VALIDITY EVIDENCE FOR STATISTICS EDUCATION INSTRUMENTS: FINDINGS AND BEST PRACTICES

Charlotte Bolch[1], Hartono Tjoe[2], Stephanie Casey[3], Douglas Whitaker[4], Leigh Harrell-Williams[5], Chris Engledowl[6], Taylor Mulé[5], and Justine Piontek[5]
[1]Midwestern University, Glendale, AZ
[2]The Pennsylvania State University, Reading, PA
[3]Eastern Michigan University, Ypsilanti, MI
[4]Mount Saint Vincent University, Halifax, Nova Scotia
[5]The University of Memphis, Memphis, TN
[6]New Mexico State University, Las Cruces, NM
cbolch@midwestern.edu

Appropriate validity evidence is essential when interpreting scores from tests and instruments. In statistics education, the application of modern measurement theory is limited and not well integrated into research as support for instrument interpretation. The Validity Evidence for Measurement in Mathematics Education project is documenting validity evidence for mathematics and statistics education instruments through a structured literature review to create a searchable instrument database. This paper highlights the status of the work documenting validity evidence for statistics education instruments measuring constructs such as teacher knowledge and attitudes. Many "custom" single-study measures incorporated items from multiple validated instruments and/or added new items without providing evidence for the new instrument. Preliminary information about the types of validity claims and evidence identified from standardized coding is reported.

INTRODUCTION

Validity evidence is a critical component of the interpretation of quantitative assessments in education. A clear understanding of validity and validation lends support to the measurement of that which item developers seek to quantify (Bostic et al., 2019a). The more substantial the interpretation and the use of these instruments and their scores, the greater their potential to improve the quality of evaluation and research findings (Bostic et al., 2019b).

Validity is "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (American Educational Research Association [AERA] et al., 2014, p. 11). The argument-based approach to validity evidence is increasingly evident given that evaluation of validity evidence claims is indispensable—not only those claims inherent to the validity evidence, but also those involving proposed interpretation and use of scores from existing quantitative instruments (Krupa et al., 2019). Consequently, an instrument is in general only as constructive as the degree to which its uses and interpretations may become generalizable (Hill & Shih, 2009).

The Validity Evidence for Measurement in Mathematics Education (VM²Ed) project aims to examine, draw on, synthesize, and curate a framework of validity evidence for existing mathematics and statistics education instruments. Through this project, we strive to determine the consistency and uniformity of this validity evidence in order to bridge the test intentions of the researchers who are item developers and the score uses and interpretations of the researchers who are item users.

The VM²Ed statistics education synthesis group has documented that, in the field of statistics education in particular, the application of modern measurement theory is not only limited, but also appears to be independent of instrument-interpretation research. The current paper highlights the process and status of the effort to document validity evidence for statistics education instruments that measure constructs. Specifically, we focus on statistics education instruments that measure the knowledge and attitudes of students and teachers. In doing so, we hope to bring awareness to issues within the field of statistics education regarding validity evidence for instruments and promote best practices for instrument development and appropriate documentation of validity evidence for existing instruments.

METHODS

The VM²Ed statistics education synthesis group performed three initial rounds of work with regards to identifying the statistics education content to be initially added to the searchable repository

