THE SEARCH FOR VALIDITY EVIDENCE FOR INSTRUMENTS IN STATISTICS EDUCATION: PRELIMINARY FINDINGS

Douglas Whitaker, Charlotte Bolch, Leigh Harrell-Williams, Stephanie Casey,
Corinne Huggins-Manley, Christopher Engledowl and Hartono Tjoe
Mount Saint Vincent University, Canada
Midwestern University, USA
University of Memphis, USA
Eastern Michigan University, USA
University of Florida, USA
New Mexico State University, USA
Pennsylvania State University, USA
douglas.whitaker@msvu.ca

Interpreting results from instruments requires appropriate validity evidence. However, evolution in the fields of educational measurement and statistics education means that the validity evidence supporting instruments is often narrowly focused. For the Validity Evidence for Measurement in Mathematics Education project, we are systematically documenting validity evidence for instruments used to measure constructs in statistics education (such as knowledge and attitudes) for students and instructors. The researchers identified instruments measuring statistics-specific constructs, where and how these instruments were used, and validity evidence supporting their use. A structured literature review approach was used to identify instruments developed since 2000 and studies that used them or contained relevant validity evidence. Validity evidence for each instrument was documented using a standardized system. Preliminary information about the instruments identified, the frequency of their published use, and the amount of published work containing validity evidence will be presented.

INTRODUCTION

Appropriate and relevant validity evidence is necessary for the interpretation of results from instruments and assessments. Validity is defined by the *Standards for Educational and Psychological Testing* as "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). Validity is a judgment based on the degree to which empirical evidence and theoretical frameworks support the adequacy and appropriateness of inference and actions based on test scores (Messick, 1989). The contemporary approach to validity is counter to the tripartite view of validity (criterion-related, content, and construct) and instead supports an argument-based approach (AERA et al., 2014; Krupa et al., 2019) where the intended interpretation(s) and use(s) of the instrument or test scores are specified and then validity evidence is collected to support the claims made in this argument.

For the *Validity Evidence for Measurement in Mathematics Education* (V-M^{2}ED) project (NSF Grant No. DRL #1920619 & #1920621), statistics education and mathematics education researchers with an interest in educational measurement are working in small synthesis groups, systematically documenting the validity evidence available for instruments used in their respective fields. The ultimate goal for this project is a searchable database of instruments designed for mathematics and statistics education that includes the validity evidence supporting the uses of each instrument. The focus of the group for statistics education (K-20) instruments and tests was to identify instruments used to measure constructs important in statistics education (e.g., knowledge and attitudes) for students and instructors. The other five synthesis groups focus on mathematics education instruments and tests in elementary (K-6) grades, secondary (7-12) grades, undergraduate/graduate education, teacher education tests, and teacher education instruments. Data science education is nascent and few data science-specific instruments with validity evidence exist (Unfried et al., 2021).

METHOD

Each synthesis group is documenting validity evidence over three rounds of searching. The Round 1 search focused on identifying instruments/tests to include in the database, and the Round 2 search focused on identifying sources that may provide validity evidence for the instruments. The Round 3 search will focus on identifying specific validity claims and supporting evidence from the

In: R Helenius, E Falck (Eds.), *Statistics Education in the Era of Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
Aug-Sept 2021, Online conference. iase-web.org/Conference_Proceedings.php

Round 2 sources. The purpose of this paper is to quantify the number of statistics education instruments with and without validity evidence based on the literature searches done in Rounds 1 and 2; we intend to complete Round 3 by Fall 2022. Documentation of the validity evidence will use a framework standardized across all groups in the larger V-M^{2}ED project. Additionally, systematic checking of the work done in Round 2 will occur simultaneously with Round 3.

