability estimates were expected given the values of the factor loadings across conditions (see Table 3). Notice the relationship between each item and the corresponding factor differ negligibly across instruction condition.

Table 2

Model Fit Indices for One-Factor and Two-Factor Models of the Student Opinion Scale

Model	*df*	χ^2_{SB}	$RMSEA_{SB}$	CFI_{SB}	SRMR	Correlation Residuals > |.15|
Institutional						
One-Factor	35	309.97*	0.16	0.78	0.110	7
Two-Factor	34	142.23*	0.09	0.93	0.067	0
Feedback						
One-Factor	35	187.84*	0.12	0.89	0.089	4
Two-Factor	34	97.00*	0.07	0.96	0.057	1
Personal						
One-Factor	35	366.92*	0.16	0.84	0.091	6
Two-Factor	34	166.36*	0.09	0.95	0.054	0

Note. χ^2_{SB} = Satorra-Bentler chi-square; $RMSEA_{SB}$ = robust root mean square error of approximation; CFI_{SB} = robust comparative fit index; SRMR = standardized root mean square residual. Chi-square difference tests were not conducted to compare the ill-fitting one-factor model to the two-factor model. Nested model difference tests are useful when evaluating competing models that represent the data well; thus, a nested model comparison is unnecessary here when one of two models grossly misrepresents the data (Bandalos & Finney, 2010).
*$p < .001$.

In short, effort and importance items were well represented by two correlated factors of effort and importance, not one over-arching motivation factor, when students were told their scores were available to them personally and when students were told their personal scores could be viewed by faculty.

Discussion and Implications

Given previous test-taking motivation research, support for the two-factor structure of test importance and expended effort in the Institutional condition was not surprising. The SOS has consistently been shown to be comprised of two moderately correlated yet distinct factors when examinees are told test scores are used solely for institutional accountability purposes. Our results reinforce the idea that when test scores have no personal relevance to students, test-taking motivation is *not* unidimensional in structure, and should not be scored as one total score. We found identical results when we increased the personal relevance or consequence for students. In short, effort and importance items were well represented by two correlated factors of effort and importance, not one over-arching motivation factor, when students were told their scores were available to them personally and when students were told their personal scores could be viewed by faculty.

Moreover, for the Institutional condition, 37.21% of the variance was shared between the effort and importance factors. When additional consequences were added in the Feedback and Personal conditions, 45.69% and 47.19% of the variance was shared, respectively. These results suggest as consequences are increased, effort and importance become only slightly less distinct. Despite the slight convergence of the two factors, most of the variance associated with

1 Formal measurement invariance tests were conducted to assess not only the equivalence of the factor structure across conditions (the focus of the current study), but also the equivalence of the factor pattern coefficients (i.e., equal unstandardized factor loadings, which is typically referred to as metric invariance), the covariance between Effort and Importance factors, and the correlation between Effort and Importance factors. All models (configural invariance, metric invariance, factor covariance invariance, and factor correlation invariance) fit well in an absolute sense (i.e., adequate values of fit indices) and, each model did not fit worse than the baseline configural model. Hence, in addition to the SOS having the same two-factor structure across motivation instruction conditions, the SOS items also had equal saliency to the factors across conditions (i.e., metric invariance) and an equivalent relationship between Effort and Importance factors across conditions.

Table 3
Standardized Factor Pattern Coefficients Across Motivation Instruction Conditions

	Institutional		Feedback		Personal	
Item	Effort	Importance	Effort	Importance	Effort	Importance
1. Doing well on this test was important to me.		.71		.72		.78
3. I am not curious about how I did on this test relative to others.		.54		.51		.59
4. I am not concerned about the score I receive on this test.		.69		.58		.69
5. This was an important test to me.		.66		.64		.70
8. I would like to know how well I did on this test.		.72		.58		.66
2. I engaged in good effort throughout this test.	.72		.80		.79	
6. I gave my best effort on this test.	.81		.77		.87	
7. While taking this test, I could have worked harder on it.	.69		.74		.68	
9. I did not give this test my full attention while completing it.	.77		.78		.75	
10. While taking this test I was able to persist to completion of the task.	.53		.54		.56	

Note. Given each item represents only one factor, the values above can be interpreted as correlations and squared to indicate the amount of variance explained in the item by the factor.

Fortunately, the current study supports the use of the SOS for the continued evaluation of motivation instruction interventions.

effort and importance was not shared, further supporting the distinction between perceived test importance and expended effort in low-stakes contexts.

Given these results, in low-stakes testing contexts the SOS may be perceived as having increased utility for two purposes: (a) reporting and interpreting examinee motivation to align with the *Standards for Educational and Psychological Testing* (2014); and (b) researching the effectiveness of motivation instructions. Regarding the first purpose, as noted earlier, when gathering data for accountability purposes in low-stakes testing contexts, assessment practitioners should collect and interpret examinee motivation information to inform inferences from test scores. We realize that the instructions communicated to university students in these low-stakes contexts differ across institutions, and those differences are tied to the personal relevance of the scores for students. We have provided evidence that the SOS importance and effort scores are appropriate to report and interpret in low-stakes contexts that differ in the message conveyed to students.

Regarding the second purpose, additional study of the effectiveness of motivation instructions is needed given previous research employs small samples (e.g., Hawthorne et al., 2015; Kornhauser et al., 2014), relies on volunteers who may not represent the university population (e.g., Liu et al., 2012; Liu et al., 2015), utilizes institutions with a fairly homogenous demographic composition (e.g., Finney et al., 2016), and confounds instruction

interventions with financial incentives (e.g., Liu et al., 2012; Liu et al., 2015). Fortunately, the current study supports the use of the SOS for the continued evaluation of motivation instruction interventions.

Although obvious, we feel it is important to reiterate that the SOS is a self-report measure. Assessment practitioners must rely on examinees providing responses to the SOS that represent true levels of perceived test importance and expended effort. Measures of effort such as response time effort (RTE) do not rely on accurate self-reporting but rather actual behavior as indexed by time (Wise & Kong, 2005). If self-report measures are necessary given lack of access to computerized testing to gauge RTE (as was the case in the current study), there is evidence of the alignment between RTE and self-report SOS scores (e.g., Rios et al., 2014; Swerdzewski et al., 2011). Nevertheless, we encourage gathering multiple measures of motivation when possible to provide additional insight into the effectiveness of motivation interventions and further validity evidence for self-report measures. Moreover, this study was based on a large, representative sample of first-year students, thus results should not be generalized to other student populations. We encourage researchers to conduct additional study of the properties of the SOS in motivation instruction conditions using other student populations.

In conclusion, the expectation for institutions to collect outcomes assessment data is not expected to decline, thus low-stakes testing will likely remain prevalent in higher education contexts. Consequently, the need to report and interpret examinee motivation will remain critical, as will the need to uncover a feasible intervention to increase motivation in these contexts. Fortunately, the SOS can be utilized for both purposes, allowing assessment practitioners to focus on possible solutions to the vexing problem of examinee motivation rather than its measurement.

Appendix

Institutional Condition Test Instructions

Please make sure you have correctly filled in your name and ID number on the scan form. After you have done this please write NW-9 in the top right corner of the scan form.

The Natural World test is designed to assess your quantitative and scientific reasoning. At this university, we define these as thinking processes for obtaining and evaluating knowledge of the natural world. This instrument was developed by faculty who teach in the university's award winning General Education program. The results are used to inform and improve our General Education program.

You will have 60 minutes to complete the 66 multiple-choice items on this test. You will have a piece of scrap paper to help you during this test. If you need more scrap paper, raise your hand. Make sure to read all test directions carefully, and answer the items to the best of your ability.

Thank you in advance for your effort and concentration on this important test. You may begin.

Feedback Condition Test Instructions

Please make sure you have correctly filled in your name and ID number on the scan form. After you have done this please write NW-9 in the top right corner of the scan form.

The Natural World test is designed to assess your quantitative and scientific reasoning. At this university, we define these as thinking processes for obtaining and evaluating knowledge of the natural world. This instrument was developed by faculty who teach in the university's award winning General Education program. The results are used to inform and improve our General Education program.

You will have 60 minutes to complete the 66 multiple-choice items on this test. You will have a piece of scrap paper to help you during this test. If you need more scrap paper, raise your hand. Make sure to read all test directions carefully, and answer the items to the best of your ability.

We are pleased to let you know that you will be able to find out how you scored on the quantitative and scientific reasoning measures and what your scores tell you about these reasoning skills. Later in the semester, you will receive an e-mail providing you with instructions for accessing your scores and the interpretive information.

When you become eligible for assessment again as a sophomore or junior, we will make every effort to make sure that you are assigned to take this instrument again so you can compare those scores with the ones you earn today. It is the hope of the faculty that we will see a lot of growth and development. We are pleased to offer this feedback to you.

Thank you in advance for your effort and concentration on this important test. You may begin.

Personal Condition Test Instructions

Please make sure you have correctly filled in your name and ID number on the scan form. After you have done this please write NW-9 in the top right corner of the scan form.

The Natural World test is designed to assess your quantitative and scientific reasoning. At this university, we define these as thinking processes for obtaining and evaluating knowledge of the natural world. This instrument was developed by faculty who teach in the university's award winning General Education program. The results are used to inform and improve our General Education program.

You will have 60 minutes to complete the 66 multiple-choice items on this test. You will have a piece of scrap paper to help you during this test. If you need more scrap paper, raise your hand. Make sure to read all test directions carefully, and answer the items to the best of your ability.

We are pleased to let you know that you will be able to find out how you scored on the quantitative and scientific reasoning measures and what your scores tell you about these reasoning skills. Later in the semester, you will receive an e-mail providing you with instructions for accessing your scores and the interpretive information.

When you become eligible for assessment again as a sophomore or junior, we will make every effort to make sure that you are assigned to take this instrument again so you can compare those scores with the ones you earn today. It is the hope of the faculty that we will see a lot of growth and development. We are pleased to offer this feedback to you.

Later in the semester, your personal test scores will be released to your faculty.

Thank you in advance for your effort and concentration on this important test. You may begin.

References

Abdelfattah, F. (2010). The relationship between motivation and achievement in low-stakes examinations. *Social Behavior and Personality*, 38, 159-168.

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bandalos, D. L., & Finney, S. J. (2010). Exploratory and confirmatory factor analysis. In G.R. Hancock & R.O. Mueller, (Eds.), *Quantitative methods in the social and behavioral sciences: A guide for researchers and reviewers*. (pp. 125-155). Florence, NY: Routledge.

Barry, C. L., & Finney, S. J. (2009). Does it matter how data are collected? A comparison of testing conditions and the implications for validity. *Research & Practice in Assessment, 3*, 1–15.

Barry, C. L. ,& Finney, S. J. (2016). Modeling change in effort across a low-stakes testing session: A latent growth curve modeling approach. *Applied Measurement in Education, 29*, 46-64.

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing, 10*, 342-363.

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology, 33*, 609-624.

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research and Practice in Assessment, 8*, 69-82.

De Leeuw, E. D., Mellenbergh, G. J., & Hox, J. J. (1996). The influence of data collection method on structural models: A comparison of a mail, a telephone, and a face-to-face survey. *Sociological Methods & Research, 24*, 443–472.

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller, (Eds.), *Quantitative methods in education and the behavioral sciences: Issues, research, and teaching*. (pp. 439-492), Charlotte, NC: Information Age Publishing.

Finney, S. J., Sundre, D. L., Swain, M., & Williams, L. (2016). The validity of value-added estimates from low-stakes testing contexts: The impact of change in test-taking motivation and test consequences. *Educational Assessment, 21*, 60-87.

Hawthorne, K. A., Bol, L., Pribesh, S., & Suh, Y. (2015). Effects of motivational prompts on motivation, effort, and performance on a low-stakes standardized test. *Research & Practice in Assessment, 10*, 30-39.

Kopp, J. P., Zinn, T. E., Finney, S. J., & Jurich, D. P. (2011). The development and evaluation of the academic entitlement questionnaire. *Measurement and Evaluation in Counseling and Development, 44*, 105-129.

Kornhauser, Z. G. C., Minahan, J., Siedlecki, K. L., & Steedle, J. T. (2014, April). *A strategy for increasing student motivation on low-stakes assessments*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Lau, A. R., Swerdzewski, P. J., Jones, A. T., Anderson, R. D., & Markle, R. E. (2009). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *The Journal of General Education, 58*, 196-217.

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher, 4*, 352-362.

Liu, O. L., Rios, J. A., & Borden, V. (2015). The effects of motivational instruction on college students' performance on low-stakes assessment. *Educational Assessment, 20*, 79-94.

Marsh, H. W., Hau, K., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal, 11*, 320-341.

O'Neil, H. F., Jr., Sugrue, B., & Baker, E. L. (1995). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment, 3*, 135-157.

O'Neil, H. F., Jr., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment, 10*, 185-208.

Rios, J. A., Liu, O. L., & Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research, 161*, 69-82.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. Clogg (Eds.), Latent variables analysis: *Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: Sage.

Sessoms, J. C., & Finney, S. J. (2015). Measuring and modeling change in examinee effort on low-stakes tests across testing occasions. *International Journal of Testing, 15*, 356-388.

Steedle, J. T. (2010, April). *Incentives, motivation, and performance on a low-stakes test of college learning*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

Steedle, J. T. (2014). Motivation filtering on a multi-institution assessment of general college outcomes. *Applied Measurement in Education, 27*, 58-76.

Sundre, D. L., & Thelk, A. D. (2010). Advancing assessment of quantitative and scientific reasoning. *Numeracy, 3*, 2.

Sundre, D. L., Thelk, A. D, & Wigtil, C. (2008). *The Natural World Test, Version 9: A measure of quantitative and scientific reasoning, test manual.*

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2009). Skipping the test: Using empirical evidence to inform policy related to students who avoid taking low-stakes assessments in college. *The Journal of General Education, 58*, 167-195.

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education, 24*, 162-188.

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the student opinion scale to make valid inferences about student performance. *The Journal of General Education, 58*, 129-151.

U. S. Department of Education. (2006). *A test of leadership: Charting the future of American higher education*. Washington, D.C.

Waskiewicz, R. A. (2011). Pharmacy students' test-taking motivation-effort on a low-stakes standardized test. *American Journal of Pharmaceutical Education, 75*(3), 1-8.

Wigfield. A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*, 68-81.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*, 163-183.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1-17.

Wolf. L. F, & Smith, J. K. (1995). The consequences of consequence. Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*, 227-242.

Yu, C., & Muthén, B. (2002, April). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Zilberberg, A., Finney, S. J., Marsh, K. R., & Anderson, R. D. (2014). The role of students' attitudes and test-taking motivation on the validity of college institutional accountability tests: A path analytic model. *International Journal of Testing, 14*, 360-384.

Abstract

The Council for the Accreditation of Educator Preparation (CAEP) requires teacher preparation programs in the United States (US) to document their ability to produce teachers who can effectively promote the learning of a diverse P-12 student population (CAEP, 2013). To meet CAEP accreditation standards, leaders of teacher preparation programs are required to use multiple measures to document and report teacher candidates' learning attainments. Among others, CAEP reviewers accept surveys as an appropriate measure to evaluate program effectiveness. The purpose of this study is to examine considerations related to the use of surveys to effectively measure teacher candidate dispositions towards culturally responsive teaching practices. Study findings identify key factors associated with the use of survey data to guide programmatic and accreditation decisions.

AUTHORS

Jason C. Immekus, Ph.D.
University of Louisville

The Use of Surveys in Teacher Education Programs to Meet Accreditation Standards: Preservice Teachers' Culturally Responsive Beliefs and Practices

The Council for the Accreditation of Educator Preparation (CAEP, 2013) requires teacher preparation programs in the United States to engage in systematic self-study using multiple measures to document their ability to produce teachers who can educate a diverse P-12 student population. This accreditation framework has two important implications for teacher education programs. First, it requires programs to provide teacher candidates rich learning experiences to develop their knowledge and skills to engage in cultural responsive teaching (CRT) practices (Banks & Banks, 1995; Gay, 2002, 2010a, 2010b). For example, clinical exposure affords teacher candidates the opportunity to understand better cultural differences and refine their approaches to teaching diverse students. Second, programs need to select and use measures that yield reliable and valid data to document the extent to which they are able to prepare high- quality teachers. Because surveys are identified as an acceptable accreditation measure, routinely used in higher education to assess student outcomes, and are readily available to operationalize teacher candidates' diversity beliefs (Castro, 2010; Law & Lane, 1987; Song, 2006), it is reasonable to expect that they will serve as an important assessment tool to guide decisions related to meeting CAEP standards. However, effective survey use to measure teacher candidates' CRT beliefs and practices requires consideration of the empirical evidence needed to substantiate the interpretation and use of scores for programmatic and accreditation purposes.

CORRESPONDENCE

Email
jcimme01@louisville.edu

In response, this study examines the use of surveys to measure teacher candidates' CRT beliefs and practices for accreditation purposes. The discussion is framed within the *Standards for Educational and Psychological Testing* (or, *Standards*; American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), which serves "to provide criteria for the development and evaluation of tests and testing practices and to provide guidelines for assessing the validity of interpretations of test scores for the intended

test uses" (p. 1). Specific factors addressed include survey selection, development, and the psychometric properties of scores. For didactic purposes, empirical evidence based on data obtained on teacher candidates' CRT beliefs and practices in a large teacher preparation program, located in the California (CA) Central Valley, is provided. Notably, the considerations addressed in this paper extend to the use of surveys to measure a range of student dispositions in higher education.

Teacher Education Accreditation Standards

…this study examines the use of surveys to measure teacher candidates' CRT beliefs and practices for accreditation purposes.