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org

of instruments. Specifically, we began the first round of work in February 2020 by searching journal databases, proceedings papers from the International Conference on Teaching Statistics (ICOTS), and Google Scholar (to capture dissertations and other conference proceedings papers) to identify a list of statistics education instruments noted in publications from the year 2000 and beyond. For the second round, we excluded instruments focused on statistics outside of the statistics education domain (i.e., statistics in biology or psychology), as well as instruments whose authors focus on the interpretation of individual items rather than the instrument as a whole (i.e., ones in which test/instrument total scores are never used). We categorized each instrument as either measuring student knowledge, student attitudes (which include beliefs, perceptions, etc.), or teacher knowledge and attitudes. We then conducted a more specific search for each included instrument to identify publications citing the instrument, classifying each publication as "having used the instrument" and/or "containing validity evidence." We also noted which instruments are single use—meaning the instrument was created and used exclusively by a single researcher (or research team). Work on the second round was completed in May 2021. We are currently in the third round, which encompasses the identification and classification of pieces of validity evidence for each instrument presented in these publications. Validity evidence is classified using the five types of validity evidence (test content, internal structure, response processes, relations to other variables, and consequences of testing) specified in the 2014 *Standards for Educational and Psychological Testing* (AERA et al., 2014), using a tagging system for sources of evidence (e.g., cognitive interview for response processes) and a framework developed by the VM²Ed project team. The framework developed is a standardized data collection process to ensure all relevant information from the articles is collected so that the work of the statistics education synthesis group is incorporated into the database for the overall VM²Ed project. Additionally, we are documenting the presence of use and interpretation statements, and whether the authors make explicit claims about the validity evidence. The methods used for the coding of explicit claims was developed by the VM²Ed project team and all researchers were required to attend a training workshop. Then, multiple rounds of coding were conducted in pairs in the statistics education synthesis group to establish consistency of coding explicit claims and non-explicit claims as well as identifying types of validity evidence. Non-explicit claims do not have a connection between the validity evidence and the claim made. Within this paper, we use descriptive statistics to present a subset of results from the current round of work, focusing on the number of instruments classified as "single use" and the types and sources of evidence for the teacher-focused statistics education instruments.

RESULTS

Single Use Instruments

Table 1 presents the number and percentage of single use instruments, both overall and by instrument type. Our study found that single use instruments were the predominant kind used in manuscripts published between January 2000 and May 2021. Of the 107 statistics education instruments identified by our study, 81 (76%) are single use. The percentages of single use instruments across the three instrument types (student attitude: 77%, student knowledge: 71%, and teacher knowledge and attitudes: 76%) showed a consistent prevalence of single use instruments across all instrument types. When we analyzed the single use instruments for the types of validity evidence they provide, we found that most single use instruments have only a few types of validity evidence. Generally speaking, the researchers designing and using single use instruments provided some test content validity evidence, mostly by designing instruments that align with established frameworks or standards, reviewing the literature related to the instrument's construct, including experts in the creation of the instrument, or revising instrument items to improve their wording after piloting of the instrument. Factor analysis was commonly reported, although not necessarily described as evidence of internal structure. Computing an internal consistency measure was commonly reported to demonstrate the reliability of a single use instrument. The remaining types of validity evidence were seldom used; in particular, consequences of testing validity evidence was very rare for single use instruments.

Table 1. Number and percentage of single-use instruments

	Number of Single Use Instruments	Total Number of Instruments	Percentage of Single Use Instruments
Student Attitudes Instruments	36	47	77%
Student Knowledge Instruments	33	43	77%
Teacher Knowledge & Attitudes Instruments	12	17	71%
Total	81	107	76%

Summary of Results for Teacher-Focused Statistics Education Instruments

The coding of the framework for sources and types of evidence was broken into two components: one for explicit claims made and one for non-explicit claims made. A total of 17 teacher-based instruments were identified. An overview of the evidence for non-explicit and explicit claims for teacher-education instruments is provided in Table 2. Of these instruments, seven have been used in populations outside of the intended population of in-service and pre-service teachers. For explicit and non-explicit claims, it was possible that multiple sources of evidence might be coded for one instrument. In addition, it was possible that a source of evidence could have had multiple types of evidence. Table 2 shows the types of evidence and sources of evidence for articles for the teacher-based statistics instruments for explicit and non-explicit claims. There were 90 distinct types of evidence for non-explicit claims. The most common source of evidence identified from the articles specifying non-explicit claims about the instruments was test content (64.44%). Common types of evidence reported for test content included "alignment with frameworks/standards/theory/learning trajectory," "data from experts," and "revision process." Other types of evidence sources were internal structure (8.89%), relation to other variables (12.22%), and response process (10%). More limited types of evidence for non-explicit claims made were reliability (2.22%) and consequences of testing (2.22%).

