WERA Washington Educational Research Association
Linking Research, Data, and Assessment to Impact Policy and Practice

The WERA Educational Journal

Volume 8, Number 1 November 2015

CONTENTS

Just Pressing Buttons? Validity Evidence for the STAR and Smarter Balanced Summative Assessments

Jack B. Monpas-Huber, Marysville Public Schools

As long as American public school students continue to take tests, those in the field of educational measurement who develop and use tests are exhorted to uphold high standards of practice. For testing experts working in public school districts, this means being very clear about the purposes of various assessments and then being able to describe and produce evidence of validity for those purposes. This paper considers challenges to the validity of the STAR and Smarter Balanced assessments on the grounds of content and anomalies in administration. It then examines correlations between the test scores as evidence of validity. The results show strong correlations.

Introduction

Students in Washington public schools continue to take a steady stream of tests. Last spring, thousands of students across the state took the Smarter Balanced Assessments (SBA) in English language arts and mathematics as their end-of-year test. For many students, these high stakes tests were preceded by various district tests designed to predict their achievement on the SBA and provide intervention to students who showed signs of risk.

As long as testing continues, the field of educational measurement continues to hold its professionals to high standards of practice (AERA, APA & NCME, 2014). It holds validity as the most important consideration, and it exhorts developers and users of tests to strive for valid interpretations and uses of tests (AERA, APA & NCME, 2014). This means, first, articulating a validity argument—the interpretations and uses of a test in a given context (Kane, 2006)—then gathering different kinds of evidence for the validity of those interpretations and uses (AERA, APA & NCME, 2014; Kane, 2006; Messick, 1989). In my experience as a measurement professional charged with managing testing programs in American public school districts, this has meant three tasks: (1) clearly and publicly articulating the purposes of particular assessments and the different decisions informed by the data from these assessments by stakeholders at different levels; (2) framing vocal challenges to the validity of test uses as hypotheses to be investigated with data; and (3) being ready to conceptualize and produce evidence that the data can support those decisions. Validity work is not merely theoretical; it is important on a practical level in order to respond effectively to clearly articulated threats or challenges to validity (Lissitz, 2009). In this paper, I respond to a challenge from local educators to the validity of a district's use of online assessments to make decisions about students. I first consider the purposes and decisions of the assessments. Then I examine correlations between the scores of the assessments as evidence of validity.

Validity Argument for the STAR and Smarter Balanced Assessments

Last spring, over six thousand students in Marysville Public Schools students took the Smarter Balanced Summative Assessments (SBA) in English language arts and mathematics. The purpose of those tests was to measure students' proficiency with the knowledge and skills outlined in the Common Core State Standards in English language arts and mathematics. For high school students, the stakes for these tests are high; students will soon have to pass the tests in order to earn a high school diploma. For younger students, the tests can inform decisions about course placement; students who score below proficiency could be scheduled for an intervention class in place of an elective. For elementary students, scores below proficiency can place students on a list for additional instructional supported services provided by federal and state categorical programs.

For most of these students in my district as well as many of their peers in other districts, the spring SBAs were merely the last in a series of online tests—the STAR Reading and STAR Math tests—that the students took throughout the year to help educators benchmark student progress toward proficiency in reading and mathematics.

These tests had three purposes. One is to function as a universal screener of all students to identify students at risk of not reaching proficiency. A second is progress monitoring—to measure struggling students' progress toward proficiency in response to instructional interventions. A third is to provide outcome measurement for goal setting and program evaluation at the school and district levels. These uses of the STAR assessments to predict student achievement on the SBAs are based on "curriculum-based measurement", a theory of action that high quality, frequent growth measurement along with instructional interventions can help teachers and schools make more informed instructional decisions that will enable more students to meet grade level expectations (Stecker, Fuchs, & Fuchs, 2005).

Validity Challenge: "They're Just Pressing Buttons"

Although online testing is established in the field of testing, it is still very new to many educators in Washington State. As a result, some if not many educators view online testing with some skepticism. Reading assessment is a good example. Reading assessments like the Dynamic Indicators of Basic Early Literacy System (DIBELS) (Good & Kaminsky, 2002) or the Fountas & Pinnell Benchmark Assessment System (Fountas & Pinnell, 2010) require teachers to carefully listen as their students read a passage for a period of time, while observing errors that the students make. Some teachers who have used and understand one-on-one reading assessments like those are skeptical of an online mode of delivery that requires students to read a passage on a screen and respond to a series of selected-response items. These concerns are compounded when students have difficulties demonstrating their knowledge and skills on online tests, as some students did during administration of the SBAs last spring. After days of testing, the SBAs ran long for some students, and fatigue set in. Taken together, to some if not many educators these concerns amount to a deep skepticism of the value of the data from these assessments. Typical remarks included: "I saw students clicking through" and "They're just pressing buttons."