In 2013, CAEP was established as the agency responsible for the accreditation of teacher education programs in the United States. Within this framework, teacher preparation programs must demonstrate success across five key areas identified as necessary to promote high-quality teacher candidates to meet the learning needs of a diverse P-12 student population. The first three standards are based on the National Research Council (2010) report on factors directly associated with student outcomes, and include:

- Standard 1: Content and Pedagogical Knowledge
- Standard 2: Clinical Partnerships and Practice
- Standard 3: Candidate Quality, Recruitment, and Selection
- Standard 4: Program Impact
- Standard 5: Provider Quality Assurance and Continuous Improvement

Each standard addresses a key component in the training and preparation of teachers to advance the learning of a diverse P-12 student population. The first three standards address the learning outcomes, clinical exposure and experiences, and quality, recruitment, and selection of teacher candidates. Standards 4 and 5 provide a framework for teacher preparation programs to document program impact, as well as quality assurance and continuous improvement efforts.

Standard 1 addresses the content and pedagogical knowledge that teacher candidates are expected to have upon graduation. Among other competencies, teachers must be able to understand how learners develop, use knowledge of students' culture and community differences to create an inclusive learning environment, and utilize effective learning strategies to maximize student learning. The approach to training teacher candidates is critical as the preparation of "culturally responsive teachers with the willingness and abilities to teach in these more diverse school contexts represents, perhaps, the most daunting task facing teacher educators today" (Castro, 2010, p. 198). Therefore, the collection and analysis of different data types with acceptable levels of reliability and validity is paramount for teacher preparation programs to proactively support teacher candidates' abilities to meet the classroom needs of a diverse student population.

Assessment of Preservice Teacher Dispositions

Surveys are widely used among college and university faculty and administrators to measure a range of student outcomes for programmatic and accreditation purposes. Specifically, surveys can be designed or adapted to meet programmatic needs and can be incorporated into electronic assessment systems. Also, the psychometric properties of scores can be evaluated. To facilitate effective survey use in teacher preparation programs, factors related to their selection, development, and the psychometric properties of scores are presented.

Decisions related to survey selection and use should be based on how well the survey aligns to the program outcome it seeks to measure. This requires that program outcomes are clearly defined within a theory of action identifying how they are impacted by program inputs and activities. Gay (2010b) defines CRT as "using the cultural knowledge, prior experiences, frames of reference, and performance styles of ethnically diverse students to make learning encounters more relevant to and effective for them" (p. 31), which provides a framework to identify and evaluate existing measures. There are several measures related to the assessment of teachers' diversity beliefs and practices, including: *Multicultural Teaching Scale* (Wayson, 1993); *Bogardue Social Distance Scale* (Law & Lane, 1987); *Cultural Diversity Awareness*

Inventory (Henry, 1986; Larke, 1990); and the *Culturally Responsive Teaching Self-Efficacy* (CRTSE) and *Culturally Responsive Teaching Outcomes Expectancy Scale* (CRTOE; Siwatu, 2007), respectively. Inspection of the instruments indicates varying perspectives, populations, and approaches used to develop and validate the measures. Therefore, prior to the selection of an existing measure, it is critical to clearly delineate the dispositions that will be operationalized through its use, including the psychometric properties of scores (Immekus, Tracy, Yoo, Maller, French, & Oakes, 2004).

The use of an existing instrument may not be feasible for a number of reasons such as length, cost, or alignment with outcomes. For example, the misalignment of program outcomes and a survey's purpose suggests the need to explore the development of a program-specific measure. As per the *Standards* (Standard 1.1; AERA et al., 2014), the first step in scale development is identifying the instrument's purpose. For instance, the purpose of a programmatic survey could be: "to assess the dispositions of teacher candidates to engage in CRT practices." Subsequent considerations related to instrument development include: teacher candidate characteristics, dimensions of CRT-related practices, administrative constraints (e.g., time), and intended inferences and uses of scores, among others. Characteristics of quality items include that the question and response process is "scripted" so that candidates can answer the question, that the question is equally meaningful across diverse respondents, and that answers can be interpreted similarly across respondents (DeVellis, 2012; Fowler, 2014). Fowler and Cosenza (2008) identify that to answer survey questions accurately, respondents must be able to (a) understand the question, (b) retrieve the information to answer the question, (c) answer appropriately, and (d) answer accurately. The item writing process should engage a range of program stakeholders (e.g., program coordinators) to ensure that obtained results can be used for program decision making. Evidence-based strategies, such as focus groups, cognitive interviews, and review by subject-matter experts, can be used to support the development of a quality instrument (Clark & Watson, 1995; Fowler, 2014).

Investigating the psychometric properties of obtained scores also provide teacher preparation programs a basis to understand the quality of the data. Reporting the psychometric properties of scores used for accreditation purposes is also a CAEP requirement. The *Standards* (AERA et al., 2014) provide a valuable resource to guide decision makers on the types of evidence that can be used to judge the quality of survey scores. Empirical studies indicate that the development of measures of CRT beliefs and practices that yield psychometrically sound scores is an ongoing area of focus (e.g., Siwatu, 2007; Yang & Montgomery, 2011). Consequently, it cannot be assumed that the psychometric properties of scores generalize beyond the context and population in which they have been reported. As such, the types of reliability and validity evidence to gather and report will depend on the intended interpretations and uses of scores.

Thus, despite specific types of test score validity evidence, no one approach is sufficient in and of itself. Instead, documentation of the validity of survey scores within teacher preparation programs is needed throughout all phases of their use.

Reliability deals with test score consistency and addresses the degree to which scores contain unexplained (random) error (Nunnally & Bernstein, 1994; Traub & Rowley, 1992; Thompson, 2003). As such, reliability provides evidence on score precision. There are many approaches to evaluate the reliability of scores derived from surveys (e.g., internal consistency, test-retest), which depend on the sources of errors believed to affect scores (e.g., raters, time). For scores based on an item set, internal consistency reliability is perhaps the most widely used and reported measure of reliability (e.g., Cronbach's coefficient alpha; Streiner, 2003; Thompson & Vacha-Haas, 2000), with estimates above .80 desired (see Henson, 2001). On the other hand, test-retest reliability can be used to examine the stability of survey scores over time. There are multiple measures of reliability, and programs must consider the sources of error (e.g., content, sampling) when selecting and developing the appropriate measure. Therefore, reliability provides one type of evidence on the quality of an instrument's scores, and provides the basis to examine the validity of obtained scores.

Validity is an evolving concept that addresses the extent to which scores represent the measured trait (e.g., diversity beliefs). Kane (2008) states that "[To] validate an interpretation or use of measurements is to evaluate the rationale, or argument, for the claims being made, and this in turn requires a clear statement of the proposed interpretations and uses and a critical evaluation of these interpretations and uses" (p. 17). The *Standards* (AERA et al., 2014) identify five sources of evidence to examine the validity of the interpretations and uses of scores. These include: test content, response processes, internal structure, relations

to other variables, and consequences of testing. These sources of evidence indicate there is no uniform approach to establishing score validity. For example, at the initial stages of survey selection or development, evidence of validity for test content can be gathered using procedures based on the judgments of subject matter experts (e.g., Clark & Watson, 1995; McKenzie, Wood, Kotecki, Clark, & Brey, 1999). Evidence based on internal structure addresses the interrelationships among items or the extent to which items function differently across diverse groups (i.e., differential item functioning). Along these lines, exploratory factor analysis can be used to guide decisions on the retention of items during scale development (Reise, Waller, & Comrey, 2000), whereas confirmatory factor analysis may be used to formally test an instrument's internal structure (Thompson, 2004). Thus, despite specific types of test score validity evidence, no one approach is sufficient in and of itself. Instead, documentation of the validity of survey scores within teacher preparation programs is needed throughout all phases of their use.

Therefore, the collection and analysis of different data types with acceptable levels of reliability and validity is paramount for teacher preparation programs to proactively support teacher candidates' abilities to meet the classroom needs of a diverse student population.

While these considerations can assist programs in selecting and developing surveys that yield psychometrically sound scores, there are noted errors associated with their use. Dillman, Smyth, and Christian (2014) identify four types of errors associated with survey use: coverage, sampling, nonresponse, and measurement. Each type of error is unique and can impede the quality of survey data for accreditation purposes. For example, coverage error occurs if a program restricts data collection activities to only include teacher candidates exposed to specific clinical experiences. In this instance, the sample data may not represent the entire population of teacher candidates in the program. A consequence of this is sampling error, which occurs when the sample data differs from that based on all teacher candidates. Coverage and sampling error can be reduced by ensuring that all teacher candidates have equal likelihood of being included in data collection activities. Nonresponse error is always a concern in survey research and happens when respondents choose not to answer certain questions. Ensuring confidentiality of answers, sending follow-up requests to nonrespondents, and using short surveys can help minimize nonresponse error. Lastly, measurement error deals with the accuracy of the answers. Approaches to reduce measurement error include question clarity and articulating how the data will be used to promote the likelihood that teacher candidates will answer the questions honestly (e.g., minimize social desirability; Fowler, 2014). These sources of error should be considered once surveys have been selected for use to identify strategies to minimize their effect on the interpretation and use of scores.

Used appropriately, surveys offer teacher education programs valuable tools to document the impact of their program to produce quality teachers to meet the learning needs of their P-12 students. Ewell (2013) identifies ten principles related to evaluating the quality of accreditation measures. For example, survey data should be relevant, actionable, of interest to stakeholders, and reliable and valid. Therefore, the quality of accreditation measures is a key indicator of the extent to which they can be used to guide programmatic decisions.

Study Purpose

Situated within these considerations, empirical evidence is reported on the use of surveys to measure preservice teachers' dispositions towards CRT practices within a large teacher education program, located in the culturally rich California Central Valley. Specifically, the program sought to examine the utility of surveys to gather data on teacher candidate diversity beliefs as they progressed in the program. Furthermore, survey results were to be used in conjunction with other evidence (e.g., writing samples) to document teacher candidates' attainment of state-level teacher credentialing requirements. The research questions included:

1) To what extent do the psychometric properties of survey scores support their interpretation and use to measure teacher candidates' CRT practices?

2) What are the dispositions of teacher candidates towards CRT practices?

Methods

A cross-sectional survey design was used to measure dispositional beliefs towards CRT practices among candidates who were at two different phases of their training (completion of their first or last semester in the program) at a large teacher education program in a public university

Used appropriately, surveys offer teacher education programs valuable tools to document the impact of their program to produce quality teachers to meet the learning needs of their P-12 students.

system, located in the California Central Valley. Data was gathered upon completion of the fall semester during the 2010-11 (Year 1) and 2011-12 (Year 2) academic years. Year 1 data was to pilot test the surveys, whereas Year 2 data was to examine the generalizability of Year 1 results. The program sought to identify surveys to gather baseline and periodic data on teacher candidates' dispositions towards CRT practices.

In Year 1, 331 Single Subject credential students (52.6% Female) completed the *Teacher Disposition Index* (TDI; Schulte, Edick, Edwards, & Mackiel, 2004). Of these candidates, 15.11% were final semester completers; the remaining were first semester completers. The racial/ethnic characteristics of teacher candidates responding to the survey were: 58.9% White; 20.7% Latino; 5.4% Asian. The majority of candidates held Bachelor's degrees (89.9%) and were native English speakers (85.9%).

In addition, a separate sample of teacher candidates (N = 208; 74% female) completed the *Culturally Responsive Teaching Self-Efficacy Scale* (CRTSE; Siwatu, 2007) and *Culturally Responsive Teaching Outcomes Expectancy Scale* (CRTOE; Siwatu, 2007). The sample was evenly split according to the number of candidates who completed their first and last semesters in the program. Of first semester completers, 78.8% held a bachelor's degree, 73.1% were native English speakers, and 47.1% were pursuing a Single Subjects credential, compared to 52.9% seeking a Multiple Subjects credential. Also, 46.2% were White, 26% Latino/a, 14.4% Asian, 10.6% reporting two or more races, and 1.9% were African American. Of the last semester completers, 50% were female, 85.6% native English speakers, and all were pursuing a Single Subjects credential. The majority (78.8%) had a bachelor's degree, and race/ethnicity included: 58.7% white, 21.2% Latino/a, 12.5% two or more races, 3.8% Asian, and 29% African American, respectively.

Year 2 data was obtained on the CRTSE and CRTOE among 268 candidates (67.5% female) who were all first semester completers. Program enrollment type included: 45.9% Single Subject, 53.7% Multiple Subjects, and 0.4% missing. Native English speakers comprised 80.6% of the sample, and 50% were white, 29.6% Latino/a, 10.8% two or more races, 7.8% Asian, 1.1% Native Hawaiian or Other (0.4% missing).

Instrumentation

The TDI (Schulte et al., 2004) is a 45-item self-report survey designed to measure preservice teachers' diversity beliefs (e.g., respect cultures of all students), and includes two sub-scale scores: Student-Centered (SC; 25 items) and Professionalism, Curriculum-Centered (PCC; 20 items). Schulte et al. (2004) reported acceptable internal consistency reliability across scores (> .84), with factor analytic results supporting the scale's two-factor structure. For this study, Cronbach's coefficient alpha exceeded .90 across subscale scores.

Siwatu's (2007) CRTSE survey was used to measure preservice teachers' self-efficacy to engage in culturally responsive teaching practices. It includes 40 items that require respondents to provide their answers on a Likert scale (1 = *Strongly Disagree* to 5 = *Strongly* Agree). Siwatu reported that exploratory factor analytic results supported a one-factor model. For this study, Cronbach's coefficient alpha exceeded .95 across Year 1 and 2 data.

Siwatu's (2007) CRTOE survey was used to operationalize preservice teachers' belief in their outcome expectancy beliefs to produce positive outcomes for diverse students. It includes 26 items asking respondents to indicate their ability to positively impact educationally relevant outcomes on a Likert scale (1 = *Strongly Disagree* to 5 = *Strongly Agree*). Siwatu reported that exploratory factor analytic results supported a one-factor model and acceptable internal consistency. For this study, Cronbach's coefficient alpha exceeded .94 across Year 1 and 2 data.

Data Analysis

Descriptive and inferential statistics were used to describe the characteristics of the teacher candidates and examine sub-group differences (e.g., gender, language) on obtained scale scores. Pearson Product Moment correlations were used to examine the relationship

among scores. Inferential statistics included the use of a t-test to examine average score differences across gender, language, and semester completers (i.e., first vs. last semester). Effect sizes were used to characterize the magnitude of the difference between scores (Cohen, 1988).

Results

TDI.

Initial inspection of the data included examining item response frequencies and an item analysis. The item response frequencies indicated that teacher candidates rated themselves at the higher end of the response continuum, with less than 3% responding using the lowest two score categories (i.e., *Strongly Disagree*, *Disagree*). Regardless of semester completed, median item values were a 4 or 5, indicating the high dispositional beliefs among teacher candidates.

Table 1 reports descriptive statistics on the TDI across first and last semester completers for Year 1. As shown, regardless of semester completer, candidates reported high ratings across SC and PCC subscales in excess of 4.50, indicating a high disposition towards diversity beliefs. Correlations indicated that the two subscales were strongly correlated across first semester (r = .92) and final semester completers (r = .85), indicating scores were nearly indistinguishable across samples. Based on a TDI composite score, no score differences were reported across gender, language, or semester completers (ps > .05).

Table 1

Year 1 Descriptive Statistics for the Teacher Disposition Index across First (N = 281) and Last (N = 50)[A] Semester Completers

Scale Score	Mean	*SD*	Minimum	Maximum
SC	4.55 (4.55)	.46 (.32)	1.00 (3.60)	5.00 (5.00)
PCC	4.63 (4.67)	.48 (.31)	1.00 (3.95)	5.00 (5.00)
Total Scale Score	4.58 (4.60)	.46 (.31)	1.00 (3.75)	5.00 (5.00)

[A] Values in parenthesis. *SD* = Standard Deviation. SC = Student-Centered. PCC = Professionalism, Curriculum-Centered.

CRTSE and CRTOE.

A preliminary item analysis indicated that there was a restriction of range across the CRTSE and CRTOE item responses. Specifically, for any given item, less than 6% of respondents selected the lowest two response categories (i.e., *Strongly Disagree*, *Disagree*). Furthermore, all but one item (Item 14) on the CRTSE reported a median score of 5; only Items 4 and 8 of the CRTOE had median scores of 4 compared to 5 for the other 24 items.

While surveys will invariably serve as an important accreditation measure, there are a range of key considerations that teacher preparation programs need to address to substantiate their selection and use.

Table 2 reports Year 1 CRTSE and CRTOE scores across first and last semester completers. As shown, CRTSE and CRTOE average scores were nearly identical, as well as scores across those in the program for one semester compared to those completing their last semester in the program. The scores were also highly correlated, r = .88, across semester completers. Of note, first semester completers reported a slightly higher CRTSE score than last semester completers, although not statistically significant (p > .05).

Inferential statistics were used to examine the presence of statistical score differences across the teacher candidate sub-groups of gender and language, including phase in the program. Females (n = 128) were found to have statistically higher average CRTSE scores (M = 4.74, SD = .28) than males (n = 75; M = 4.58, SD = .45), $t(201)$ = -3.11, p < .01, with a small reported effect size (ES = .35). No score differences were found across language or semester completers (ps > .05).

Table 2

Year 1 Descriptive Statistics for the Culturally Responsive Teaching Self-Efficacy Scale (CRTSE) and *Culturally Responsive Teaching Outcomes Expectancy Scale (CRTOE) among First and Last[a] Semester Completers*

Scale Score	Mean	*SD*	Minimum	Maximum
CRTSE	4.70 (4.66)	.36 (.36)	3.24 (3.59)	5.00 (5.00)
CRTOE	4.66 (4.66)	.35 (.36)	3.50 (3.54)	5.00 (5.00)

N = 208. *SD* = Standard Deviation.
[a] Values in parenthesis.