One aim of the paper is to identify the number of instruments identified by the intended population, instrument type, and item type. The second aim of the paper was to focus on three statistics education instruments and describe using counts the numbers of articles using each instrument and whether or not they seem to provide evidence for its use. The researchers started by identifying instruments measuring statistics-specific constructs by searching databases with specific search terms and examining sources where statistics education research is published (e.g., *Statistics Education Research Journal*, *Journal of Statistics Education*, and the International Conference on Teaching Statistics proceedings). In Round 1 of the search, only instruments published since 2000 were included. Once an instrument was found, the following information was recorded: name of instrument, citation of the article, abstract, grade level, and whether the instrument is intended to be administered to students or teachers. The exclusion criteria for this part of the search process included:

- instruments related to probability knowledge, instruments not focused on statistics that happened to be used in a statistics study (e.g., a math attitudes scale used with statistics students), and instruments focused on biostatistics and/or the health sciences
- instruments that were only collections of individual items analyzed separately (i.e., the responses were not summed or averaged)
- instruments that were primarily used in sources employing a language other than English (e.g., studies with Spanish-speaking students published in English for which the documentation of the instruments was largely published in Spanish)

Instruments that were intended only for local use were also excluded from the primary search (e.g., final exams within a specific course).

Before Round 2 started, the identified statistics education instruments were grouped into three broad categories: 1) student attitudes, beliefs, and perceptions, 2) student knowledge, and 3) teacher instruments. The second round of the instrument search consisted of a detailed literature search for any peer-reviewed research articles, conference proceedings papers, and dissertations mentioning the instrument identified from round one, with smaller groups assigned to each of the three categories. No year limit was imposed in this round; instruments that had not been included in the first round were now included when they were discovered, regardless of year of publication. Once an article or conference paper was identified that referenced the instrument, the following information was collected about the article: citation, abstract, grade level, whether the article was peer reviewed, whether the article actually used the instrument: was a conceptual paper with a substantial focus on the instrument germane to validity evidence, or neither, and whether the article seemed to contain validity evidence based on the title, abstract, or methods section. A source was judged to seem to have validity evidence based on a cursory examination by a team member looking for descriptions of the instrument development process, relevant statistical results, etc.; we erred on the side of inclusion in Round 2.

RESULTS

Across both Rounds 1 and 2, a total of 111 instruments were identified: 50 related to student attitudes, beliefs, and perceptions (SA); 45 related to student knowledge (SK), and 16 related to teachers (TCH). Many of these instruments were seldom used; often they were only used in a single source (the one in which it was first described) or in a small number of articles by the same research team. A few instruments were cited by hundreds of sources, such as the Comprehensive Assessment of Outcomes in a First Statistics course (CAOS; delMas et al., 2007), the Statistics Anxiety Rating Scale (STARS; Cruise et al., 1985), and the Survey of Attitudes Toward Statistics (SATS; Schau, 1992, 2003). There was considerable variation in the number of citations among the remaining instruments, which is a function of both the age of the instrument and the extent to which it has been adopted by the field. While the instruments were initially grouped into broad categories based on the population with which they were intended to be used, more detailed information about the populations with which they were actually used was recorded (see Table 1).

As seen in Table 1, instruments that are intended for use with students (SA, SK) are largely

used with students, though there was appreciable use of these instruments with pre-service teachers and in-service teachers. Conversely, instruments designed for K-12 teachers (i.e., K-6 and 7-12 teachers) were only used with populations of K-12 teachers. The two instruments intended for K-12 teachers that were used with graduate students were intended for use with graduate teaching assistants - the Graduate Student Statistics Teaching Inventory (GSSTI; Justice et al., 2017) and the Graduate Students' Experiences Teaching Statistics inventory (GETS; Justice, 2017).

Table 1. Number of instruments used with each population by intended population

Population of use	SA	SK	TCH	Total
Elementary/Primary/K-6 Students	3	5	0	8
Secondary/7-12 Students	13	13	0	26
Undergraduate Students	37	34	0	71
Graduate Students	16	7	2	25
Pre-Service Teachers (Undergrad/MAT/etc.)	6	1	1	8
Elementary/Primary/K-6 Teachers	2	3	5	10
Secondary/7-12 Teachers	4	3	8	15
Tertiary Instructors	5	0	6	11
Other	3	3	1	7

Note. Some instruments were used with multiple populations.