Validity Evidence: Correlations between STAR and SBA Scores

These concerns are understandable. Some students did "click through" one or both sets of online assessments due to lack of understanding of the test or its content, low motivation, low ability to construct responses on performance assessments, and/or test fatigue. The impact of these threats to validity is undeniable. If the test score fails for whatever reason to reflect the student's true knowledge and skills, then the score cannot support the weight of an instructional decision about the student; and when aggregated, the scores produce data that can lead to anomalous or misleading results. The task of the measurement professional is to assess the damage of these threats to the quality of the data. It is to undertake validation work.

Validation is scientific investigation into the meaning of test scores (Messick, 1989). This means suspending the anecdotes to take an empirical approach by reformulating validity threats as hypotheses and then testing them against the data from the assessment. In this case, if too many students answered randomly, then the distribution of scores would be random and not related to achievement. The way to test this hypothesis is to examine correlations between these test scores and scores from other tests of the same content area knowledge and skills.

Correlations are an important form of validity evidence in two ways. First, they are evidence of criterion-related validity, specifically predictive validity (Carmines & Zeller, 1979; Allen & Yen, 2002). This is when a predictor test, such as a STAR screener, is used to predict scores on a future outcome test, called a criterion test, such as SBA. Correlations between predictor and criterion scores are called validity coefficients (Allen & Yen, 2002). Strong validity coefficients between STAR and SBA scores would be evidence for the predictive validity of the STAR tests.

The second is that correlation is a measure (and therefore evidence) of reliability. Reliability is the extent to which a measurement instrument will produce the same or similar results across repeated measurements. Reliability is not

separate from validity; rather, it is an essential form of validity evidence (Messick, 1989) because an unreliable assessment does not measure anything.

Methods

To investigate the impact of these challenges to validity of the data, I examined the correlations among scores from the STAR and SBA assessments from last school year, 2014-15.

The population was students attending Marysville Public Schools in school year 2014-15 who took both the STAR Assessments and the Smarter Balanced Summative Assessments. Specifically, 3,669 students in grades 3-8 and 10-11 took the fall, winter and spring STAR Reading assessments and the Smarter Balanced summative assessments of English Language Arts; and 3,378 students in these same grades (except 10) took the fall, winter and spring STAR Math assessments and the Smarter Balanced summative assessments of mathematics. Marysville students are similar to the state as a whole. Table 1 presents basic demographic information about Marysville students compared to the state.

Table 1: Demographic Information, Marysville and State of Washington

	Marysville	Washington
October 2014 Enrollment	11,398	1,075,107
May 2015 Enrollment	11,227	1,070,756
% Male (October 2014)	50.9	51.5
% Hispanic (October 2014)	21.2	21.7
% American Indian / Alaska Native (October 2014	6.1	1.5
% Black / African American (October 2014)	5.0	4.5
% Native Hawaiian / Other Pacific Islander (October 2014)	0.8	1.0
% White (October 2014)	55.5	57.0
% Two or More Races (October 2014)	9.9	7.1
% Free or Reduced-Price Meals (May 2015)	46.2	45.0
% Special Education (May 2015)	15.0	13.4
% Transitional Bilingual (May 2015)	8.0	10.4

STAR Reading and Mathematics Assessments

The STAR assessments are computer-adaptive assessments of reading comprehension and mathematical ability. Both assessments use 34 items to locate students on a vertical scale spanning across grade levels from kindergarten to grade 12. Scores on both scales range from 0 to 1400. The tests take students approximately 20 minutes to complete (Renaissance Learning, 2014; Renaissance Learning, 2015).

Because 2014-15 was the first year of implementation of the STAR assessments in Marysville, and because teachers were encouraged to use STAR to progress monitor students, many students had multiple STAR scores spanning across months. For this analysis, I examined only students who had a complete set of STAR scores from fall, winter and spring, and a SBA scale score. For each student, for both the STAR Reading and STAR Math tests, I defined the fall STAR score as the maximum STAR score earned in the months of October and/or November, the winter STAR score as the maximum STAR score earned in the months of January and/or February, and the spring STAR score as the maximum STAR score earned in the months of May and/or June.

Smarter Balanced Summative Assessments

The Smarter Balanced assessments are online assessments of English language arts and mathematics. Both assessments have two parts: a computer-adaptive test (CAT) and a performance task (PT). The CAT took students through a series of selected-response and technology-enhanced items designed to produce a reliable student scale score. The PT required students to experience a teacher-led classroom activity designed to provide common background and vocabulary before asking students to engage in a complex task designed to call upon multiple skills. The information gathered from the CAT and the PT placed students on vertical scales of achievement ranging from grade 3 to 11. Scores on both scales range from 2000 to 3000. Each content area assessment was estimated to take students approximately four hours to complete.

Results

The Pearson correlations between the STAR and SBA scores are reported in Table 2. With the exception of Grade 11, which saw very low student participation, correlations are strong across the board. In English language arts, correlations range from a low of .76 to a high of .81, and the spring correlations are generally higher than the fall and winter correlations. The mathematics correlations are higher than the ELA correlations, ranging from a low of .77 to a high of .87, and the spring correlations are generally higher than the fall and winter correlations.

