The search for validity evidence for instruments in statistics education: preliminary findings

IASE Satellite Conference

2 September 2021

Douglas Whitaker
Mount Saint Vincent University

Charlotte Bolch
Midwestern University

Leigh Harrell-Williams
University of Memphis

Stephanie Casey
Eastern Michigan University

Corinne Huggins-Manley
University of Florida

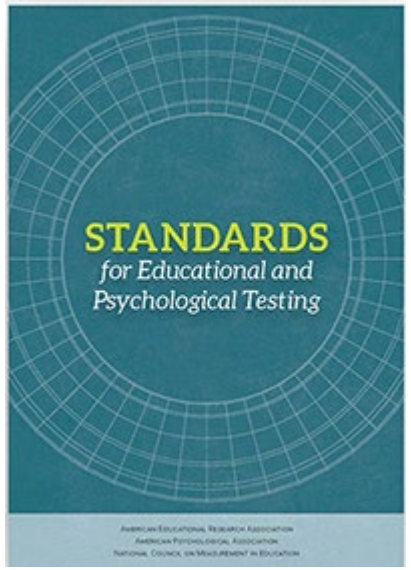
Christopher Engledowl
New Mexico State University

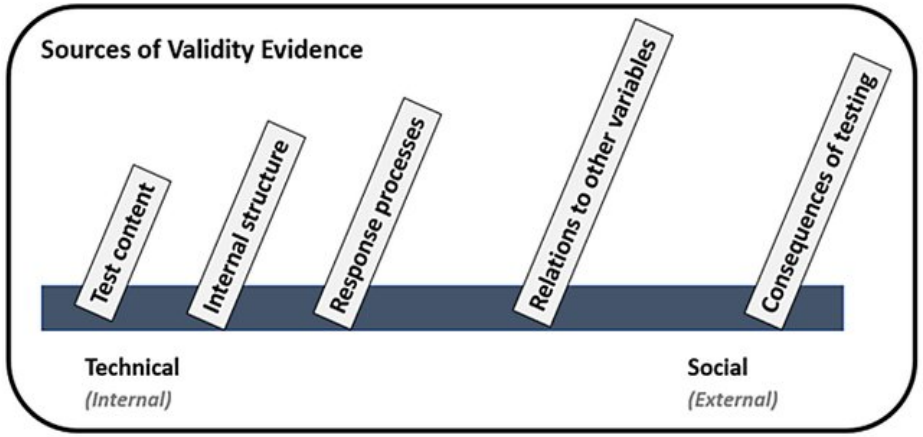
Hartono Tjoe
Pennsylvania State University

Why talk about validity and validity evidence?

Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014)

- Validity is “the **most fundamental consideration** in developing tests and evaluating tests” (p. 11)
- “Validity is a unitary concept. **It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use.** Like the 1999 Standards, this edition refers to types of validity evidence, rather than distinct types of validity.” (p. 14)



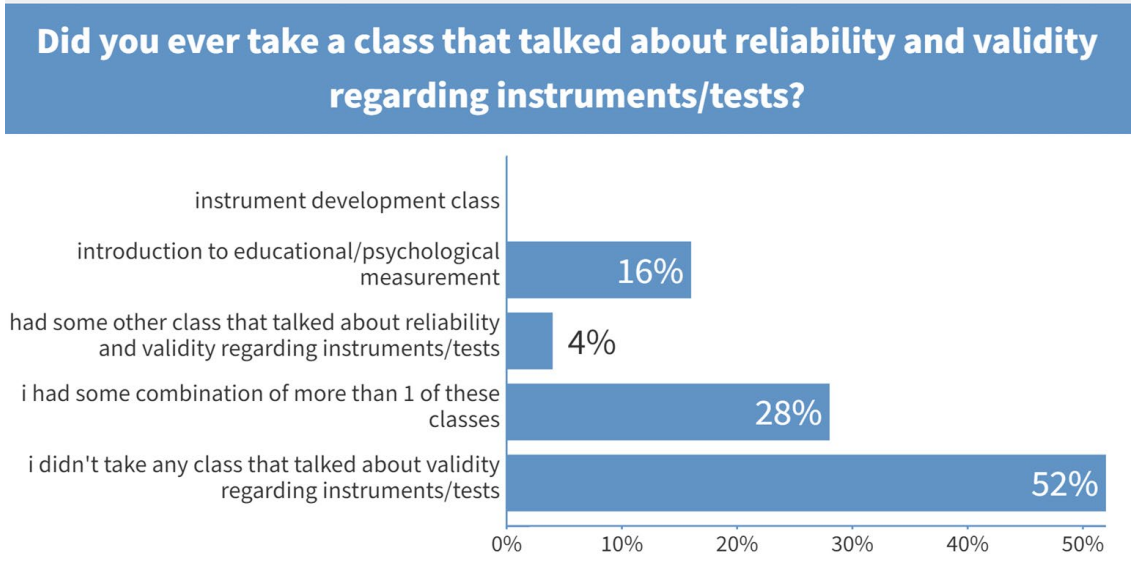


Source of figure: Anunciação & Portugal (2020)

Why talk about validity and validity evidence?