Instruments were also categorized based on their design; multiple instrument types could be used to classify an instrument. Table 2 presents that counts of the types of instruments in each group, and Table 3 presents the counts of the item types used by instruments in each group. As shown in Table 2, Likert/Rating Scales and Summative Assessments were the most prevalent instrument type among those examined in Round 2. Across the SA and TCH instruments, only a single instrument measuring attitudes, beliefs, or perceptions used was not primarily a Likert/Rating Scale: the GETS (Justice, 2017), which was classified as a Survey because it used a multiple-choice item format. (The other TCH instruments that were classified as Summative or Survey were not intended for measuring attitudes, beliefs, or perceptions.)

Table 2. Number of instruments of each instrument type

Instrument Type	SA	SK	TCH	Total
Likert/Rating Scale	47	3	11	61
Summative	0	36	2	38
Survey	0	3	4	7
Diagnostic	0	6	0	6
Formative	0	7	0	7
Observation	0	2	0	2
Missing	1	0	0	1

Note. Some instruments were classified as having multiple types.

Table 3. Number of instruments using different item types

Item Type	SA	SK	TCH	Total
Free response	2	19	3	24
Multiple choice	2	34	6	42
Short answer	0	10	0	10
Likert scale	49	4	13	66
Yes/No	1	0	0	1
Other	0	2	0	2
Missing	1	0	0	1

Note. Some instruments included multiple item types.

The types of items used by each instrument were also recorded and are shown in Table 3.

While almost all of the SA instruments were classified as being a Likert/Rating Scale, these instruments sometimes included items in other formats to complement the primary Likert-type items. Among the SK instruments, multiple choice formats were the most widely used, followed by free response and short answer. Likert scales were also the most widely used format among the TCH instruments, reflecting that a majority of these instruments assessed attitudes, beliefs, or perceptions. Other TCH instruments that measured knowledge or other characteristics of instructors used free response and multiple-choice item types.

We will now present detailed information about the sources examined for three instruments: the SATS family of instruments (Schau, 1992, 2003) from the SA group, the Levels of Conceptual Understanding in Statistics (LOCUS) family of instruments (Jacobbe et al., 2014; Whitaker et al., 2015) from the SK group, and the Self-Efficacy for Teaching Statistics (SETS) family of instruments (Harrell-Williams et al., 2014a, 2014b) from the TCH group. These instruments were chosen because they typified instruments that had many sources that were examined in Round 2. Two tables are shown below: Table 4 records the number of sources that used the instrument and the number of sources that seemed to provide validity evidence for the instrument, and Table 5 shows the number of sources that seem to provide (or not provide) validity evidence for each of the populations it was used with only for sources that actually used the instrument.

Table 4. The numbers of sources using each family of instruments and whether or not they seem to provide evidence for its use

	Does the source seem to provide the validity evidence?					
	SATS (SA)		LOCUS (SK)		SETS (TCH)	
Was each instrument used in the source?	Yes	No	Yes	No	Yes	No
Yes	110	150	7	11	10	2
No	0	282	0	2	0	6
Total	110	432	7	13	10	8

Note. Some sources may have used more than one instrument.

Table 5. The number of sources that do and do not seem to provide validity evidence for each population only for sources that used the family of instruments

	Does the source seem to provide the validity evidence?					
	SATS (SA)		LOCUS (SK)		SETS (TCH)	
Population of use	Yes	No	Yes	No	Yes	No
Elementary/Primary/K-6 Students	0	0	0	0	0	0
Secondary/7-12 Students	1	4	***3***	***3***	1	0
Undergraduate Students	***81***	***120***	1	3	0	0
Graduate Students	5	10	0	0	0	0
PSTs (Undergrad/MAT/etc.)	4	4	0	0	***9***	***2***
Elementary/Primary/K-6 Teachers	1	1	0	0	0	0
Secondary/7-12 Teachers	0	2	1	4	2	0
Tertiary Instructors	0	0	0	0	0	0
Other (write in column to the right)	6	4	4	5	0	0
Missing	16	10	0	0	0	0

Note. Some instruments were used with multiple populations within the same source. The original population for which validity evidence was documented is indicated with ***bold italics***.