For the explicit claims made in articles about teacher-based instruments, there were a total of 47 sources of evidence. Multiple sources of evidence were sometimes used to support a claim. The most common source of evidence identified for explicit claims was internal structure (38%), with types of evidence reported as "cluster analysis," "factor analysis" (both confirmatory factor analysis and principal component analysis), and Rasch modeling. Only a single source of evidence was identified for consequences of testing looking at item functioning (2.13%). The other sources of evidence identified for explicit claims were relations to other variables (17.02%), reliability (23.4%), and test content (19.15%). Overall, most explicit claims were made about reliability and internal structure, while most non-explicit claims were about relationship to other variables and test content.

To understand how a specific teacher-focused instrument was coded for the framework and how the types of evidence were identified, an example instrument is used. An example of a family of instruments having had a broad range of sources of validity evidence was the Self-Efficacy to Teach Statistics in High School (SETS-HS) (Harrell-Williams, Lovett, Lee, et al., 2019) and in Middle School (SETS-MS) (Harrell-Williams et al., 2014) instruments (Harrell-Williams, Lovett, Lesser, et al., 2019). A total of nine articles were identified that had sources of validity evidence for the SETS instrument. Most of the sources of validity evidence were relations to other variables, test content, and response process for non-explicit claims. For explicit claims, the main sources of evidence were reliability ("internal consistency–Rasch reliability") as an example of a type of evidence; internal structure, which focused on Rasch modeling; and relations to other variables using "correlation analysis."

DISCUSSION

As statistics education research literature has grown, so too has the popularity of using instruments and tests to measure outcomes of interest. Current best practices for developing and using instruments and tests (e.g., AERA et al., 2014) emphasize that score interpretations for specific uses should always be supported by validity evidence. Although there are many available instruments developed for use in statistics education research, the state of validity evidence for these instruments is heterogenous at best: some instruments have been widely adopted and have large bodies of validity evidence supporting their use, whereas many/most are single use instruments created and used only by a single researcher or project. Although such single use instruments can be valuable in some cases, the

majority of such instruments are presented with little to no validity evidence, leaving the field with a fragmented ability to effectively build from such research. Thus, we advocate the collection and presentation of validity evidence for these instruments, which are quite prevalent in the literature. Prior to the VM²Ed project, there had been no systematic effort to document the available instruments in statistics education outside of meta-analyses focused on specific constructs. The database of available instruments in statistics and mathematics education resulting from the work of the VM²Ed project will provide a centralized resource for documenting available statistics education instruments and tests together with the validity evidence supporting them.

Table 2. Evidence for non-explicit and explicit claims for teacher-education instruments

Sources of Evidence	Non-explicit claim *n* (%)	Explicit claim *n* (%)
Internal Structure		
Cluster analysis	1 (1.11%)	1 (2.13%)
Factor Analysis–Confirmatory Factor Analysis (CFA)	3 (3.33%)	7 (14.89%)
Factor Analysis–Exploratory Factor Analysis/Exploratory Structural Equation Modeling	1 (1.11%)	-
Factor Analysis–Principal Component Analysis (PCA)	1 (1.11%)	2 (4.26%)
Rasch modeling	2 (2.22%)	8 (17.02%)
Subtotal	8 (8.89%)	18 (38.3%)
Relations to Other Variables		
Correlation analysis	6 (6.67%)	7 (14.89%)
Structural Equation Model (SEM)	1 (1.11%)	-
Discriminant validity	1 (1.11%)	1 (2.13%)
Statistical Testing (e.g., *t*-test, regression, and chi-square)	3 (3.33%)	-
Subtotal	11 (12.22%)	8 (17.02%)
Reliability		
Internal consistency or alternatives–Cronbach's Alpha	2 (2.22%)	10 (21.28%)
Kuder-Richardson formula 20	-	1 (2.13%)
Subtotal	2 (2.22%)	11 (23.4%)
Response Process		
Focus groups	2 (2.22%)	-
Think alouds	4 (4.44%)	-
Student written work	3 (3.33%)	-
Subtotal	9 (10%)	-
Test Content		
Alignment with frameworks/standards/theory/learning trajectory	16 (17.78%)	2 (4.26%)
Construct definition	2 (2.22%)	1 (2.13%)
Data from experts	14 (15.56%)	1 (2.13%)
Field work	3 (3.33%)	-
Fairness of content	1 (1.11%)	-
Literature review	7 (7.78%)	-
Participant-generated content	1 (1.11%)	-
Revision process	14 (15.56%)	5 (10.64%)
Subtotal	58 (64.44%)	9 (19.15%)
Consequences of Testing		
Explicit intended uses and interpretations and warning against inappropriate uses	2 (2.22%)	-
Item functioning such as DIF–unknown subgroups had to know	-	1 (2.13%)
Subtotal	2 (2.22%)	1 (2.13%)
Total	90	47