These results suggest targeted research is needed on the dispositions of incoming teacher candidates regarding their CRT beliefs and practices to guide the selection or development of a more appropriate instrument.

Year 2 data was used to examine across-year trends in teacher candidate CRTSE and CRTOE scores. Similar to Year 1 findings, less than 4% of the candidates selected the lowest two response categories (i.e., *Strongly Disagree*, *Disagree*) for any given item. In most cases, responses were restricted to response options 3 (*Unsure*) to 5 (*Strongly Agree*). As reported, data was only collected on first semester program completers who were either in the Single or Multiple Subjects credential programs.

Table 3 reports descriptive statistics on the CRTSE and CRTOE across program area candidates. As shown, regardless of credential type, candidates reported high scores across measures with slightly higher CRTSE scores. Strong, positive correlations were reported between CRTSE and CRTOE scores for Single Subject (r = .83) and Multiple Subject (r = .75) candidates.

Statistical comparisons were made across program type, gender, and language. Multiple Subject candidates' scores were statistically significantly higher on both the CRTSE (t[218] = -5.83, p < .01) and CRTOE (t[221] = -5.53, p < .01). Effect sizes for the CRTSE (*ES* = .75) and CRTOE (ES = .69) were moderate, respectively. No score differences were found across gender (ps > .05). Among language groups, non-native English-speaking teacher candidates reported a higher average CRTOE score (M = 4.73, SD = .32) than native English-speaking candidates (M = 4.58, SD = .40), t(91) = -2.98, p < .01, with a small effect size (*ES* = .38).

Table 3

Year 2 Descriptive Statistics for the Culturally Responsive Teaching Self-Efficacy Scale and *Culturally Responsive Teaching Outcomes Expectancy Scale[a] across Program Types*

Program Type	*N*	Mean	*SD*	Minimum	Maximum
Single Subject	119 (123)	4.49 (4.47)	.42 (.42)	3.33 (3.58)	5.00 (5.00)
Multiple Subjects	142 (144)	4.76 (4.72)	.32 (.31)	3.16 (3.73)	5.00 (5.00)

[a] Values in parenthesis. *N* = Sample size. *SD* = Standard Deviation.

Discussions and Recommendations

Initiatives to improve teacher effectiveness across P-12 education have resulted in a dramatic shift in how teacher preparation programs are held accountable for the quality of their graduates. This is reflected in the recent adoption and implementation of the CAEP standards for the accreditation of teacher education programs in the United States. One hallmark of this accreditation model is the requirement of teacher preparation programs to engage in self-study practices to collect and analyze data based on multiple measures to document their capacity to prepare teachers that can effectively promote the learning of an increasingly diverse P-12 student population. A critical component is the use of data that meets high-quality standards to yield information that is substantive and meaningful to guide a range of program activities. While surveys will invariably serve as an important

accreditation measure, there are a range of key considerations that teacher preparation programs need to address to substantiate their selection and use.

Toward this end, key issues associated with employing surveys as an accreditation measure were presented within the context of their use to measure teacher candidate dispositions towards CRT beliefs and practices, aligned with CAEP Standard 1 that addresses the ability of teachers to recognize and value student diversity. Beyond accreditation purposes, promoting teacher candidates' CRT practices is critical in light of the noted demographic differences between teachers (predominantly white females) and their students (e.g., Banks & Banks, 1995; Castro, 2010; Gay, 2010a, 2010b). Within teacher education, surveys can provide a convenient and effective approach to investigate program features that most effectively promote candidate outcomes—notwithstanding the attention and consideration that must be taken to ensure that the survey's purpose and program outcomes are well aligned. Therefore, the selection and use of surveys as measures of candidate quality and program impact should only be made after determining the inferences and uses of obtained scores. Such decisions can be supported by professional standards (e.g., *Standards*; AERA et al., 2014), and there are many user-friendly resources to guide program stakeholders in scale selection and development (e.g., Clark & Watson, 1995; DeVellis, 2012; Fowler, 2014; Hinkin, 1995, 1998).

These findings raise a pertinent question related to the development and use of standardized surveys across teacher preparation programs that differ in terms of geography and in their recruitment and selection of culturally diverse students.

This study sought to examine the utility of existing surveys to measure teacher candidate diversity beliefs in a large teacher education program. Survey data was to be used as documentation of teacher candidates' attainment of California's teacher credential requirements. Conceptualization of CRT practices led to the selection of the existing measures of three surveys to be considered for programmatic use. Whereas previous research supported the instruments' psychometric properties, candidate responses in this study resulted in the limited utility of the data. That is, candidate scores had a severe restriction of range at the high end of the continuum, with very few of the respondents selecting the lowest two response categories for any item (regardless of the instrument). Consequently, this limited the use of procedures to pursue specific test score validity studies (e.g., factor analysis). Furthermore, first semester completers reported scores comparable to candidates preparing to exit the program. These results suggest targeted research is needed on the dispositions of incoming teacher candidates regarding their CRT beliefs and practices to guide the selection or development of a more appropriate instrument. For example, the context of this study was in a teacher education program in a regional university located in the culturally rich California Central Valley. As represented in the study's sample, more than 20% of the candidates identified as Latino/a. Also, non-native English speakers reported higher CRTSE scores than native English speakers, suggesting a heightened sense of self-efficacy to engage in culturally responsive practices. Indeed, such findings are noteworthy, and provide areas for future research beyond the sample in which the study data was based. Additional research is underway to investigate the extent to which candidates' exposure to cultural diversity prior to program enrollment may explain these findings. These findings raise a pertinent question related to the development and use of standardized surveys across teacher preparation programs that differ in terms of geography and in their recruitment and selection of culturally diverse students.

Indeed, while surveys can offer teacher preparation programs an efficient and effective approach to gathering program and accreditation data, there are important considerations related to their use…

The findings of this study have direct implications for teacher education programs seeking to use surveys. First, teacher preparation programs are encouraged to evaluate and select surveys that are aligned with their program objectives and then to conduct studies to judge the quality of their scores. Whereas existing scales afford programs access to empirical evidence on their development and validation, this information may not generalize to the context or population in which they may be used (Immekus et al., 2004). As such, the selection and use of surveys for program purposes should be recognized as a process that takes time. In this study, two years of data were gathered on existing instruments to understand their utility. Another issue is the use of multiple measures to document evidence of the preparation of quality teachers. A challenge associated with the use of different electronic assessment systems is that they may not facilitate institutions' ability to merge diverse data types to conduct studies in a timely manner. Such factors identify areas of continued research and consideration in the use of surveys as accreditation measures.

While CAEP standards are noteworthy in their effort to encourage teacher education programs to use rigorous data to improve teacher quality, there are clear challenges to this endeavor. First, the selection and use of quality measures requires time to determine their adequacy based on the principles outlined by Ewell (2013). Second, the vague nature of accreditation standards (e.g., content knowledge) requires teacher education programs to articulate these outcomes (e.g., multicultural education) within the context of their program. This may be especially challenging when the literature is inconclusive regarding how certain outcomes are defined and measured, or which types of clinical experiences are most effective for promoting quality teachers. Third, teacher education programs may use more than one electronic assessment system to collect and organize candidate data (e.g., dispositions, grades). In this instance, there are specific logistics (e.g., student identifiers) that must be identified and addressed to integrate data from different electronic assessment systems. Fourth, when existing surveys are unavailable or their scores lack acceptable psychometric properties, programs will need to determine the appropriateness of creating an institution-specific measure. Such an endeavor may span multiple semesters to gather enough data to evaluate the instrument's quality. By no means an exhaustive list, these are some of the readily apparent issues related to the effective use of surveys as accreditation measures.

Indeed, while surveys can offer teacher preparation programs an efficient and effective approach to gathering program and accreditation data, there are important considerations related to their use—they are beneficial due to their administrative convenience, ability to be integrated into electronic assessment systems, and potential to evaluate the psychometric properties of obtained scores. Notwithstanding these strengths, programs should adhere to professional guidelines and practices regarding their selection and use to ensure that they yield substantive and meaningful information. This is critical in light of the need for continued research on the strategies teacher education programs can use to most effectively promote preservice teachers' diversity beliefs (Castro, 2010; Song, 2006). As such, surveys hold much promise to strengthen teacher education training but require thoughtful consideration in their selection and use.

References

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

Banks, J. B., & Banks, C.A.M. (Eds.). (1995). *Handbook of research on multicultural education*. New York: Macmillan.

Castro, A. J., (2010). Themes in the research on preservice teachers' views of cultural diversity:Implications for researching millennial preservice teachers. *Educational Researcher, 39*(3), 198–210.

Clark, L. A. & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*, 309–319.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Council for the Accreditation of Educator Preparation (2013). *CAEP Accreditation Standards*.

DeVellis, R. (2012). *Scale development: Theory and applications* (3rd ed.). Thousand Oaks, CA: Sage.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). Internet, mail, and mixed-mode surveys: The tailored design method (4th ed.). Hoboken, NJ: Wiley.

Ewell, P. (2013). *Principles for measures used in the CAEP accreditation process*.

Fowler, F. J. (2014). *Survey research methods* (5th ed.). Thousand Oaks, CA: Sage.

Fowler, F. J., & Cosenza, C. (2008). Writing effective questions. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman, (Eds.), *International handbook of survey methodology* (pp. 136–160). New York: Taylor & Francis.

Gay, G. (2002). Preparing for culturally responsive teaching. *Journal of Teacher Education, 53*(2), 106–116.

Gay, G. (2010a). Acting on beliefs in teacher education for cultural diversity. *Journal of Teacher Education, 61*, 143–152.

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

Henry, G. (1986). *Cultural Diversity Awareness Inventory*. Hampton, VA: Hampton University Mainstreaming Outreach Project. (ERIC Document Reproduction Service No. ED 282 657)

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*, 177–189.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management, 21*(5), 967–988.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1*, 104–121.

Immekus, J. C., Tracy, S., Yoo, J. E., Maller, S. J., French, B. F., & Oakes, W. C. (2004). Developing self-report instruments to measure ABET EC 2000 Criterion 3 professional outcomes. *Proceedings of American Society of Engineering Education, USA*, 3230.

Kane, M. T. (2008). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.; pp. 17–64). Westport, CT: Praeger.

Law, S. G., & Lane, D. S. (1987). Multicultural acceptance by teacher education students: A survey of attitudes toward 32 ethnic and national groups and a comparison with 60 years of data. *Journal of Instructional Psychology, 14*(1), 3–9.

Larke, P. J. (1990). Cultural diversity awareness inventory: Assessing the sensitivity of preservice teachers. *Action in Teacher Education, 12*(3), 23–30.

McKenzie, J. F., Wood, M. L., Kotecki, J. E., Clark, J. K., & Brey, R. A. (1999). Establishing content validity: Using qualitative and quantitative steps. *American Journal of Health Behavior, 23*, 311–318.

National Research Council (2010). *Preparing teachers: Building evidence for sound policy*. Committee on the Study of Teacher Preparation Programs in the United States, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). New York, NY: McGraw-Hill.

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12*, 287–297.

Schulte, L.E., Edick, N., Edwards, S., and Mackiel, D. (2004). The development and validation of the *Teacher Dispositions Index. Essays in Education, 12*.

Siwatu, K. O. (2007). Preservice teachers' culturally responsive teaching self-efficacy and outcome expectancy beliefs. *Teaching and Teacher Education, 23*, 1086–1101.

Song, K. M. (2006). Urban teachers' beliefs on teaching, learning, and students: A pilot study in the United States of America. *Education and Urban Society, 38*(4), 481–499.

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*, 99–103.

Thompson, B. (2003).Understanding reliability and coefficient alpha, really. In B. Thompson (Ed.), *Score Reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: Sage.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*, 174–195.

Traub, R. E., & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement: Issues & Practice, 10*, 37–45.

Wayson, W. W. (1993). *Multicultural Teaching Scale*, Synergetic Development Inc. Ohio: Plain City.

Yang, Y., & Montgomery, D. (2011). Exploratory and confirmatory factor analyses of the Multicultural Teaching Scale. *Journal of Psychoeducational Assessment, 29*, 261–272.

Abstract

The Mapmark standard setting method was adapted to a higher education setting in which faculty leaders were highly involved. Eighteen university faculty members participated in a day-long standard setting for a general education communications test. In Round 1, faculty set initial cut-scores for each of four student learning objectives. In Rounds 2 and 3, participants used a Mapmark item map to consider information from four student learning objectives at one glance and to integrate this information into a single cut-score. Participants and faculty leaders reported that the process was intuitive, and there was support for a defensible cut-score from the majority of participants and faculty leaders. Practical suggestions and implications are discussed.

AUTHORS

S. Jeanne Horst, Ph.D.
James Madison University

Christine E. DeMars, Ph.D.
James Madison University

CORRESPONDENCE

Email
hortstsj@jmu.edu

Higher Education Faculty Engagement in a Modified Mapmark Standard Setting

In higher education, setting a standard on an assessment can assist faculty and administrators to distinguish between students who are or are not meeting learning objectives. Standard-setting is the process of selecting cut-scores on a test that will separate examinees' scores into achievement categories (Cizek, 2001; Cizek, Bunch, & Koons, 2004). This facilitates the interpretation of scores in a criterion-referenced fashion, because each category is accompanied by a description of what examinees in that category typically know or can do. For example, on certification exams, the cut-score may be used to indicate whether an examinee has at least adequate knowledge or skills to perform in a job or profession.

Standard-setting has long played a role in primary and secondary education, from the minimal competency or graduation tests common in the 1970s-1990s, to the many statewide tests created in response to the No Child Left Behind (NCLB) legislation, and now to tests under development for the Common Core standards (e.g., Borque & Hambleton, 1993; Tong, Patterson, Swerdzewski, & Shyer, 2014). Even when cut-scores are not used for purposes of passing a test, proficiency categories help students and instructors understand what a score means (AERA, APA, & NCME, 2014, Chapter 5). Although less common, standard-setting is also helpful in higher education. Although higher education scores are typically reported as percent-correct, depending on the difficulty and content-coverage of a test, the percent-correct score may have different meanings. For example, on a test designed to measure a wide range of difficulty spanning four years of education in a major, first-year students scoring 60% may have exceeded the expectations faculty set based on the first-year curriculum. However, if the test only covers foundational concepts students should know before entering the program, this same 60% is likely below the faculty's standard. Proficiency

categories help clarify what a score of 60% means in each of these contexts. In this paper, we will describe a standard-setting workshop for university faculty to set a cut-score on a required communications test. We will discuss the ways the procedure was adapted to meet the needs of the faculty and highlight unique features of the higher-education context.

Standard Setting Procedures

For example, in the current study education professionals requested that participants examine separate ordered item booklets for each of four objectives, rather than one comprehensive ordered item booklet. For this reason, the Mapmark standard setting procedure offered an appealing alternative.

Many methods have been developed for setting standards. Common to most are (a) the development of performance standards (i.e., qualitative descriptions of performance levels, or what students should know and be able to do at the particular level) and (b) the setting of cut-scores (i.e., the score at which an examinee is said to have met the standard; Kane, 1998, 2001). In this study, following the development of performance standards by faculty experts, we used a modification of the Mapmark method, which is closely related to the bookmark method. Mapmark has been used at the national level for setting standards related to the National Assessment of Educational Progress (ACT, Inc., 2007). For purposes of contrast, it is important to briefly introduce one of the most commonly used standard setting methods, the Angoff standard setting method.

Angoff Standard Setting Procedure. Although there are several variants, the Angoff standard setting procedure typically requires standard setting participants (i.e., experts or judges) to conceptualize a "hypothetical minimally acceptable person" (Cizek et al., 2004, p. 40). During the standard setting, participants view test items and make judgments about whether they believe the hypothetical examinee could correctly answer each item. Often participants indicate the proportion of minimally acceptable students who would correctly answer each item. Alternatively, in one common variant of the Angoff procedure, participants respond yes (1) or no (0) regarding whether the hypothetical examinee could correctly answer each item (Impara & Plake, 1997). The cut-score is determined from the average across the items and participants. For example, if the average rating across items and participants is .58, then the cut-score would be 58% correct (Cizek et al., 2004). Other common modifications of the Angoff procedure include multiple rounds (typically two or three) of judgements. Between rounds, workshop leaders facilitate discussions about differences in cut-score judgements. Before the final round of judgements, participants generally receive feedback about their own and others' cut-scores, as well as information about student performance relative to the cut-score, termed *impact* because this information can be used to assess the impact of the cut-score on students.

Inherent within the Angoff method is the assumption that participants are able to adequately conceptualize the knowledge, skills, and abilities of the hypothetical minimally-acceptable examinee, and are able to predict how well that examinee would be able to perform on each item (Impara & Plake, 1998). Moreover, as may be expected, participants *do not* always accurately conceptualize the abilities of the minimally-acceptable examinee (Impara & Plake, 1997, 1998). The bookmark standard setting method attempted to simplify the cognitive task required of Angoff participants by providing booklets of items ordered by empirical difficulty.