The most striking feature of Tables 4 and 5 are the numbers of articles that used an instrument but seem to not provide validity evidence supporting its use—especially when used with a population other than for which it was intended. As noted by the *Standards*, "Validation is the joint responsibility of the [instrument] developer and [instrument] user" (AERA et al., 2014, p. 13). For example, a small but growing number of studies are using the SATS family of instruments with K-12 teachers despite the fact it was intended for use with students in introductory statistics courses. No direct validity evidence supporting its use with K-12 teachers is provided by the instrument developers, suggesting that the onus of providing validity evidence in these new populations falls to the users of the

instruments. While validation arguments for using the SATS family of instruments with other student populations such as pre-service teachers and secondary students may be more straightforward than validation arguments supporting their use with teachers, the onus for providing validity evidence rests with the users of the instruments because these are still populations other than the intended population of students in introductory statistics courses at the university level with direct validity evidence that was not provided by the instrument developers.

Similarly, the LOCUS family of assessments was developed for use with students in grades 6-12; use of the LOCUS assessments in other populations such as undergraduate students or with secondary school teachers is beyond the scope of the validity arguments provided by the test developers. Among the SETS family of instruments, only two articles do not seem to provide validity, but these used the SETS instruments with their intended population and cite the validation articles.

DISCUSSION

Among the more than one hundred instruments identified that are specific to statistics education, 96 were designed to measure constructs of interest with student populations. However, there was interest in using some of these student instruments with other populations, especially with teachers. For example, 20 articles used the SATS with populations other than undergraduate students or graduate students a total of 27 times; these studies tended to be more recent, with the oldest published in 2004 and half published since 2014. While some of the sources that used instruments with populations outside of the intended one were classified as providing validity evidence supporting this, many other sources did not provide such evidence. The 2014 *Standards* makes clear that validity is not an inherent property of an instrument and that "statements about validity should refer to particular interpretations for specific uses" (AERA et al., 2014, p. 11). While instrument developers are responsible for providing initial validity evidence consistent with uses they intend for an instrument, validation is an ongoing process that also involves evidence from users of instruments.

V-M^{2}ED project's ongoing systematic documentation of the validity evidence supporting the uses of instruments in statistics education is already providing empirical evidence that, in many cases, users of instruments are not directly contributing to the body of validity evidence for an instrument. In the next round of reviews, the specific types of validity evidence provided by each source will be documented and categorized. The resulting database should be useful for anyone seeking to select or use an instrument for use with students or teachers in mathematics or statistics education.

Beyond the results presented above, two problematic patterns emerged that were noted in discussions by the review team; empirical evidence of these patterns will be gathered in the next round of the review. First, instruments are being used outside of their intended population without providing validity evidence to support this use. Second, many instruments have been developed that intend to measure the same constructs with the same populations as other existing instruments. While it will not be quantified until the next round of the systematic search, many of these instruments discovered were only used a few times, often by the initial developers. Increasing the number of instruments intended to measure the same construct without a clear, articulable reason why a new instrument is needed complicates the field and makes comparisons across studies more difficult. It is possible that the field may be better served by these efforts being spent instead on documenting validity evidence supporting (or not supporting) the use of existing instruments in various populations.

While pervasive issues about documenting validity evidence for instruments may not be remedied quickly, individual researchers can strengthen the field of statistics education by adopting best practices when using existing instruments and developing new ones and resources supporting this are becoming more common. Flake and Fried (2020) provide an accessible overview of what they term *questionable measurement practices* that researchers using instruments should be aware of, and readers seeking examples of validation studies consistent with contemporary measurement practices can find some in two books published as part of the V-M^{2}ED project (Bostic et al., 2019a, 2019b).