Table 2: Pearson Correlations of Scores on STAR Benchmark Tests to Smarter Balanced Summative Assessment Scores

	English Language Arts				Mathematics			
Grade	N	Fall	Winter	Spring	N	Fall	Winter	Spring
3	653	.78	.78	.79	608	.82	.83	.85
4	573	.78	.78	.81	640	.81	.84	.87
5	579	.76	.79	.80	513	.83	.83	.85
6	550	.78	.79	.79	561	.82	.84	.86
7	449	.79	.80	.79	569	.81	.81	.82
8	443	.79	.80	.81	432	.79	.77	.79
10	261	.78	.79	.74	n/a	n/a	n/a	n/a
11	161	.61	67	.66	55	.50	.59	.52

Note. All correlations are statistically significant at $p < .001$.

Discussion

These correlations are good news. Together they offer important evidence of validity and reliability for the STAR and Smarter Balanced summative assessments. Strong correlations are also evidence of reliability (an essential form of validity evidence) because they mean that only a small fraction of the total variance in observed scores is due to variance from guessing rather than variance in true achievement (Carmines & Zeller, 1979). In more practical terms, strong correlations essentially mean that the two assessments have *sorted* students very consistently. This is an important point for two reasons. One is that it means we can use scores from a screener to make decisions about students even if the screener is not exactly the same as the outcome measure. To invoke the example of the reading assessment used earlier, some reading educators are quick to question the value of an online screener because it does not look or feel exactly like the criterion test using a more authentic form of reading performance. Strong correlations of a screener to the criterion measure suggest that the screener functions like the criterion even if it is

not exactly the same assessment. The second is that strong correlations of a benchmark screener to a criterion mean we can still make decisions from the screener even if students are learning and scoring higher on subsequent administrations of the benchmark screening tests, especially on a vertical scale where there is no grade level ceiling. This is because correlation is about sorting; on a reliable assessment, students who score two standard deviations above the mean in the fall should, even after three months of teaching and learning, score somewhere close to two standard deviations above the mean in the winter. A strong correlation between the screener and the criterion means a student who scores two standard deviations above the mean on the STAR screener also scores somewhere close to two standard deviations above the mean on the content area SBA. To return to the original issue of validity threats, these strong correlations suggest that the assessments produced data of sufficient quality to support instructional decisions about students and higher-level analysis of achievement gaps and program effectiveness despite some anomalies in the administration of the tests. In short, the STAR and SBA data, despite glitches, are not just noise but are in fact giving us real signals about achievement.

This information is reassuring and encouraging, but it should not be viewed as complete for two reasons. One is that this analysis of correlations occurred at the level of scale scores, but far more educators make decisions about students on the basis of performance levels (which are often color coded in databases and data displays) rather than scale scores. This means additional validity work is needed at that level. Such work should include diagnostic accuracy analysis of the sensitivity and specificity rates of different STAR cut scores distinguishing "At or Above Benchmark" from "On Watch" and "Intensive." At this writing, Renaissance Learning is collecting SBA data from districts throughout the Smarter Balanced Consortium in order to conduct a linking study that will produce new STAR benchmark cut scores predicting success on SBA (E. Stickney, personal communication, September 25, 2015).

The second is that strong correlation is not perfect prediction, insofar as some educators view prediction as validity evidence at all. This means that even with strong correlations between assessment scores, educators will still see some surprises and outliers that will capture attention and discussion. This is inevitable, because educators at different levels hold different conceptions of evidence (Coburn & Talbert, 2006), with teachers tending to place more stock in data collected from the classroom rather than standardized test scores (Guskey, 2007). Correlation and diagnostic accuracy work are helpful because they illustrate the pattern as well as the outliers that capture attention and discussion. Hopefully this work will stimulate thinking about validity and good practice of educational measurement in district and schools.

References

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Prospect Heights, IL: Waveland Press.

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

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Newbury Park: Sage Publications.

Coburn, C. E., & Talbert, J. (2006). Conceptions of evidence use in school districts: Mapping the terrain. *American Journal of Education*, 112(4), 469-485.

Fountas, I. C., & Pinnell, G. S. (2010). *Fountas & Pinnell benchmark assessment system*. Portsmouth, NH: Heinemann.

Good, R. H., & Kaminski, R. A. (Eds.). (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Institute for the Development of Education Achievement. Available at http://dibels.uoregon.edu.

Guskey, T. R. (2007). Multiple sources of evidence: An analysis of stakeholders' perceptions of various indicators of student learning. *Educational Measurement: Issues and Practice*,19-27.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education and Praeger Publishers.

Lissitz, R. W. (Ed.) (2009). *The concept of validity*. Charlotte, NC: Information Age Publishing.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.

Renaissance Learning. (2014). *STAR Reading technical manual*. Available at https://resources.renlearnrp.com/us/manuals/sr/srrptechnicalmanual.pdf.

Renaissance Learning. (2015). *STAR Math technical manual*. Available at https://resources.renlearnrp.com/us/manuals/sm/smrptechnicalmanual.pdf.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools*, *42*(8), 795-819.

About the Author

Jack B. Monpas-Huber is the Director of Assessment, Student Information Systems, and Highly Capable Program in the Marysville Public Schools in Marysville, Washington.