Bookmark Standard Setting Procedure. The bookmark standard setting procedure was developed for purposes of minimizing the cognitive tasks and number of judgments required of standard setting participants (Mitzel, Lewis, Patz, & Green, 2001). The central feature of the bookmark method is the ordered item booklet, which consists of test items presented in order of item difficulty. Additionally, participants are provided an item map, which is a table that summarizes the item location information (Mitzel et al., 2001). Standard setting participants place a bookmark at the page at which a minimally-competent examinee would have *mastered* the items prior to the bookmark and would have *not mastered* the items following the bookmark. To "master" an item refers to the point at which the *just-competent* examinee would answer the item correctly, roughly 67% of the time (70-75% with guessing).[1]

Bookmark standard settings typically involve three rounds, similar to many Angoff standard settings. Following orientation, participants review each item in small groups. Participants attempt to identify the knowledge, skills, and abilities required of each item, and the features of each item that make it more difficult than previous items (Mitzel et al., 2001). Following Round 1, participants individually place bookmarks. During Round 2, small group

participants discuss the group's bookmarks in light of the characteristics of the items that fall within the group's range, as well as what students should know at the various proficiency levels. Based on small group discussions, participants again place a bookmark. Following Round 2, the median for each small group and the total group is presented, along with impact data (the percentage of students who would have achieved each performance level). Round 3 involves a discussion among the entire group of participants, following which participants again place individual bookmarks; the final cut-score is the median of these bookmarks. The final cut-score and impact data are presented.

One benefit of the bookmark method over other methods is that item difficulties have been empirically computed, allowing panelists to focus on the *content* of the items (Shulz & Mitzel, 2011). However, one quandary is how to manage the ordered item booklets when test developers desire close attention to items by objectives or domains. For example, in the current study education professionals requested that participants examine separate ordered item booklets for each of four objectives, rather than one comprehensive ordered item booklet. For this reason, the Mapmark standard setting procedure offered an appealing alternative.

Mapmark Standard Setting Procedure. The Mapmark method enhances the bookmark standard setting procedure by assigning the item map a central role in the process (Schulz & Mitzel, 2011). However, unlike the item map provided in the bookmark method, which is simply a list of empirical information about each item in the item booklet, the item map in the Mapmark method presents the information visuo-spatially. By providing spatial information for panelists to judge the distance between the difficulty of the items (see Figure 1), the Mapmark method offers "holistic feedback" on the entire test (Schulz & Mitzel, 2011, p. 168). Round 1 bookmarks are placed in ordered item booklets, as in the bookmark method, but in successive rounds the bookmarks are placed on the item map. Sometimes there are large score gaps between items in the item booklet. In the bookmark procedure, participants must choose a specific item for the cut-score, but in the Mapmark procedure participants can choose to place the cut-score anywhere on the scale, even at scores to which no item difficulties are mapped. As seen in Figure 1, in one glance, panelists are able to focus on the spread of difficulty across domains or objectives. This particular feature of the Mapmark standard-setting procedure was of interest to the current study, in which we were interested in simultaneously presenting information on four separate communication learning objectives.

Context for the Current Study

At a mid-sized public university in the Mid-Atlantic region all students are required to take a basic communications course that covers four learning objectives: (a) Construct messages consistent with the diversity of communication purpose, audience, context, and ethics; (b) Respond to messages consistent with the diversity of communication purpose, audience, context, and ethics; (c) Explain the fundamental processes that significantly influence communication; (d) Utilize information literacy skills expected of ethical communicators. The course is part of the General Education program, which is divided into five components called Clusters. The communications course is part of Cluster 1: Foundations, which includes critical thinking, writing, communication, and information literacy. The current Cluster 1 coordinator is also a Speech Communications professor and the former course director.

The proficiency classifications are used specifically for assessment purposes, to help faculty to judge whether curricular/ instructional changes are needed, and for external accountability reporting.

All basic communication students take a common 100-item course-embedded final exam, which includes 25 items mapped to each of the four learning objectives. The exam is administered in a proctored computer lab. There are approximately 70-80 sections of the course each semester, with 4,000-4,500 students per year. Each instructor can choose the specific

1 Selecting the appropriate response probability (RP) value can be controversial and can influence the order of items in the ordered item booklet. The RP plays a role in determining the location of items when an item response theory model other than Rasch is employed, and influences the description of the standard setting procedure to workshop participants. Participants seem able to adjust the bookmark to partly but not fully compensate for changes in the RP (National Academies of Sciences, 2005, Ch. 5). Traditionally, the bookmark procedure included .67 RP (Mitzel et al., 2001); however, other response probabilities have been investigated (Karantonis & Sireci, 2006). For example, a practitioner may choose to select .50 RP, in which to "master" an item the just-competent examinee would answer the item correctly roughly 50% of the time. However, it is argued that because .67 is above .50, it is more consistent with arguing that a just-competent examinee has mastered an item than .50 RP (Karantonis & Sireci, 2006).

Proportion Correct at Scale Score					Items near Scale Score, by page #			
Scale Score	Obj. 1	Obj. 2	Obj. 3	Obj. 4	Obj. 1	Obj. 2	Obj. 3	Obj. 4
≤200	48%	53%	48%	65%	1,2,3,4	1,2,3,4,5,6	1	1,2,3,4,5,6,7,8,9
210	49%	54%	49%	65%	.	.	.	.
220	50%	55%	50%	66%	.	.	.	.
230	52%	56%	51%	67%	5	7	.	.
240	53%	57%	52%	68%	.	.	.	.
250	54%	58%	53%	69%	6	8	.	10
260	55%	59%	54%	69%	.	9	.	11
270	56%	60%	55%	70%	.	.	2	12
280	58%	61%	56%	71%	.	.	.	13,14,15
290	59%	62%	57%	71%	.	.	.	.
300	60%	63%	58%	72%	7	10	.	.
310	61%	64%	59%	73%	.	.	3	.
320	63%	65%	60%	73%	.	11	4,5	.
330	64%	66%	61%	74%	.	.	6	16
340	65%	67%	63%	75%	.	.	7	.
350	67%	68%	64%	75%	.	12	8	.
360	68%	69%	65%	76%	8,9,10	.	9,10	.
370	69%	70%	66%	76%	.	.	.	.
380	70%	71%	67%	77%	11,12	13,14	.	.
390	72%	72%	68%	77%	13	15,16	.	.
400	73%	73%	69%	78%	14	17	.	.
410	74%	74%	70%	78%	15,16	.	11	.
420	75%	75%	71%	79%	17	.	.	17
430	76%	76%	72%	79%	.	.	12,13	.
440	77%	76%	73%	80%	18	18	14,15	.
450	78%	77%	74%	80%	.	.	16	.

Figure 1. *Mapmark item map. The complete item map extended to a score of 800.*

learning activities, but all sections use the same textbook and cover the same objectives. The basic course director, a Speech Communications professor, facilitates consistency across the many instructors, and oversaw faculty who wrote the test items.

Rationale for Standard Setting

The cut-score corresponding to the proficiency standard is not used to determine whether students pass or fail the course. The continuous score on the final exam, not the dichotomous proficiency classification, is incorporated as one part of each student's course grade, along with presentations and other in-class assignments. The proficiency classifications are used specifically for assessment purposes, to help faculty to judge whether curricular/instructional changes are needed, and for external accountability reporting.

Thus, faculty leaders wanted the procedure modified to separate the learning objectives, yet yield a single cut-score. Therefore, the Mapmark standard setting procedure was chosen as a viable standard setting method.

Because the context of the current study differs from the traditional K-12 standard setting, it is important to carefully define three roles: faculty leaders, workshop leaders, and participants. For the purpose of program evaluation and accountability reporting, the course director and Cluster 1 coordinator requested the assistance of faculty at the university's assessment office in setting a proficiency standard on the final exam. The term *faculty leaders* will be used to refer to the Cluster Coordinator and the course director. The term *workshop leaders* will be used to refer to the personnel who did the psychometric work, prepared materials, and helped facilitate the workshop. These labels are arbitrary because both groups are faculty and both groups participated in leading the workshop, but short labels are needed for description. The workshop leaders played the role typically fulfilled by testing company staff when setting standards for statewide K-12 tests or certification tests. The faculty leaders, on the other hand, have no direct parallel. Because of the scale of statewide K-12 tests, curriculum leaders are generally not personally known by the standard setting participants the way the faculty leaders were in this context. Because the standard-setting took place in a single university, and most of the participants taught General Education courses, the faculty leaders were viewed as colleagues. Finally, the term *participants* will be used to refer to faculty members who served as content experts throughout the workshop.

The faculty leaders had participated in other standard setting workshops at the university, using a modified bookmark procedure (for example, DeMars, Sundre, & Wise, 2002). In previous standard settings, all items were included in one ordered item booklet,

regardless of the objective to which the item aligned. Faculty leaders felt it was confusing combining items mapped to four separate learning objectives into one ordered item booklet, making it difficult to discuss what each item was measuring and why it might be harder than the item before it. Thus, faculty leaders wanted the procedure modified to separate the learning objectives, yet yield a single cut-score. Therefore, the Mapmark standard setting procedure was chosen as a viable standard setting method.

Purpose

The purpose of this study was to illustrate a variation on the Mapmark standard setting procedure designed to highlight multiple learning objectives assessed by one test. A secondary purpose was to illustrate standard setting within a higher-education context. The context of the current study was unique, relative to traditional standard settings, given that faculty leaders were highly involved in the process. Moreover, faculty leaders felt strongly that items should be considered by learning objective. Also unique to the higher education setting was the length of the standard setting workshop. Rather than several days, the current standard setting was conducted in one day, to minimize demands on faculty time. The current study summarizes this adaptation of the Mapmark standard setting procedure in a higher education context.

Method

Modification of the Mapmark Procedure

Because the faculty were dissatisfied with previous standard-settings, in which items with different learning objectives were interspersed within the ordered item booklets, workshop leaders and faculty leaders discussed ways of separating the task by learning objective. Faculty agreed that they wanted a single cut-score on the test as a whole, not four separate standards. One option was for the participants to set four separate standards using the bookmark procedure and combine them at the end of the workshop. One concern was that, with the shorter ordered-item-booklets resulting from dividing the items by objective, there would be many score gaps within each booklet. Imagine that the just-proficient student envisioned by a particular participant has the skills corresponding to a scaled score of 328. The standard-setting participant does not know the value 328, but can, hypothetically, envision skills and knowledge at this level. But there may not be any items close to this level; perhaps there is a large gap between an item located at 280 and another located at 362. Another problem is that if each standard were set in isolation, the standards for each learning objective would likely end up at very different points on the proficiency continuum and the mean would not represent the desired proficiencies well. This might be hidden from the participants by using a method that sets the standard on the percent-correct metric, such as the Angoff method; participants would assume that objectives where they set the percent-correct cut-score high were easier than objectives where they set the percent-correct cut-score low. Of course, hiding the incongruity from the participants does not make it go away. Setting the cut-score on the percent-correct metric could also be problematic when the test forms changed; the cut-score might correspond to a different percent-correct when the new form was equated. The faculty leaders also were comfortable with the Bookmark method and did not want to replace it.

The context of the current study was unique, relative to traditional standard settings, given that faculty leaders were highly involved in the process.

The Mapmark procedure provided a way to incorporate the learning objectives because it displays the expected percent-correct by objective or content area. Although participants using the Mapmark procedure generally use a single item booklet in Round 1, with items from different objectives interspersed, we modified Round 1 to include four separate ordered-item booklets, and participants set four separate bookmarks. During Round 2, participants received feedback on where their bookmarks for the different objectives fell relative to the scale scores and to bookmarks set by others. Each participant then set a single bookmark directly on the overall scale in successive rounds.

Preparation

Performance-level descriptors were written by the faculty leaders. Detailed descriptors are important for helping standard setting participants envision students who just meet the criteria for each performance level (Kane, 1998, p. 134; 2001, p. 59). Without written descriptors, participants will implicitly define the performance levels for themselves, which

can lead to wide variation in interpreting the performance levels. Perie (2008) provided practical suggestions on developing performance-level descriptions.

The faculty wanted a single cut-score on the test as a whole, which implies a unidimensional scoring model. It seems somewhat cognitively inconsistent to emphasize the uniqueness of the learning objectives yet score the test using a unidimensional model. To make sure that a single score on the test was meaningful, we ran a multidimensional 3-parameter-logistic (3PL) confirmatory factor model. The latent (disattenuated) correlations among the first three factors were estimated to be 1. The factor tapping Objective 4 was estimated to be correlated .83 with the other three factors. The RMSEA[2] was .01 for both the 4-dimensional model and the 1-dimensional model, suggesting both models fit acceptably. Thus, it seemed reasonable to follow the faculty desire for a single score (unidimensional model).

Materials were prepared for Round 1 following the usual bookmark procedures. Based on the unidimensional 3PL[3] calibration, the item location was calculated. The item location was defined as the ability at which an examinee would have a 2/3 probability of correct response, not counting correct guessing (Lewis, Green, Mitzel, Baum, & Patz, 1998), also referred to as .67 RP. Recognizing that the choice of RP can be controversial, we chose the .67 RP (i.e., 2/3 probability of correct response), which aligns with the original description of the Bookmark method (Mitzel et al., 2001) and is consistent with findings suggesting that participants more easily conceptualize .67 as examinee mastery of items (Karantonis & Sireci, 2006). The item locations were linearly transformed to the scaled scores used in score reporting, ranging from 200 to 800. In a typical bookmark or Mapmark standard setting, items are ordered by location. In this modification, items were separated by objective and ordered within each objective. Each item was printed on a separate page, along with information about the proportion of students in the upper and lower thirds of the score distribution who chose each option.

In a typical bookmark or Mapmark standard setting, items are ordered by location. In this modification, items were separated by objective and ordered within each objective.

For Rounds 2 and 3, an item map was assembled showing scaled scores in increments of ten. At each scaled score, the expected proportion correct was displayed for each objective, followed by the page numbers of items that mapped to that scaled score after rounding. An example of the first part of the scale range is shown in Figure 1—the complete scale range was printed out on a single sheet of 11 by 17 paper for each participant. Figure 1 illustrates, for example, that students who scored 300 would have average raw scores of 60% on Objective 1, and 63%, 58%, and 72% on Objectives 2, 3, and 4, respectively. About 2/3 of the students at score 300 have mastered the 7th item in Objective 1, plus a few more would get it right by guessing. Higher proportions of the students at score 300 have mastered the first 6 items in Objective 1, and lower proportions have mastered the harder items ordered after item 7. This item map helps the participants put the separate learning objectives back into the context of the test as a whole. Score gaps are also evident in Figure 1. For example, using the Mapmark item map, participants could place the cut-score at a score of 370, which would not be possible using the bookmark procedure because there are no items located near that score.

Workshop Activities

The 18 participants completed the test prior to the workshop so that the entire standard-setting could take place in a single day. After providing an overview of the day's activities, faculty leaders provided a context for the test's use within the general education program and discussed the development of the test. Workshop leaders discussed item writing, the way in which distractors contribute to an item's difficulty, and introduced activities that would occur throughout the day. Prior to the beginning of the session, the entire group discussed performance level descriptors. Given that the task was to set one cut-score, there were two performance-level descriptors written by the faculty leaders.The Developing student was described as:

2 The RMSEA used here is based on marginalizing estimations from full-information methods down to bivariate moments so that fit indices developed for limited-information methods can be estimated (Maydeu-Olivares & Joe, 2014).

3 More precisely, a bifactor model was used with secondary factors to account for dependence between some pairs of items, with the parameter estimates projected onto the primary factor (Kahraman & Thompson, 2011) to produce a unidimensional scale.

> "Students below the proficient category have not demonstrated the skills necessary to be able to recognize the fundamental processes that significantly influence communication. Students at this level have not demonstrated an ability to ethically construct and respond to messages consistent with the diversity of communication purposes, audiences, and contexts. They may be unable to utilize information literacy skills or to construct and/or respond to messages effectively or ethically. This category denotes partial but insufficient mastery."

The Proficient student was described as:

> "Students meeting this standard are able to explain the fundamental processes that significantly influence communication. Students at this level demonstrate an ability to ethically construct and respond to messages consistent with the diversity of communication purposes, audiences, and contexts. Students who achieve this standard are able to utilize information literacy skills expected of ethical communicators. Although further development is expected, students achieving this level or higher have the knowledge necessary to communicate effectively within the [institution] academic community."

Participants were each provided a notebook that included: agenda, background context, performance level descriptions, and the four ordered item booklets, one per learning objective.

Round One. Participants divided into four table groups. Starting with Objective 1, participants followed the usual bookmark procedure for Round 1. Each group discussed what each item measured and why it was more difficult than the previous item. A separate item map was provided for each objective, so participants could see when the locations of adjacent items were similar and not spend time trying to discern nonexistent or small differences in item difficulty. Table leaders encouraged full participation from everyone at their table. After all tables discussed Objective 1 items, the bookmark process was explained. After placing bookmarks for Objective 1, table groups discussed Objective 2 items, placed bookmarks, and proceeded through the remaining Objectives. Workshop leaders calculated scale scores for (a) each participant's four bookmarks, (b) mean ratings across each participant's four bookmarks, (c) each table's median rating, and (d) each table's lowest and highest average bookmark scale score.

Round Two. After a lunch break, table group results and Mapmark item maps were explained. Once participants demonstrated that they understood the Mapmark item map, they were encouraged to flag the place on the scale next to the bookmark they selected for each objective and their table's lowest and highest bookmark. Table leaders directed participants' attention to the items between the table's lowest and highest bookmarks. Participants discussed the knowledge, skills, and abilities they believed the items were measuring and whether just-Proficient students should be expected to master the content represented by the items. After small group discussion, each participant individually placed *one* Round 2 bookmark, indicating the scale score appropriate for a just-Proficient student. Workshop leaders tabulated each participant's response and provided the median small group scale score.

Round Three. Following a break, the entire group resumed for discussion. Workshop leaders presented a summary of each table's median scale score as well as impact data for the entire group's median cut-score. The impact data were based upon data from the previous year's administration of the test, and indicated the percent of examinees scoring at or above Proficient level based on the Round 2 median bookmark. Following discussion, participants were instructed to place their third and final bookmarks. Workshop leaders tabulated the data and presented the final cut-score and impact data. Faculty leaders and workshop leaders led discussion with participants about their satisfaction with the final cut-score and the day's experiences. Participants completed an evaluation prior to leaving.