CONCLUSION

Within the V-M^{2}ED project, the statistics education group reviewed documentation of statistics education instruments/tests and identified over one hundred instruments designed to measure various constructs in students and teachers. Validity evidence about an instrument/test "is best thought of as a program of research in which one attempts to obtain a body of evidence that, taken as a whole,

would support the intended uses of and inferences from the test scores" (Bandalos, 2018, p. 263). Our findings illustrate that (1) there is great opportunity for statistics education researchers to contribute to the body of validity evidence for using existing instruments, especially when used in a new population or context than intended by the instrument developers, and (2) the project's searchable database may bring greater attention to those instruments that might be under-utilized and curb the creation of new instruments that overlap with existing instruments.

REFERENCES

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

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences.* New York City, NY: Guilford Press.

Bostic, J. D., Krupa, E. E., & Shih, J. C. (Eds.). (2019a). *Assessment in Mathematics Education Contexts: Theoretical Frameworks and New Directions* (1st ed.). Routledge.

Bostic, J. D., Krupa, E. E., & Shih, J. C. (Eds.). (2019b). *Quantitative Measures of Mathematical Knowledge: Researching Instruments and Perspectives* (1st ed.). Routledge.

Cruise, R. J., Cash, R. W., & Bolton, D. L. (1985). Development and validation of an instrument to measure statistical anxiety. *Proceedings of the Section on Statistical Education, American Statistical Association,* 92–98.

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28–58.

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

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.

Harrell-Williams, L., Sorto, M. A., Pierce, R., Lesser, L. M., & Murphy, T. J. (2014). Using the sets instruments to investigate sources of variation in levels of pre-service teacher efficacy to teach statistics. *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9).*

Jacobbe, T., Case, C., Whitaker, D., & Foti, S. (2014). Establishing the validity of the LOCUS assessments through an evidenced-centered design approach. In K. Makar & R. Gould (Eds.), *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014).*

Justice, N., Zieffler, A., & Garfield, J. (2017). Statistics graduate teaching assistants' beliefs, practices and preparation for teaching introductory statistics. *Statistics Education Research Journal, 16*(1), 294-319. https://iase-web.org/documents/SERJ/SERJ16(1)_Justice.pdf

Justice, N. (2017). *Statistics Graduate Students' Professional Development for Teaching: A Communities of Practice Model* [University of Minnesota].

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.

Lem, S., Onghena, P., Verschaffel, L., & Van Dooren, W. (2013). The heuristic interpretation of box plots. *Learning and Instruction, 26*, 22–35. https://doi.org/10.1016/j.learninstruc.2013.01.001

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). American Council on Education.

Schau, C. (1992). *Survey of Attitudes Toward Statistics (SATS-28).* http://evaluationandstatistics.com/

Schau, C. (2003). *Survey of Attitudes Toward Statistics (SATS-36).* http://evaluationandstatistics.com/

Unfried, A., Posner, M., Bond, M., Kerby-Helm, A., Bolon, W., Whitaker, D., & Batakci, L. (2021). *Why Do We Need Yet ANOTHER Instrument Measuring Student Attitudes?* Presentation at the 2021 Joint Statistical Meetings (JSM), Virtual. http://sdsattitudes.com/wp/jsm2021/

Whitaker, D., Foti, S., & Jacobbe, T. (2015). The levels of conceptual understanding in statistics (LOCUS) project: Results of the pilot study. *Numeracy, 8*(2), Article 4.

Zimmerman, W. A., & Austin, S. R. (2018). Using attitudes and anxieties to predict end-of-course outcomes in online and face-to-face introductory statistics courses. *Statistics Education Research Journal, 17*(2), 68–81. https://iase-web.org/documents/SERJ/SERJ17(2)_Zimmerman.pdf