The VM²Ed instrument database has direct and indirect implications for both developers and users of instruments. The proliferation of single use instruments presents numerous challenges for statistics education researchers. First, researchers seeking to use an instrument to measure a specific construct face challenges in determining if a suitable instrument even exists. There are dozens of single use instruments for measuring aspects of statistics attitudes alone, and without a centralized location for instruments, it is conceivable that researchers may fail to locate appropriate instruments and resort to creating new ones. Second, even when researchers do locate an instrument that measures the construct of interest, single use instruments tend to have limited validity evidence supporting their interpretations. Consequently, researchers may not be satisfied with the available validity evidence for some reason and may choose to develop their own instruments with the specific validity evidence they are seeking. Third, the results stemming from two instruments ostensibly measuring the same construct may be difficult to compare, and researchers seeking to synthesize the results of studies of that construct might be faced with irreconcilable findings or findings that can only be compared after a laborious investigation, which could lead to fragmented research in some areas.

CONCLUSION

Although a database of statistics education research instruments should make finding appropriate instruments easier for researchers, a perspective shift regarding which types of research activities are valuable is also necessary. That is, studies that contribute to the validity evidence of existing instruments should be recognized as essential research activities, and the development of new instruments without a clear, articulable need should be viewed with skepticism. When there is a need for new instrument development work, rigorous development processes should be used, and multiple forms of validity evidence should be provided by the developers so that there is a foundation of validity evidence upon which to build.

REFERNCES

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

Bostic, J., Krupa, E., & Shih, J. (Eds.). (2019a). *Assessment in mathematics education contexts: Theoretical frameworks and new directions*. Routledge.

Bostic, J., Krupa, E., & Shih, J. (Eds.). (2019b). *Quantitative measures of mathematical knowledge: researching instruments and perspectives*. Routledge.

Harrell-Williams, L. M., Lovett, J. N., Lee, H. S., Pierce, R. L., Lesser, L. M., & Sorto, M. A. (2019). Validation of scores from the high school version of the self-efficacy to teach statistics instrument using preservice mathematics teachers. *Journal of Psychoeducational Assessment*, *37*(2), 194–208. https://doi.org/10.1177/0734282917735151

Harrell-Williams, L. M., Lovett, J. N., Lesser, L. M., Lee, H. S., Pierce, R. L., Murphy, T. J., & Sorto, M. A. (2019). Measuring self-efficacy to teach statistics in grades 6–12 mathematics teachers. In J. Bostic, E. Krupa, & J. Shih (Eds.), *Assessment in mathematics education contexts: Theoretical frameworks and new directions* (pp. 147–171). Routledge.

Harrell-Williams, L. M., Sorto, M. A., Pierce, R. L., Lesser, L. M., & Murphy, T. J. (2014). Validation of scores from a new measure of preservice teachers' self-efficacy to teach statistics in the middle grades. *Journal of Psychoeducational Assessment, 32*(1), 40–50. https://doi.org/10.1177/0734282913486256

Hill, H. C., & Shih, J. C. (2009). Examining the quality of statistical mathematics education research [Research commentary]. *Journal for Research in Mathematics Education*, *40*(3), 241–250. https://doi.org/10.5951/jresematheduc.40.3.0241

Krupa, E. E., Carney, M., & Bostic, J. (2019). Argument-based validation in practice: Examples from mathematics education. *Applied Measurement in Education, 32*(1), 1–9. http://doi.org/10.1080/08957347.2018.1544139