Because we were adapting the Mapmark method to our context, it was crucial to evaluate the appropriateness of the method.

Results and Validation

Scores are on a scale from 200-800, with a mean of 500 and standard deviation of 100. The recommended cut-score following Round 3 was 480. Impact data computed from the previous year's administration of the test indicated that with this cut-score, 58% of students

taking the basic communications course would have been classified as Proficient. Although 58% Proficiency may seem stringent, faculty leaders and participants expressed strong support for the score.

Other distinctive features of the process were that faculty leaders were highly involved throughout the standard setting, and that, with the exception that we required participants to complete the test prior to the standard setting, the standard setting occurred in only one day.

In the context of describing the choice of an appropriate standard setting method, Kane (1998) noted, "it is not easy to evaluate how well a standard-setting procedure is working" (p. 130). That is, standard settings are fraught with subjectivity and arbitrary decisions (Kane, 1994). Cut-scores are representative of the value judgments of the standard setting participants (AERA, APA, & NCME, 2014, p. 101). At best, evaluation of the effectiveness of a standard setting method involves consideration of the appropriateness within the context and purpose for the standard setting, and evaluation of the validity of inferences drawn from application of the standard. The current context was an educational setting, in which faculty leaders were highly involved in the process and would use the information for improvement of their program, rather than high-stakes student pass/no-pass decisions. As such, we felt the strongest evidence would be to adopt the validity argument approach to evaluating the appropriateness of the adaptation of the Mapmark to the current context. At least three forms of validity evidence are recommended: procedural, internal consistency, and external evidence (Kane, 1994, 2001).

Procedural Evidence

Kane (1998) stressed that cut-scores are set, not estimated. There is no "true" cut-score. Thus, procedural evidence often plays a large role in validating the cut-score (Kane, 1994, 1998, 2001). Because we were adapting the Mapmark method to our context, it was crucial to evaluate the appropriateness of the method. We attempted to stay true to the traditional bookmark and Mapmark procedures, as well as general best practices described within the standard setting literature (e.g., Hambleton, 2001; Plake, 2008). And, although anecdotal, standard setting participants seemed to easily grasp the concept of the Mapmark item map. For purposes of assessing procedural validity, we administered a paper-pencil questionnaire immediately following the standard setting.

Table 1

Responses to Satisfaction Questions (Procedural Validity)

Satisfaction with final cut-scores
100.0% (18) Satisfied/Very Satisfied
0.0% (0) Neither satisfied nor dissatisfied
0.0% (0) Dissatisfied/Very Dissatisfied
Satisfaction with standards-referenced nature of cut-scores
94.5% (17) Satisfied/Very Satisfied
5.6% (1) Neither satisfied nor dissatisfied
0.0% (0) Dissatisfied/Very Dissatisfied
Satisfaction with consideration of values/opinions
88.9% (16) Satisfied/Very Satisfied
5.6% (1) Neither satisfied nor dissatisfied
5.6% (1) Dissatisfied
0.0% (0) Very Dissatisfied
Defending the cut-point
83.3% (15) would defend the cut-point
16.7% (3) would not defend the cut-point
Round 3 bookmark changes
38.9% (7) changed bookmark but not as a result of the impact data
38.9% (7) changed bookmark based on the impact data or others' reactions to it
22.2% (4) did not change bookmark
Confidence in Bookmark Procedure for setting valid standards
72.2% (13) Confident/Very Confident
27.8% (5) Neutral
0.0% (0) Not Confident /Not at all Confident
Agreement with item ordering in booklets
88.9% (16) Generally/Somewhat Agreed
5.6% (1) Neither Agreed nor Disagreed
5.6% (1) Somewhat Disagreed
0.0% (0) Generally Disagreed

Table 2

Response to Workshop Setting (Procedural Validity)
Organization of workshop
94.4% (17) Very Organized/Organized
0.0% (0) Neither Organized nor Disorganized
5.6% (1) Disorganized
0.0% (0) Very Disorganized
Quality of general Bookmark training
44.4% (8) Excellent
38.9% (7) Good
16.7% (3) Fair
0.0% (0) Poor
0.0% (0) Fail
Quality of workshop leaders
50.0% (9) Excellent
38.9% (7) Good
11.1% (2) Fair
0.0% (0) Poor
0.0% (0) Fail
Overall Value of Workshop as Professional Development Experience
66.7% (12) Excellent
27.8% (5) Good
5.6% (1) Fair
0.0% (0) Poor
0.0% (0) Fail
Value of Interacting with peers in the group
83.3% (15) Excellent
11.1% (2) Good
5.6% (1) Fair
0.0% (0) Poor
0.0% (0) Fail
Value of constructing better classroom tests (1 missing)
52.9% (9) Excellent
29.4% (5) Good
17.6% (3) Fair
0.0% (0) Poor
0.0% (0) Fail
Value of targeting instruction (2 missing)
37.5% (6) Excellent
31.3% (5) Good
31.3% (5) Fair
0.0% (0) Poor
0.0% (0) Fail

Overall, program participants were satisfied with the workshop. Only one person expressed dissatisfaction with the extent to which participant opinions were considered and valued. A majority of participants (83.3%) stated that they would defend the final cut-score. Three participants who indicated they would not defend the cut-score also indicated that they had changed their cut-score in Round 3; two reported changing their cut-scores as a result of something other than the impact data, and one participant reported changing his/her cut-score based on impact data. All who elected not to change their bookmark at Round 3 were among those who indicated that they would defend the cut-score if asked. See Tables 1 and 2 for a summary of responses.

Most participants expressed confidence in the validity of the standard setting process. A large majority of participants generally or somewhat agreed with the item ordering found in the booklets. Although not indicated in the numeric data, one respondent reported feeling that Round 1 evaluation of Objective 1 was a training session, resulting in less valid Objective 1 bookmarks than subsequent objectives' bookmarks. However, individual objective bookmarks were simply used as a starting point for the exam's cut-score and no single objective in Round 1 should have a large influence on the final cut-score. Although most participants expressed satisfaction with the process, confidence in the cut-score, and appreciation for the workshop as a form of professional development, it

was clear that the process was not perfect. Three participants stated they would not defend the cut-score. However, the proportion of participants who defended the cut-score was similar to the proportion who indicated the same during a prior year's bookmark standard setting. Moreover, confidence in the order of the ordered item booklets increased in the current Mapmark standard setting (88.9%) relative to the prior year's bookmark standard setting (68%), in which items were combined across objectives. However, given that the prior year's standard setting involved a different test and different participants, comparisons across years were made cautiously. The majority of participants indicated they would use the information gained through the standard setting process to enhance their pedagogy.

Both faculty leaders had also been involved in the prior year's bookmark standard setting and noted that the Mapmark was an improvement over the bookmark method. In particular, the Mapmark allowed participants to consider each of the four Objectives individually, while at the same time setting one cut-score. In their estimation, the Mapmark method was a success. However, it was also important to evaluate other forms of evidence.

Internal-Consistency Evidence

In addition to procedural evidence, evaluation of internal consistency of participants' ratings is also a component of a sound validity argument (Kane, 1994; 2001). Figure 2 portrays individual participants' cut-scores across the three rounds. Note that variation in Group 1 participants' cut-scores decreased across the three rounds (e.g., cut-scores converged). In contrast, the remaining groups' ratings converged at Round 2, following table discussions. However, following Round 3 discussions, some participants changed their cut-scores. Although there was still variation in participants' final cut-scores, the least variability was following the Round 3 discussion.

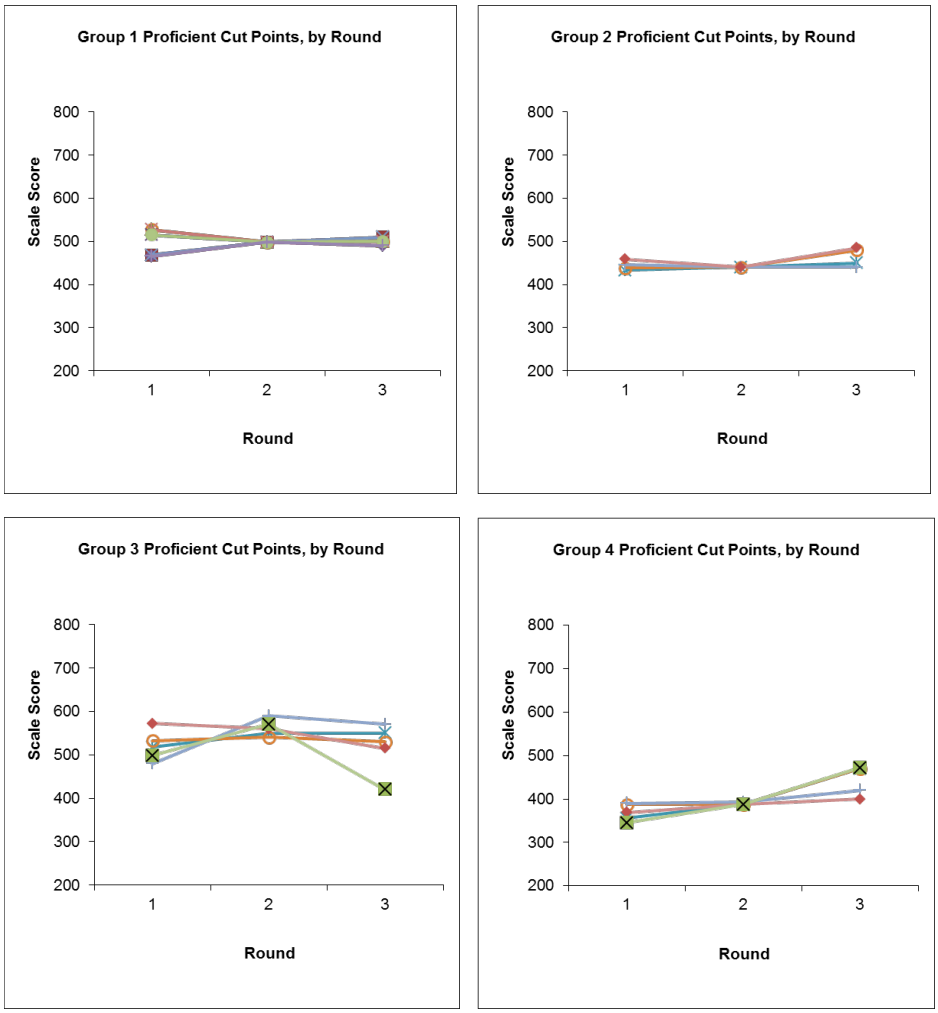


Figure 2: *Variance in Bookmarks*

Hypothetically, the standard error of the cut-score would be the standard deviation of the final cut-score set at each of an infinite number of workshops, with different participants at each workshop. Because the table groups are relatively independent at Round 2, data from Round 2 are typically used in the estimate of the standard error (Lewis et al., 1998; Mitzel et al., 2001), calculated as:

$$SE = \sqrt{\frac{s^2}{N}\left[1+(n-1)r\right]},$$

where s^2 is the variance of the cut-scores, N is the total number of participants, n is the number of groups, and r is the intraclass correlation, which adjusts the SE to take into account dependency within group. If the median Round 3 cut-scores from different workshops were more alike than the median Round 2 cut-scores from different tables within the same workshop, this would be an overestimate of the standard error (or conversely, it would be an underestimate if groups were more alike within workshops than between workshops). In Figure 2, it is evident that the variance within groups is much smaller than the variance across groups; the intraclass correlation is 0.92. Thus, the estimated standard error of the cut-score was 32.9; it would have been 16.6 simply using the unadjusted standard error of the mean.

External Evidence

Finally, the collection of external validity evidence contributes to a strong validity argument (Kane, 1994, 2001). One form of external validity evidence for the current test is whether the cut-score can aid in identifying groups of students that may need extra support. Anecdotal and empirical evidence (i.e., average percent correct) at the university in which the current study was conducted identified several groups that seem to struggle with passing the test. For the purpose of understanding student performance on the test, examination of Developing/Proficient rates using the cut-score were computed, identifying groups who are still in the Developing category. Analysis of the previous year's data indicated that there was a large group of international students (70.4% Fall 2013; 77.8% Spring 2014) identified as Developing (not yet Proficient). Across both semesters, male students, on average, scored below the cut-score; whereas female students' average was above the cut-score. In sum, the external evidence that was available pointed to meaningful interpretations when applying the cut-score.

In sum, recognizing that further study and direct comparisons with other standard setting methods should be conducted, we cautiously recommend the modified Mapmark process for use by higher education practitioners when evaluation of items by objective or domain is desired.

Discussion and Conclusions

The current study presents an application of the Mapmark standard setting procedure to a higher education setting, during which a standard was set for a test mapping to multiple learning objectives. Other distinctive features of the process were that faculty leaders were highly involved throughout the standard setting, and that, with the exception that we required participants to complete the test prior to the standard setting, the standard setting occurred in only one day. In general, faculty leaders and participants expressed appreciation for the process and most supported the standard that was set. Nonetheless, the process was not perfect and the subjectivity and arbitrariness inherent within any standard setting was evident in the procedural validity feedback from participants.

The cut-score adopted in the current study is used for program assessment purposes. However, there are other reasons that higher education assessment practitioners may want to create a cut-score. For example, unlike the current study, in a previous standard setting we set a cut-score for our university's information literacy assessment test, in which the cut-score is used for pass/fail determinations. Students who do not meet the cut-score are required to repeat the test, until they have mastered the test at a proficient level of competency. Another use for cut-scores within higher education is for university placement. For example, performance on foreign language or mathematics tests frequently determine placement into the appropriate level of language or mathematics course. The procedures described in this study are applicable across these standard-setting contexts.

Future Study and Limitations

As mentioned by Kane (1994), "There is no gold standard. There is not even a silver standard" (p. 448). Comparing cut-score classification with a direct behavioral assessment would provide validity evidence for the performance descriptor and the cut-score. Conducting a standard setting for the communications test using another standard setting method (e.g., Angoff) and comparing results would provide further external validity evidence (Kane, 1994). However, doing so in an applied context where participant time is costly would be prohibitive and outside the mission of practitioners at the university. Continued application of the method and ongoing evaluation of validity evidence for resulting cut-scores is warranted.

Practical Suggestions

In sum, recognizing that further study and direct comparisons with other standard setting methods should be conducted, we cautiously recommend the modified Mapmark process for use by higher education practitioners when evaluation of items by objective or domain is desired. The concept of the holistic item map was easily grasped by workshop participants and the process resulted in a cut-score that was approved by most participants. The following are some practical suggestions that one may want to consider if planning a similar standard setting.

Detailed performance level descriptors should be reviewed at the beginning of the standard setting and be provided for participants to consult throughout the session. Without detailed descriptors, participants may rely on their own personal definitions of competence, resulting in greater variation in cut-scores than desired (Kane, 1998; 2001). Flexibility in the schedule is also recommended. Given that Round 1 involves the careful identification of the knowledge, skills, and abilities required to correctly answer each item, it is important to allow participants enough time to fully complete this step. Allowing some flexibility within the schedule permits organizers the opportunity to lengthen the time allotted to the various rounds, as needed.

Finally, assessment practitioners who conduct standard settings within higher education may want to consider involving faculty leaders throughout the process. The faculty leaders' involvement lent credibility—they were curricular leaders and colleagues to the participants. Faculty leaders provided a perspective that resonated with participants, they supported and defended the assessment process, and they were able to provide an educational perspective to the discussion. Consequently, Round 3 discussions were lively and collegial. Faculty members who teach downstream from the communications course counted the experience as professional development and expressed appreciation for knowing what to expect of students' communication knowledge, skills, and abilities. Nonetheless, when including faculty leaders it is important to consider whether unwanted influence on ratings is introduced through their participation. In the current study, we felt that course director participation enhanced the process and outweighed any potential sources of bias. However, there may be situations in which this is not the case, and assessment practitioners would want to take sole responsibility for the workshop.

Conclusion

The current study offers support for an adaptation of the Mapmark standard setting method to a higher educational setting. Inclusion of the Mapmark item map in Rounds 2 and 3 of the bookmark standard setting allowed participants to consider information from all four objectives at one glance. Participants and faculty leaders reported that the process was intuitive, and there was support for a defensible cut-score from the majority of participants and the faculty leaders.

References

ACT, Inc. (2007). *Developing achievement levels on the 2006 national assessment of educational progress in grade twelve economics: Progress report.* Iowa City, IA: Author.

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

Borque, M. L., & Hambleton, R. K. (1993). Setting performance standards on the national assessment of educational progress. *Measurement & Evaluation in Counseling & Development, 26*, 41–47.

Cizek, G. J. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 3–51). Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31–31.

DeMars, C. E., Sundre, D. L, & Wise, S. L. (2002). Standard setting: A systematic approach to interpreting student learning. *Journal of General Education, 51*, 1-20.

Hambleton, R.K. (2001) Setting performance standards on educational assessments and criteria for evaluating the process. In G.J. Cizek (Ed.), *Setting performance standards* (pp. 89–116). Mahwah, NJ: Lawrence Erlbaum.

Impara, J.C., & Plake, B.S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 353–366.

Impara, J.C., & Plake, B.S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35*, 69–81.

Kahraman, N., & Thompson, T. (2011). Relating unidimensional IRT parameters to a multidimensional response space: A review of two alternative projection IRT models for scoring subscales. *Journal of Educational Measurement, 48*, 581–601.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 62*, 425–461.

Kane, M. (1998). Choosing between examinee-centered and test-centered standard-setting methods. *Educational Assessment, 5*, 129–145.

Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G.J. Cizek (Ed.), *Setting performance standards* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice, 25*(1), 4–12.

Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (April, 1998). *The bookmark standard setting procedure: Methodology and recent implementations*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49*, 305–328.

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The bookmark procedure: Psychological perspectives. In G.J. Cizek (Ed.), *Setting performance standards* (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum.

National Academies of Sciences (2005). *Measuring literacy: Performance levels for adults, interim report* Available from http://www.nap.edu/catalog/11267/measuring-literacy-performance-levels-for-adults

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice, 27*(4), 15–29.

Plake, B.S. (2008). Standard setters: Stand up and take a stand! *Educational Measurement: Issues and Practice, 27*(1), 3–9.

Schulz, E.M., & Mitzel, H.C. (2011). A Mapmark method of standard setting as implemented for the National Assessment Governing Board. *Journal of Applied Measurement, 12*, 165–193.

Tong, Y., Patterson, B., Swerdzewski, P., & Shyer, C. (2014, April). *Standard setting for a Common Core aligned assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.

Book Review

Pedigree:How Elite Students Get Elite Jobs. Laura A. Rivera. Princeton, NJ: Princeton University Press, 2015. 392 pp. ISBN: 9780691155623. Hardcover, $35.

REVIEWED BY:
Jamie Alea, Ph.D.
San Jose University

Laura A. Rivera's *Pedigree: How Elite Students Get Elite Jobs* gives us a glimpse into the world of top-tier investment banks, management consulting firms, and law firms and the ways in which their hiring practices reproduce economic privilege. Rivera examines how elite reproduction occurs in labor markets after students graduate from institutions of higher education. Specifically, she investigates how access to the highest-paying elite jobs is closely tied to socio-economic class, cultural resources, and social connections. Her research challenges the belief held by most Americans that individual merit and hard work, not socioeconomic class and social connections, are the most important factors for social mobility.

Rivera describes the transmission of economic privilege from one generation to the next as occurring through the educational system. She states that although access to higher education has expanded for all, children from the most affluent families lead university attendance—even more so at elite colleges and universities. This phenomenon persists into graduate education where over half of the students at top-tier business and law schools come from families from the top 10 percent of incomes nationally (Fisher, 2012). Rivera argues that higher education has *become* the mechanism of social stratification and inequality.

Rivera outlines other factors, such as social connections and cultural resources, which increase candidates' chances of securing the highest-paying elite jobs. Rivera found that parents' social connections can give their children an advantage by providing access to important opportunities such as social networks, insider tips, and coveted internships. Similarly, Rivera's study revealed that a shared world view, class-specific tastes, values, and interaction styles give candidates from affluent backgrounds the upper hand and frequently influence hiring decisions.

Rivera studied hiring methods in elite professional service firms through a combination of interviews and participant observation. Her qualitative study focuses on the phenomenon of elite reproduction through access to elite jobs and high incomes. Rivera conducted 120 semi-structured interviews with professionals involved in undergraduate and graduate hiring in top-tier consulting, banking and law firms. She refers to the firms as *elite professional services* (EPS) firms. These professionals were made up of hiring partners, managing directors, and mid-level employees charged with conducting interviews and screening resumes. The interviews concentrated on the evaluators' approach to assessing candidates—which qualities they sought and details of their interactions with the candidates at each stage of the hiring process. Rivera also presented fictitious candidates' resumes to the hiring professionals in order to tease out what evaluative criteria the participants used and how they interpreted the resumes.

Rivera's study also includes participant observation of recruitment activities. She sought to better understand how EPS firms seek new employees and the ways in which they communicate the qualities that they look for in candidates. Rivera also presented herself as a graduate student who was interested in employment opportunities with these firms. She attended recruitment presentations and diversity job fairs over a period of six months. In addition, she conducted fieldwork as a recruiting intern within the recruiting department of one of the EPS firms over a nine-month period. Rivera shadowed the recruitment team through full-time and summer intern recruitment in order to observe candidate selection directly and to note patterns outside of the evaluators' consciousness.

This book takes the reader through the entire hiring process, from recruitment to the final offer of employment or rejection. Rivera systematically outlines the ways in which candidates are evaluated and how hiring decisions are made, revealing a "golden pipeline" of prestigious universities from which these EPS firms recruit and select their employees. In most cases, only students who attended a university with strong ties to these firms or had social connections to individuals within the firms were selected for interviews—prerequisites which are strongly associated with parental socioeconomic status.

In a time when the cost of the college degree is rising and social mobility is stagnant, Rivera challenges us to honestly look at the ways in which we address social and economic inequality through our current system of education.

Pedigree exposes the ways in which elite reproduction occurs in the hiring process through its close ties to elite universities. Rivera asserts that despite the perception that elite schools have rigorous merit-based admissions concerned only with finding the best and the brightest these schools are socio-economically homogeneous; they have a student body disproportionately made up of students from affluent families. Further, Rivera describes a hiring process which exclusively recruits from elite universities, thus restricting access by diverse non-affluent applicants. The candidates are almost exclusively from affluent backgrounds and the high-paying positions are systematically offered to students from the same prestigious universities with ties to the EPS firms.

Rivera explains how a small number of candidates from less affluent backgrounds were able to secure employment with these EPS firms through various pathways. Some non-elite applicants had a serendipitous match with

RESEARCH & PRACTICE IN ASSESSMENT

an interviewer who valued non-traditional applicants; others received insider coaching or benefited from cultural osmosis. Some applicants relied on mimicry, caricaturing class difference, or compensatory credentialing. Rivera describes cases where candidates from non-elite backgrounds had the fortune of being interviewed by people who championed candidates that would most likely be rejected by the hiring team. Other non-elite candidates were able to secure employment with the EPS firms through insider coaching from friends, romantic partners, or classmates who had connections to or were themselves inside the firms. They received insider or class-based knowledge about how to navigate the interview process of the EPS firms.

Less affluent candidates sometimes used mimicry in order to emulate the self-presentation and interaction styles of the elite individuals they knew. This strategy helped non-elite candidates to connect with the interviewers from the EPS firms. Rivera presents some instances when cultural osmosis was used by individuals from non-elite backgrounds who had attended elite primary or secondary schools and learned cultural signals and knowledge through immersion in privileged environments. Other tactics included caricaturing difference, in which a candidate exaggerates class difference to convey a rags-to-riches narrative. This strategy tended to evoke admiration from the interviewers but did not always pay off. Last, compensatory credentialing was sometimes used by candidates of non-elite backgrounds to gain legitimacy. This approach involves obtaining a third-party organization's certification of a job candidate's worth. Examples include the United States Military and non-profit organizations tied to elite universities. The non-profit organizations provide job opportunities in investment banking and law firms to under-represented students. These third-party organizations are respected institutions that vouch for the applicants and are given credibility for their level of discipline, rigor, or cultural capital.

Rivera's research would be of interest to anyone working in higher education, as it tests some of the preconceived notions of the impact that higher education has on one's future while focusing on equality—a strongly shared value across colleges and universities. Her work is helpful to assessment professionals because it challenges us to rethink the relationship between social mobility and educational attainment. While educational attainment is thought to be a driver of social mobility, Rivera's research exposes higher education as the pathway for the transmission of privilege for some and a glass ceiling for others. Although Rivera's study focuses on students attending prestigious universities and an elite group of employers, her findings underscore the importance of correctly identifying the necessary skills to enter all ranks of employment for college graduates.

All institutes of higher learning have vested interests in understanding what outcomes employers are looking for—which skill sets, content knowledge, and dispositions are expected in order to successfully enter the labor market. Rivera's work reveals that among elite employers a combination of interpersonal skills and dispositions, such as well-roundedness, social skills, and "polish, "[1] were more important than content knowledge in hiring decisions. As such, assessing the extent to which institutes of higher learning are adequately preparing students for the labor force may be increasingly pertinent to measuring the value of a college degree. This also underscores the importance of involving students in co-curricular experiences, in addition to their academics. Co-curricular programs often focus on outcomes such as leadership and professionalism that could help less-advantaged students prepare for the workforce and obtain higher-paying jobs following graduation.

Rivera's research would be of interest to anyone working in higher education, as it tests some of the preconceived notions of the impact that higher education has on one's future while focusing on equality—a strongly shared value across colleges and universities.

Rivera aptly shines a light on a hidden system of exclusive networks between top-tiered institutes of higher learning and the most prestigious, highest-paying jobs. Her research suggests that most moderate- and low-income students, even those attending elite colleges and universities, still believe that a college degree, hard work, and persistence will result in their social and economic mobility. In fact, social mobility research supports her assertion that the current system serves to transmit privilege rather than equalize social and economic differences (Haverman & Smeeding, 2006). The original goal of increasing access to higher education for moderate- and low-income students as a means to improve social and economic differences and close income gaps has fallen short of expectations (Haverman & Smeeding, 2006). Conversely, Rivera's work suggests that middle- and lower-income students are actually unprepared to compete in a contest in which they are unaware of the rules.

This book is significant because it pulls back the curtain and provides some justification for current skepticism about the viability of achieving the "American Dream" through educational attainment. In a time when the cost of the college degree is rising and social mobility is stagnant, Rivera challenges us to honestly look at the ways in which we address social and economic inequality through our current system of education.

Pedigree thoroughly investigates how elite students get elite jobs, illuminating the reproduction of privilege through the educational system. Rivera concludes that "successfully reducing class inequalities or increasing social mobility requires addressing biases in both" (p. 274). Her research encourages us to think about the ways in which

1 Polish is defined as a job candidate's style of communication and self-presentation.

higher education is closely linked to social and economic stratification and the difficulty in disentangling the two.

References

Fisher, D. (2012, May). Poor students are the real victims of college discrimination. *Forbes*.

Haverman, R., & Smeeding, T. M. (2006). The role of higher education in social mobility. *The Future of Children 16*, 125-150.

Rivera, L.A. (2015). *Pedigree: How elite students get elite jobs*. Princeton, NJ: Princeton University Press.

Book Review

Service-Learning Essentials. Questions, Answers, and Lessons Learned.
Barbara Jacoby. San Francisco: Jossey-Bass, 2015.
322 pp. ISBN-13: 978-1118627945. Paperback, $35.

REVIEWED BY:
Agnieszka Nance, Ph.D.
Tulane University

Service-Learning Essentials is a manuscript authored by one of the most influential scholars in the field of community engagement, Barbara Jacoby. Jacoby's contributions to this scholarly field range from a breadth of research and publications to presentations, speeches, and teaching. As Jeffrey Howard acknowledges in the foreword to the manuscript: "She is one of the icons in the service-learning movement… [and] has had a panoramic and on-the-ground view of our work" (p. xiii). The importance of the volume is underscored by its publisher, Campus Compact, a national organization "dedicated solely to campus-based civic engagement."

Jacoby clearly decided to focus on only one aspect of engagement scholarship, the pedagogy of service-learning, rather than more broadly on what she calls "civic learning" (i.e., active citizenship or public service). She sees the "tremendous potential of service-learning to prepare our students to be active participants in our democracy and our work on behalf of social change" (p. xvii).

Operating from that perspective, Jacoby's goal is a thorough explanation of the tenets of the pedagogy for the purpose of intentional and thoughtful use in the academic context. Her own intentionality comes through in the innovative format of the book. Rather than using narratives, she opted for a Q&A format with added references for further reading.

As the title suggests, the book addresses the most fundamental aspects of the pedagogy of service-learning across a continuum. Consisting of nine chapters, Jacoby's text moves from introducing the pillars of the pedagogy to far more complex themes, such as assessment and the role of service-learning in higher education. As a theorist and practitioner of service-learning, Jacoby thoroughly addresses each of the questions comprising separate and independent chapters.

After providing the reader with theoretical foundations in Chapter One, Jacoby presents best practices of critical reflection in Chapter Two, providing a context for both curricular and co-curricular settings. In Chapters Three and Four, she delves deeper into the curriculum, this time looking at community partnerships and multidisciplinary aspects of the pedagogy.

The role of assessment is presented mainly in Chapter Six. (However, Chapter Nine also includes an important discussion of the significance of research and assessment.) Quoting Furco and Holland, Jacoby acknowledges that service-learning "requires *evidence above passion* [emphasis added]." Her methods reach beyond the traditional approach to include assessment, research, and evaluation as necessary parts of scholarship. To stress the validity and significance of an evidence-based approach, Jacoby argues that assessment of the pedagogy is "essential to secure its future" (p. 254).

Jacoby presents various methods for assessing service-learning. She is careful not to advocate for any specific approach and instead presents considerations for choosing one particular method. What is refreshing about this chapter is the inclusion of the community partners' perspective. It is worth noting that Jacoby includes the question of the value of service-learning for the community–a topic often less emphasized by service-learning scholars–as well its impact on systems and partnerships. At the core of service-learning, community partners play the role of co-educators and should equally benefit from the partnership with the university.

Chapter Six is an overview of the assessment of service-learning. Jacoby does not focus on details; her purpose is a broad introduction to the topic. This approach would benefit an audience less familiar with assessment, practitioners interested in analyzing the impact of their work, and new adapters of the pedagogy.

Jacoby's text also emphasizes the institutionalization and complexities of service-learning. The last two chapters introduce the problematic aspects of incorporating service into the curriculum such as considering issues of diversity, understanding systems of power, students' resistance to the idea of service learning requirements, or recognition for faculty in academia. As a veteran of the field, Jacoby is clearly aware of the shortcomings of the pedagogy, pointing out the need for critical reflection, better engagement with K-12 education, and greater efforts to standardize service-learning pedagogy (for instance, by the Carnegie Foundation).

The strength of this book lies in its usability. It has the potential to benefit beginning scholars as well as seasoned pedagogues, trainers, and graduate students. The structure of each chapter allows the reader to select parts of higher interest and identify additional sources listed in each sub-section. Jacoby takes a balanced approach: She provides a panoramic view of the pedagogy from various standpoints, covering the fundamentals as well as the latest developments and examples. The value of *Service-Learning Essentials* is in its practicality and clarity.

References

Jacoby, B. (2015). Service-learning essentials. *Questions, answers, and lessons learned*. San Francisco: Jossey-Bass.

Notes in Brief

The ever-increasing internationalization of study programs and global mobility of students call for greater transparency of and valid information on the knowledge and skills students acquire over the course of their studies. Several theoretical and methodological challenges arise from the immense diversity of degree courses, study programs, and institutions. A recent review of the literature has revealed a substantial lack of research on assessment practices in higher education, especially on domain-specific and generic competency models, as well as on measurement methods and valid instruments for competency assessment. The German Federal Ministry of Education and Research initiated the national research program Modeling and Measuring Competencies in Higher Education (KoKoHs) to address these challenges. This article describes the assessment practices, aims, and conceptual and methodological framework of KoKoHs and presents the main results of the first funding phase of the program.

AUTHORS

Prof.Olga Zlatkin-Troitschanskaia
Johannes Gutenberg University

Prof. Hans Anand Pant
Humbolt University of Berlin

Dr. Christiane Kuhn
Johannes Gutenberg University

Miriam Toepper
Johannes Gutenberg University

Corinna Lautenbach
Humbolt University of Berlin

Assessment Practices in Higher Education & Results of the German Research Program Modeling and Measuring Competencies in Higher Education (KoKoHs)

CORRESPONDENCE

Email
lstroitschanskaia@uni-mainz.de

Earlier approaches to competency assessment in higher education were limited mostly to prerequisite admissions tests, data on learning opportunities, and subjective measures (Kuhn & Zlatkin-Troitschanskaia, 2011). Recent analyses have revealed that accredited higher-education institutions lack sufficiently reliable and valid instruments to assess students' learning outcomes and that there are tremendous differences in competency evaluation in higher education within departments and institutions, as well as among institutions nationally and internationally. Hence, not surprisingly, the results of many studies indicate that most certificates and grades in higher education are hardly comparable on the national level let alone on an international level (Zlatkin-Troitschaskaia, Shavelson & Kuhn, 2015). Given the increasing internationalization and global mobility of students, it is imperative there is transparency of students' knowledge and skills across various study models and countries.

Valid assessment of competencies in higher education form the basis for transparency and comparability of academic degrees, which are stipulated aims of policy reform programs. Therefore, in 2010, the German Federal Ministry of Education and Research (BMBF) initiated the national Modeling and Measuring Competencies in Higher Education (KoKoHs) research program, which addresses the political and practical challenges of conducting competency assessment in higher education. We present an overview of the main outcomes following the

end of the first five-year funding phase of the KoKoHs program. First, we outline the structure, aims, and theoretical and methodological framework of KoKoHs. Second, we present the accumulated results in the areas of competency modeling, test development, and validation. Third, we report on key activities of the program that will shape the future of competency assessment in higher education in Germany, including the dissemination of results, internationalization of KoKoHs networks, and provision of support for young researchers. We conclude by outlining challenges and perspectives for the second funding phase of KoKoHs.

KoKoHs: Structure and Aims

The KoKoHs program provides systematic, internationally compatible and fundamental research on competency assessment in higher education (Zlatkin-Troitschanskaia, Kuhn, & Toepper, 2014). During the first phase, the program included 70 projects with 220 researchers at more than 50 institutions of higher education in Germany and Austria. Selected during an external review process, each 24 cross-university collaborative project was required to bring together domain experts, teaching methodology experts, and research methodology experts from at least two universities. KoKoHs projects involved more than 50 international experts (from universities, testing institutes, etc.) from 20 countries including the United States, Australia, Japan, and South Korea. The first phase ran from 2011 to 2015. Having received positive external evaluation, the program is continued for another five years (2016 to 2020).

The general purpose of the KoKoHs program is to model and assess systematically domain-specific and generic competencies of students in higher education. KoKoHs projects take into account curricular and job-related requirements, transform theoretical competency models into suitable measuring instruments, and validate test score interpretations. To enable meaningful cooperation and promote cross-project synergies during the first phase, KoKoHs focused on student competencies in one generic cluster (self-regulation and general research competencies) and four domain-specific clusters comprising some of the most popular fields of study in Germany:

- engineering, including electrical engineering and mechanical engineering;
- economics and social sciences, including teacher training in economics and social sciences;
- educational sciences, including psychology; and
- teacher training in science, technology, engineering, and mathematics (STEM subjects).

Conceptual and Methodological Framework

In the KoKoHs projects, competencies were defined as latent cognitive and noncognitive underpinnings of performance (Ewell, 2005; Rychen, 2004). The KoKoHs program adopted Weinert's (2001) definition of competencies as "cognitive abilities and skills that individuals possess or acquire in order to solve certain problems as well as the aligned motivational, volitional and social dispositions and skills to apply the solutions in different situations successfully and responsibly" (p. 27-28). This general definition was specified for competencies acquired in higher education. During the first phase, KoKoHs projects focused predominantly on (latent) cognitive abilities and skills and specified them for their respective fields of study (Alexander, 1997; Alexander, Winters, Dinsmore, & Parkinson, 2011).

Given the increasing internationalization and global mobility of students, it is imperative there is transparency of students' knowledge and skills across various study models and countries.

Models of knowledge and skills were operationalized through measuring instruments and tested in empirical assessments. Validation efforts aimed to establish validity of interpretation and answer the key question: What can we infer from the (cognitive) representations elicited by the assessment of the actual competencies of students? This approach is always challenging: The underlying cognitive abilities and skills—ideally also the corresponding noncognitive (e.g., affective-motivational) aspects—need to be operationalized through representative, practice-oriented, and often domain-specific tasks; assessments need to represent specific situational contexts and be free of potential bias, such as measurement errors or influences of construct-irrelevant test-taking behavior (Kulikowich & Alexander, 1994; 2003).

KoKoHs projects take into account curricular and job-related requirements, transform theoretical competency models into suitable measuring instruments, and validate test score interpretations.

The general assessment framework in KoKoHs was based on the Assessment Triangle by Pellegrino, Chudowsky, and Glaser (2001), which covers three fundamental aspects of assessment: "a model of student *cognition* and learning in the domain, a set of beliefs about the kinds of *observations* that will provide evidence of students' competencies, and an *interpretation* process for making sense of the evidence"(p. 44) (see also Shavelson, 2013; Webb, Shavelson, & Steedle, 2012). These three aspects corresponded with key objectives of KoKoHs:

1. Define the construct to be assessed (*cognition*).
2. Develop and use suitable models and measuring instruments (*observation*).
3. Draw valid inferences from the assessment data (*interpretation*).

The Assessment Triangle provided the cornerstones for an assessment connecting theoretical constructs of students' competencies with empirical evidence; that is, developing estimates based on limited instances of students' knowledge and skills in an argument-based approach of "reasoning from evidence" (Mislevy, 1994). For more specific, practical orientation, KoKoHs project teams adopted the evidence-centered assessment approach and test development concept (Mislevy & Haertel, 2006; Hattie, Jaeger, & Bond, 1999), which includes the following steps:

- *Domain analysis and modeling:* In the assessment of competencies in higher education, initial steps included analyzing and defining the domain and modeling the domain-specific construct to be assessed.
- *Assessment framework:* An assessment framework was defined, which served to operationalize the theoretical model and develop items for the test instruments.
- *Assessment implementation:* The instruments were tested empirically.
- *Assessment delivery:* The test scores were analyzed using various psychometric models. Analyses always included evaluations of fit of the data to the theoretical constructs and to the corresponding data interpretations. The conclusive evaluation of the tests with regard to various validation criteria served as a basis for further decisions (see also Pant, Rupp, Tiffin-Richards, & Köller, 2009).

Validation is of paramount importance in KoKoHs. Validation efforts followed the International Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014; Kane, 2013).

Overview of the Main Results

The following is a synopsis of the results of the KoKoHs program at the end of the first funding phase and before the beginning of the second funding phase. Results are summarized and presented for the three main areas of work in the program: development of competency models; development of test instruments; and validation. Furthermore, we describe the efforts taken to reach the following three strategic aims of the program: to achieve national and global visibility; to ensure internationalization and compatibility of results; and to help young researchers establish a specialized research community.

Competency Models, Assessments, and Validation

The teams of the 24 collaborative projects developed 41 competency models of generic and domain-specific competencies in higher education. Content validity (including curricular validity) was ensured in the KoKoHs projects through analyses of almost 1,000 documents such as module descriptions and study regulations from more than 250 institutions of higher education throughout Germany. Furthermore, analyses of items and tasks from almost 1,500 documents (e.g., exams, exercises, lecture notes) informed the construction of test items as shown in Table 1.

Table 1

Competency Models Developed and Validation

Theoretical competency models	41
Document analyses of	
curricula, regulations, standards	910
exams, exercises, lecture notes	1,350
project and lab reports	48
Validation	
expert interviews	556
cognitive labs	459

In addition to these document analyses, validation measures employed in the KoKoHs projects often included expert interviews and cognitive labs (with N~500 experts and N~500 participants, respectively, across the 24 collaborative projects). Expert interviews provided evidence of content validity; cognitive labs provided evidence of cognitive validity through analyses of fit between cognitive processes in the theoretical models and thought processes observed empirically in think-aloud interviews while participants responded to items.

Table 2

Instruments Developed and Competency Assessments Conducted

Instruments	
paper-pencil tests	63
computer-based tests	36
video-based tests	8
other tests (e.g., critical incidents)	119
Assessment surveys	
institutions	226
participants	49,904

The teams of the 24 collaborative projects also created new assessment instruments based on the competency models developed, and/or they adapted existing international instruments, if available, to meet their needs. Altogether, more than 60 paper-pencil instruments and almost 40 computer-based instruments were developed in the KoKoHs program as depicted in Table 2. During the first five-year funding phase, more than 220 researchers and several hundred student assistants were involved in project work. In addition, lecturers and students from the participating institutions in Germany actively supported the program by supervising or participating in the surveys and assessments during classes. As an incentive, all participating institutions, lecturers, and students received professionally prepared feedback from the aggregated, anonymized data. In turn, this facilitated transfer of research results and findings into higher-education practices.

Additional, more action-oriented approaches were employed in the KoKoHs projects for valid assessment of specific competency facets. For example, video-recorded role plays were used to assess explanatory knowledge of pre-service physics teachers within the domain of teacher education. The KoKoHs reseachers used—in addition to almost 10 newly developed video-based instruments—various other measuring methods, such as critical incidents, for complementary, qualitative, in-depth analyses of competency levels. Moreover, the teams of the KoKoHs collaborative projects used quantitative methods, (e.g., structural equation analysis) to gather evidence of validity aspects such as the internal structure of competency constructs.

The internal structure was differentiated according to both content requirements (e.g., knowledge and skills related to financial plans as part of business competency) and cognitive requirements (e.g., remembering, applying, or evaluating) (Anderson & Krathwohl, 2001).

Overall, the teams of the KoKoHs collaborative projects assessed domain-specific and generic competencies as well as personal and structural influence factors of approximately 50,000 students at more than 220 institutions of higher education in Germany as well as in Austria. Results in the three areas of competency modeling, test development, and validation significantly contributed to the provision of a reliable, valid, and internationally compatible basis for competency assessment in higher education in Germany.

Overall, the teams of the KoKoHs collaborative projects assessed domain-specific and generic competencies as well as personal and structural influence factors of approximately 50,000 students at more than 220 institutions of higher education in Germany as well as in Austria

Project Example

The WiwiKom project, which focuses on modeling and measuring the competencies of students and graduates of business and economics, provides one example of how the conceptual and methodological framework was implemented and how psychometric validity was gathered (Zlatkin-Troitschanskaia, Förster, Brückner, & Happ, 2014). The construct of professional competency in business and economics was defined in a theory-driven competency model based on Kane's interpretative use argument (2013). Empirical evidence gathered in the assessments was described in the validity argument; subsequently, analyses of the data indicate the modeling was adequate (Kane, 2013).

The theory-driven model of competency in business and economics developed in WiwiKom (Zlatkin-Troitschanskaia et al., 2014) differentiated seven domain-specific content dimensions and three levels of cognitive requirements. The content dimensions represented the core curriculum in business and economics, sub-divided into content areas (e.g., microeconomics, finance, etc.). The cognitive dimension specified levels of professional competency defined in terms of the mental processes (e.g., understanding, applying, etc.) necessary to respond appropriately to situational cognitive requirements of increasing complexity. The competency model served as a basis for developing the WiwiKom test instrument and validating it in qualitative and quantitative studies with a focus on the five key validity aspects, while adhering to the Standards for Educational and Psychological Testing (AERA et al., 2014).

For curricular and content validation, the test content was examined during document analyses and was compared to curricula and textbooks from 98 degree courses at 64 institutions of higher education in Germany; it also was evaluated by lecturers of business and economics during expert interviews (N=32) and in online ratings (N=78). For cognitive validation, mental processes of 32 students were examined in cognitive labs, where students were asked to think aloud while responding to test items. For item calibration, test standardization, and establishment of validity of internal test structure, three field surveys were conducted in the WiwiKom project, assessing approximately 10,000 students of business and economics from all years of study at 57 institutions of higher education in Germany. The data was analyzed using methods such as confirmatory factor analysis or IRT modeling to gather evidence on the dimensionality and gradation of the examined competency. In addition, surveys were administered to gather data on multiple personal variables (e.g., gender, prior knowledge, etc.) and institutional variables (e.g., type of institution) for analyses of the relationship between the construct and other variables.

To enhance global visibility and international networking as well as to ensure compatibility with international research and higher education practices, KoKoHs researchers established and maintained cooperation with international experts in different research areas.

Further Activities of the KoKoHs Program

In addition to the specific research goals, there were three strategic aims of the KoKoHs program which would define the long-term impact of the program.

National and Global Visibility

A major aim of KoKoHs project was to achieve national and global visibility through the dissemination of our results. The teams of the KoKoHs collaborative projects were highly productive, primarily publishing articles in high-ranking national and international journals. Moreover, approximately 250 presentations were held at national conferences, and almost 100 presentations were held at high-profile international conferences, such as annual meetings

of the *European Association for Research on Learning and Instruction* (EARLI) and the *American Educational Research Association* (AERA).

In addition to the project teams presenting and publishing the results of individual projects, results related to the entire KoKoHs program also were documented and published by the coordination project. The coordination project not only contributed numerous presentations and posters to national and international conferences with a focus on scientific topics or higher-education practice as shown in Table 3, but also published its own KoKoHs Working Papers series (with seven issues altogether, five of which were in English[1]) as well as seven thematic issues in prestigious national and international journals, some of which were coedited by renowned international cooperation partners. The KoKoHs program is the only national research program in Germany that has published an overview paper and is represented in an international edited volume on all research initiatives worldwide in the area of learning outcomes assessment in higher education.

Table 3

Project Results Disseminated

Publications	
national	134
international	65
Presentations	
national	244
international	89

The KoKoHs program has more than 50 international cooperation partners from 20 countries on four continents.

Internationalization

To enhance global visibility and international networking as well as to ensure compatibility with international research and higher education practices, KoKoHs researchers established and maintained cooperation with international experts in different research areas. International KoKoHs cooperation partners include experts from various universities, research associations, and public and non-profit higher education and research institutions, including testing institutes. The KoKoHs program has more than 50 international cooperation partners from 20 countries on four continents. During the first funding phase, cooperation between KoKoHs project teams and international partners included joint events such as the KoKoHs Affiliated Group Meeting at the 2014 AERA conference in Philadelphia (Kuhn, Toepper, & Zlatkin-Troitschanskaia, 2014), joint publications such as a special issue in the journal Studies in Higher Education (Zlatkin-Troitschanskaia & Shavelson, 2015), and in the Peabody Journal of Education (Zlatkin-Troitschanskaia, Blömeke, & Pant, 2015) as well as joint supervision of doctoral and post-doctoral projects of KoKoHs researchers.

Supporting Young Researchers

With approximately 70 doctoral projects and almost 20 post-doctoral projects conducted by KoKoHs researchers, a major focus of the program was to systematically support young researchers in building up a scientific community within empirical higher-education research in Germany. Providing the young researchers with the necessary guidance would enable them to close existing gaps in research on competency assessment in higher education. To this end, the KoKoHs coordination project organized for all young researchers a variety of systematic training opportunities and events throughout the course of the program including methodology workshops, mentoring, and networking events such as international colloquia. Workshops were held at regular intervals over the course of the program on various

1 KoKoHs Working Papers can be downloaded at http://www.kompetenzen-im-hochschulsektor.de/ 617_DEU_HTML.php. See the KoKoHs homepage in English for more in-depth information, including details about KoKoHs events.

topics related to research methodology, including a general introduction to methods of social research, item and test development, scaling and test theory, validation, and longitudinal data analysis. Networking and mentoring events such as the International Colloquium for Young Researchers in November 2013 and the international Autumn Academy in October 2014 were organized for outstanding young researchers whose submissions had been selected by international experts (Toepper, Zlatkin-Troitschanskaia, Kuhn, Schmidt, & Brückner, 2014).

These events presented young researchers with excellent opportunities for networking internationally, presenting their work to the international scientific community, and receiving feedback from renowned international experts. Further opportunities for international networking and exchange open to all researchers in the program included the international kick-off and closing conferences as part of the cooperation between individual collaborative projects and international partners.

Conclusions and Future Perspectives

KoKoHs program goals for the second funding phase include maintaining and expanding the networks established thus far, while continuing to support and draw on the expertise of the research community solidified in Germany during the first phase.

During the first funding phase, KoKoHs addressed theoretical, methodological, and empirical challenges including: systematically designing or adapting tests; considering framework conditions such as time, method, and format; analyzing data with complex psychometric methods; confirming psychometric quality criteria; and undertaking comprehensive validation. The models of competency structures and levels, the assessment designs, and the measuring instruments developed and tested so far provide a solid basis for future in-depth longitudinal multilevel analyses in random field experimental studies in higher education.

To date, few studies in higher education have employed complex methodological designs, such as longitudinal modeling, multilevel modeling, or (quasi-)experimental designs. Hence, findings on the trajectory of competencies over the course of studies are still scarce. With regard to instruments, there remains a lack of innovative measurement methods such as adaptive computer-based testing. Many challenges need to be addressed in order to overcome the unsatisfactory state of having to rely on less direct indicators (i.e., grades, degrees, and students' self-evaluations) and to complement these existing indicators with more direct assessments that allow valid conclusions to be drawn about student competencies (Zlatkin-Troitschaskaia, Shavelson, & Kuhn, 2015).

In 2015, the BMBF launched the second phase of the KoKoHs research program. The remaining theoretical, methodological, and empirical challenges will be addressed in this funding phase. These challenges include systematically designing or adapting tests under time, method, and format constraints, analyzing data with complex methods, confirming psychometric quality criteria, and undertaking comprehensive validation (AERA et al., 2014). Due to specific challenges in higher education—reliability issues related to complex models constrained by limited testing time, panel mortality in longitudinal studies, and testing based on students' performance—more complex and innovative methods of analysis need to be considered. These methods include longitudinal and multilevel analyses in random field-experimental studies, adaptive computer-based testing, and suitable psychometric techniques. KoKoHs program goals for the second funding phase include maintaining and expanding the networks established thus far, while continuing to support and draw on the expertise of the research community solidified in Germany during the first phase. More systematic international collaboration and exchange of best practices from this field (and related areas such as competency assessment in the school sector) are needed.

References

Alexander, P. A. (1997). Mapping the multidimensional nature of domain learning: The interplay of cognitive, motivational, and strategic forces. In M. L. Maehr & P. R. Pintrich (Eds.), *Advances in motivation and achievement* (Vol. 10, pp. 213¬–250). Greenwich, CT: JAI Press.

Alexander, P. A., Winters, F., Dinsmore, D. L., & Parkinson, M. (2011). The role of domain knowledge in self-regulated learning. In B. Zimmerman & D. Schunk (Eds.), *Handbook of self-regulation of learning and performance*. New York: Routledge.

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

Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing. *A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Ewell, P. T. (2005). Can assessment serve accountability? It depends on the question. In J. C. Burke & Associates (Eds.), *Achieving accountability in higher education* (pp. 1–24). San Francisco, CA: Jossey-Bass.

Hattie, J., Jaeger, R. M., & Bond, L. (1999). Persistent methodological questions in educational testing. *Review of Educational Research, 24*, 393-446.

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

Kuhn, C., & Zlatkin-Troitschanskaia, O. (2011). Assessment of competencies among university students and graduates – Analyzing the state of research and perspectives (Business Education Working Paper No. 59). Mainz: Johannes Gutenberg University.

Kuhn, C., Toepper, M., & Zlatkin-Troitschanskaia, O. (2014). *Current international state and future perspectives on competence assessment in higher education – Report from the KoKoHs Affiliated Group Meeting at the AERA Conference on April 4, 2014 in Philadelphia (USA)* (KoKoHs Working Papers No. 6). Berlin & Mainz: Humboldt University & Johannes Gutenberg University.

Kulikowich, J. M., & Alexander, P. A. (2003). Cognitive assessment. In L. Nadel (Ed.), *The encyclopedia of cognitive science* (Vol. 1, pp. 526–532). London: Nature Publishing Group.

Kulikowich, J. M., & Alexander, P. A. (1994). Error patterns on cognitive tasks: Applications of polytomous item response theory and log-linear modeling. In C. Reynolds (Ed.), *Cognitive assessment: An interdisciplinary dialogue* (pp. 137–154). New York: Plenum.

Mislevy, R. J. (1994). *Test theory reconceived: CSE technical report 376*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6–20.

Pant, H. A., Rupp, A. A., Tiffin-Richards, S. P., & Köller, O. (2009). Validity issues in standard-setting studies. *Studies in Educational Evaluation, 35*(2–3), 95–101.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.) (2001). *Knowing what students know: The science and design of educational assessment*. Washington, D.C.: The National Academies Press.

Rychen, D.S. (2004). Key competencies for all: An overarching conceptual frame of reference. In D. S. Rychen & A. Tiana (Eds.), *Developing key competencies in education: Some lessons from international and national experience* (pp. 5–34). Paris: UNESCO.

Shavelson, R. J. (2013). An approach to testing & modeling competence. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn, & J. Fege (Eds.), *Modeling and measuring competencies in higher education*. Tasks and Challenges (pp. 29–43). Rotterdam: Sense Publishers.

Toepper, M., Zlatkin-Troitschanskaia, O., Kuhn, C., Schmidt, S., & Brückner, S. (2014). *Advancement of Young Researchers in the Field of Academic Competency Assessment – Report from the International Colloquium for Young Researchers from November 14-16, 2013 in Mainz* (KoKoHs Working Papers, 5). Berlin & Mainz: Humboldt University & Johannes Gutenberg University.

Webb, N. M., Shavelson, R. J., & Steedle, J. T. (2012). Generalizability theory in assessment contexts. In C. Secolsky & B. D. Denison (Eds.), *Handbook on measurement, assessment, and evaluation in higher education* (pp. 132–149). New York, London: Routledge.

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and selecting key competencies* (pp. 45–65). Seattle, WA: Hogrefe and Huber.

Zlatkin-Troitschanskaia, O., Blömeke, S., & Pant, H. A. (2015). Competency Research in Higher Education. [Special Issue] *Peabody Journal of Education, 90*(4).

Zlatkin-Troitschanskaia, O., Förster, M., Brückner, S., & Happ, R. (2014). Insights from a German assessment of business and economics competence. In H. Coates (Ed.), *Higher Education Learning Outcomes Assessment–International Perspectives* (pp. 175–197). Frankfurt: Peter Lang.

Zlatkin-Troitschanskaia, O., Kuhn, C., & Toepper, M. (2014). Modelling and assessing higher education learning outcomes in Germany. In H. Coates (Ed.). *Higher Education Learning Outcomes Assessment–International Perspectives* (pp. 213–235). Frankfurt: Peter Lang.

Zlatkin-Troitschanskaia, O., & Shavelson, R. J. (2015). Competence assessment in higher education. [Special Issue] *Studies in Higher Education, 40*(3).

Zlatkin-Troitschanskaia, O., Shavelson, R. J., & Kuhn, C. (2015). The international state of research on measurement of competency in higher education. *Studies in Higher Education, 40*(3), 393–411.

RESEARCH & PRACTICE IN ASSESSMENT

Notes in Brief

It is difficult to assess the effectiveness of community college initiatives serving men of color when there is a lack of understanding of the nature of the programming taking place. The purpose of this study was to understand the funding streams, interventions, and objectives of programs serving men of color in the community college. This study was motivated by the belief that understanding common interventions, outcomes, and goals illuminates practitioners' perspectives of the personal and institutional barriers facing men of color and the strategies that should be employed to address these barriers. The researchers believe that the information presented in this analysis of minority male initiatives will serve as a reference for understanding common approaches taken in the field for serving men of color in community colleges.

AUTHORS

Fnann Keflezighi, M.A.
San Diego State University

Levi Sebahari
San Diego State University

J. Luke Wood, Ph.D.
San Diego State University

An Analysis of Programs Serving Men of Color in the Community College: An Examination of Funding Streams, Interventions, and Objectives

In 2010, the American Association of Community Colleges (AACC) launched a minority male initiative (MMI) database[1] to catalogue programs, interventions, and initiatives designed to enhance the success of men of color in community colleges (Christian, 2010). The database was implemented as a resource for community college personnel due to the proliferation of efforts focused on supporting men of color in community colleges. These efforts to promote success among men of color are a byproduct of dismal academic outcomes experienced by these students. Specifically, recent data indicate that only 17.1% and 15.4% of Black and Latino men, respectively, will earn a certificate, degree, or transfer from a community college to a four-year institution within three years. In contrast, 27% of White men will achieve the same academic goals within the same time frame. Outcome rates for students who are enrolled with a mixture of part-time and full-time intensity indicate that only 15% and 15.2% of Black and Latino men, respectively, will achieve their goals, while 29.7% of White men will do so (Wood, Harris, & Xiong, 2014). These data demonstrate that community colleges struggle to facilitate success for all men, particularly underrepresented men of color.

CORRESPONDENCE

Email
luke.wood@sdsu.edu

While efforts focused on supporting the academic goals of college men of color have expanded, little is known about the nature of the programming taking place and the structured support on college campuses for these efforts. As such, the purpose of this study was to understand the funding streams, interventions, and objectives of programs serving men of color in the community college.

The researchers supposed that information on funding streams could allude to which entities (e.g., associations, colleges) are most concerned about student outcomes. This study

was motivated by the belief that understanding common interventions, outcomes, and goals illuminates practitioners' perspectives of the personal and institutional barriers facing men of color and the strategies that should be employed to address these barriers. It should be noted that a primary limitation of this article is that the findings represent what is occurring in the field currently and may not necessarily represent best practices for student success. Moreover, colleges that may have programs but are not in the AACC database or do not have publically available information are not included in this analysis. The researchers believe that the information presented in this analysis of MMIs will serve as a reference for understanding common approaches taken in the field for serving men of color in community colleges.

Method and Results

While efforts focused on supporting the academic goals of college men of color have expanded, little is known about the nature of the programming taking place and the structured support on college campuses for these efforts.

Data presented in this study were derived from a content analysis of information pertaining to community college MMIs. The researchers began by reviewing documentation featured on the AACC database. Web searches were conducted to identify additional MMIs. Using available contact information from these searches, the researchers requested documentation from MMIs not in the AACC database. A document analysis was performed on the database information, program brochures, websites, grant proposals, and other program documents. Document analysis is a qualitative procedure for reviewing documents, records, reports, and other data to provide contextual insight into a specific phenomenon (Patton, 2002). Data were coded using an ideas-grouping approach, which involves the identification of recurrent statements or ideas, re-reading of documents for additional references to these ideas, and grouping of ideas into themes (Auerbach & Silverstein, 2003). All data were reviewed, coded, and analyzed with all the researchers present. Data included in this analysis were publically available.

Location and Funding

A total of 129 campus MMIs were included in this analysis. Given that some MMIs were district-level initiatives, 83 distinct programs were identified. These programs are distributed around the nation, with the highest concentration of MMIs in North Carolina (n=46), Texas (n=32), and New York (n=10). These states were followed by programs in Maryland (n=7), Connecticut (n=5), and Florida (n=5). Interestingly, California— which has the largest community college system in the nation with 112 community colleges— had the same number of identifiable programs (n=4) serving men of color as Pennsylvania and South Carolina. This is likely because California already has an existent UMOJA (Black student) and Puente (Latino student) program structure. As such, there was less of a need to establish programs for men of color because there were programs in place for students of color, in general. The size of these programs varied widely, ranging from 9 to 825 students served (M=135).

While some campus MMIs were funded through a single source (46%), the majority had multiple funding streams (54%). Commonly, three or more funding sources were levied to support initiative efforts (47% overall). MMI funding sources often came from a variety of areas including student fees, county funds, college funds, donations, and local governments. Commonly, funding for MMIs was derived from the community colleges themselves (n=39) and their foundations (n=11). Campus funds were typically derived from enrollment services and from the Office of the President. Many initiatives were also funded by private and corporate grants (n=14) and ranged greatly in funding size. A sizeable number of initiatives derived funding for efforts from student fee dollars (n=10), thereby placing the onus of funding student services that are needed for student success directly on the students. Some colleges, often those institutions with the most resources, derived funding from the Department of Education. Often, this occurred through the Predominantly Black Institution (PBI) grants program. Table 1, provides a detailed breakdown of funding streams.

1 http://www.aacc.nche.edu/Resources/aaccprograms/diversity/MinMaleStuSuccessProgs/PagesDefault.aspx

Table 1.

MMI Funding Streams

Source	*N*
Campus funds	39
Private and corporate gifts	14
College foundation	11
Student fees	10
Unspecified grants	10
Federal grants	6
National foundations	4
City council	4
System or consortium funds	3
County funds	2

This study was motivated by the belief that understanding common interventions, outcomes, and goals illuminates practitioners' perspectives of the personal and institutional barriers facing men of color and the strategies that should be employed to address these barriers.

Interventions

The types of interventions employed by MMIs varied greatly. However, the five most common services employed by MMIs were professional skills development, mentoring, college success and survival skills, service-learning, and tutoring. By far, the two most common interventions focused on professional skills development and mentoring; these interventions were employed by 69% and 65% of programs, respectively. Professional development programming was focused primarily on basic conduct training. For example, programs trained students on business etiquette, how to dress (e.g., business attire, formal wear), preparing for job interviews, resume development, and public speaking. Mentoring programs were utilized among MMIs to assist students with socio-cultural and academic transitions to college, and included faculty-to-student mentoring, peer mentoring, and being mentored by professionals in industry and government. Table 2 provides a listing of the most common interventions identified. Many interventions were academic in nature, focused on developing students through advising, tutoring, and literacy. Depending upon the program objectives, other interventions (not listed in Table 2) were employed. For example, some programs offered university tours, health and wellness workshops, financial planning workshops, and internship opportunities.

Table 2

Common Interventions Employed by MMIs

Interventions	Percentage of MMIs that offer intervention
Leadership and professional development	69%
Mentoring	65%
College success/survival skills	48%
Service learning	46%
Tutoring	34%
Academic advising	31%
Cohort study sessions	22%
Counseling	22%
Career planning	22%
Literacy and book clubs	13%

Goals and Outcomes

In this analysis, the researchers also identified commonly employed goals and outcomes of MMIs. For this study, goals referred to "broad statements that can often be incorporated as part of the strategic plan" and are not measureable (Bresciani, Gardner, & Hickmott, 2010, p. 34). In contrast, outcomes "are very detailed and examine a particular competency that we hope students will accomplish" (Bresciani et al., 2010, p. 34). The analysis interpreted competencies to include knowledge, skills, and dispositions that programs sought to foster among men.

Five primary program goals were identified across institutions through this analysis, including engagement, leadership and professional growth, socio-cultural adjustment, personal growth, and academic advancement. Many programs had a specific goal focused on fostering student engagement. It is interesting to note that, as opposed to on-campus engagement, much of the focus on *engagement* centered on civic engagement, community involvement, and developing a social justice orientation. Thus, engagement was defined within the context of one's local community. *Leadership and professional growth* were also identified as a cross-institutional goal. This goal focused on students' future careers—their readiness for and awareness of future employment opportunities. *Socio-cultural* adjustment was an identified goal, as programs sought to aid students' transitions into college climates, cultures, and expectations. Many programs had goals of fostering *personal growth*, with an intensive focus on empowerment, spiritual development, and an understanding of self through a cultural lens. As expected, most programs also had goals of fostering *academic advancement* as it related to students' access to and academic adjustments within college.

MMI funding sources often came from a variety of areas including student fees, county funds, college funds, donations, and local governments.

Within these goals, programs specified numerous outcomes including affective and performance outcomes. Only a handful of programs had outcomes focused on what students should be *learning*. Broadly, these outcomes could be characterized as understanding the meaning of a social justice orientation, learning how to be a collaborative leader, and gaining strategies for a better understanding of self and others. Because so few programs had learning outcomes and these concepts were more often used as affective outcomes, learning outcomes were not addressed in this analysis. Affective outcomes were operationalized as referring to dispositional and emotional growth; while performance outcomes referred to student engagement and student success markers. In total, 13 affective outcomes and 10 performance outcomes were identified. Additionally, the researchers created a curriculum alignment matrix, which linked program interventions with desired outcomes. This matrix, presented in Figure 1, allowed the researchers to further examine specified program outcomes in light of services being offered (Bresciani, 2006). The matrix depicts what services were identified as leading to intended outcomes. After synthesizing the outcomes, the researchers identified seven affective outcomes and six performance outcomes that were (a) recurrent across programs and (b) most clearly linked with service interventions. Affective outcomes commonly targeted by MMIs (and general definitions associated with these outcomes) include:

The types of interventions employed by MMIs varied greatly. However, the five most common services employed by MMIs were professional skills development, mentoring, college success and survival skills, service-learning, and tutoring.

- *Academic self-efficacy* – building students' self-confidence in their abilities to perform academic tasks;
- *Sense of belonging* – creating an environment of support, affirmation, and perceived value from faculty and staff;
- *Personal self-confidence* – building students' self-confidence in their abilities to perform life tasks;
- *Resilience* – empowering students to overcome and succeed in the face of barriers;
- *Locus of control* – instilling a sense of control and responsibility over their academic futures;
- *Self-esteem* – inculcating a realization of self-worth and value; and
- *Racial affinity* – developing a positive racial regard and feeling of connection to one's racial/ethnic community.

Overwhelmingly, these affective outcomes were noncognitive in nature. Only *sense of belonging* (campus ethos outcome) and *racial affinity* (identity outcome) were of primary interest to programs.

Two performance outcomes focused on students' campus engagement: engagement with faculty and the use of academic services (e.g., tutoring, advising, and counseling) on campus. Other performance outcomes were related to student success and included student retention (persistence), achievement (as operationalized through student grades), graduation (referring to the attainment of a certificate or degree), and transfer from the community college to a four-year college or university.

With respect to program interventions, goals, and outcomes, this study found several recurrent themes. The majority of programs focused on professional skills development and mentoring. These services were offered as the primary tools to address a wide range of goals, including engagement, leadership and professional development, socio-cultural adjustment, personal growth, and academic advancement. These goals translated into outcomes that were primarily affective and performance-based, with few programs placing an emphasis on learning outcomes. Performance outcomes encompassed a wide array of student success indicators (e.g., persistence, achievement, graduation). In general, the affective outcomes included noncognitive outcomes such as academic self-efficacy, personal self-confidence, resilience, locus of control, and self-esteem; with only one campus ethos outcome (sense of belonging) and identity outcome (racial affinity) being of programmatic focus.

Five primary program goals were identified across institutions through this analysis, including engagement, leadership and professional growth, socio-cultural adjustment, personal growth, and academic advancement.

Recommendations for Next Steps

Guided by the aforementioned findings, we offer two primary recommendations. First, new MMI programs should employ this study as a framework for better understanding program structures, interventions, and outcomes. While this analysis does not claim to represent promising practices in the field, it does present primary interventions and outcomes being employed at this time in higher education. This study may guide, but should not restrict, discussions on needed outcomes and associated interventions. Second, inquiry should be conducted to determine the efficacy of MMI programs in meeting their outcomes. In particular, researchers and evaluators can use the program alignment matrix (Figure 1), to determine whether identified interventions have an effect on the specified program outcomes. This may provide better insight into which interventions have an effect on performance outcomes, as well as provide insight on which performance outcomes are most influenced by targeted interventions. Third, given that little is known about the efficacy of MMI's, scholars should examine the ways (if at all) programs are being assessed. Such research can also use the program alignment matrix to determine how different outcomes are being measured and evaluated. Fourth, this analysis may inform the development of instruments that can be used to measure common program outcomes employed by MMIs. This will aid MMI leaders in articulating the effect (if any) of their programs on the populations they serve. In total, this analysis provided insight into what is taking place in the field now; further work is needed to explore the efficacy of the approaches identified herein.

	Leadership/ Professional Development	Mentoring	College Survival Skills	Service / Community Learning Opportunities	Tutoring	Academic Advising	Cohort Study Sessions	Counseling	Career Planning	Literacy/ Book Club
Affective Outcomes										
Academic self-efficacy	x	x	x		x	x	x			x
Sense of belonging (connection to faculty & staff)		x		x			x	x		x
Personal self-confidence	x	x	x			x	x	x		
Resilience			x	x			x	x		
Internal locus of control		x								
Self-esteem	x									
Racial affinity		x		x		x	x	x		
Performance Outcomes										
Use of academic services			x			x		x	x	
Student-faculty engagement	x	x	x	x					x	
Persistence (retention)	x	x	x	x	x	x	x	x		
Achievement (GPA)			x		x	x	x			x
Graduation	x	x	x			x		x	x	
Transfer	x		x			x		x	x	

Figure 1. Curriculum Alignment Matrix with Program Interventions and Commonly Targeted Outcomes

References

Auerbach, C. F., & Silverstein, L. B. (2003). *Qualitative Data: An introduction to coding and analysis*. New York, NY: New York University.

Bresciani, M. J. (2006). *Outcomes-based undergraduate academic program review: A compilation of institutional good practices*. Sterling, VA: Stylus Publishing.

Bresciani, M. J., Gardner, M. M., & Hickmott, J. (2010). *Demonstrating student success in student affairs*. Sterling, VA: Stylus Publishing.

Christian, K. (2010). *AACC launches minority male student success database*. American Association of Community Colleges.

Patton, M. Q. (2002). *Qualitative research and evaluation methods*. Thousand Oaks, CA: Sage.

Wood, J. L., Harris III, F., & Xiong, S. (2014). Advancing the success of men of color in the community college: Special issue on the community college survey of men. *Journal of Progressive Policy & Practice, 2*(2), 129-133.