- It has long been recognized that applied measurement in social science research is widely misunderstood.
- For all of the advances in the measurement field, measurement theory is not regularly or appropriately incorporated into such research (i.e., Flake & Fried, 2020).
- Additionally, measurement training remains de-emphasized in graduate program curricula (Aiken et al., 2008; Childs & Eyde, 2002).

This problem exists within statistics education: during a USCOTS 2019 Breakout Session, Harrell-Williams and Whitaker recording the following participant responses during their 2019 USCOTS breakout session on validity evidence.



Validity Evidence for Measurement in Mathematics Education (V-M^{2}Ed)

- NSF Grant No. DRL #1920619 & #1920621
  - PIs: Erin Krupa (NC State University) & Jonathan Bostic (Bowling Green State University)
  - https://sites.ced.ncsu.edu/mathedmeasures/

- Projects:
  - Conferences on validity evidence (2017 & 2020)
  - Books about contemporary validation studies (e.g., Bostic et al., 2019a, 2019b)
  - Creation of a searchable database of instruments and validity evidence

Validity Evidence for Measurement in Mathematics Education (V-M^{2}Ed)

- NSF Grant No. DRL #1920619 & #1920621
  - PIs: Erin Krupa (NC State University) & Jonathan Bostic (Bowling Green State University)
  - https://sites.ced.ncsu.edu/mathedmeasures/

- Projects:
  - Conferences on validity evidence (2017 & 2020)
  - Books about contemporary validation studies (e.g., Bostic et al., 2019a, 2019b)
  - Creation of a searchable database of instruments and validity evidence

Validity Evidence for Measurement in Mathematics Education (V-M^{2}Ed)

Validity evidence is being documented through a structured literature review:

- **Round 1**: Instruments and tests to be included are identified.
- **Round 2**: Sources (e.g., papers and presentations) that *might* provide validity evidence for the instruments/tests are identified.
- **Round 3**: Specific validity claims and validity evidence are identified from the sources found in Round 2.

Validity Evidence for Measurement in Mathematics Education (V-M^{2}Ed)

Validity evidence is being documented through a structured literature review:

- **Round 1**: Instruments and tests to be included are identified.
- **Round 2**: Sources (e.g., papers and presentations) that *might* provide validity evidence for the instruments/tests are identified.
- **Round 3**: Specific validity claims and validity evidence are identified from the sources found in Round 2.

Synthesis groups focusing on different areas are conducting this structured literature review with a common framework:

- Elementary (K-6) Math
- Secondary (7-12) Math
- Undergrad/Grad Math
- Teacher Education Instruments
- Teacher Education Tests
- Statistics Education K-20

Validity Evidence for Measurement in Mathematics Education (V-M^{2}Ed)

Validity evidence is being documented through a structured literature review:

- **Round 1**: Instruments and tests to be included are identified.
- **Round 2**: Sources (e.g., papers and presentations) that *might* provide validity evidence for the instruments/tests are identified.
- **Round 3**: Specific validity claims and validity evidence are identified from the sources found in Round 2.

Intended completion: Fall 2022

This presentation

Synthesis groups focusing on different areas are conducting this structured literature review with a common framework:

- Elementary (K-6) Math
- Secondary (7-12) Math
- Undergrad/Grad Math
- Teacher Education Instruments
- Teacher Education Tests
- Statistics Education K-20

Current Stat. Ed. K-20 Synthesis Group Members:
Charlotte Bolch, Stephanie Casey, Christopher Engledowl, Leigh Harrell-Williams, Taylor Mulé, Justine Pointek, Hartono Tjoe, Douglas Whitaker

Summary of Round 1

- Initial goal: identify instruments developed since 2000 using:
  - Database searches
  - Focused searches of:
    - *Statistics Education Research Journal* (SERJ)
    - *Journal of Statistics and Data Science Education* (JSE/JSDSE)
    - Proceedings of the *International Conference on Teaching Statistics* (ICOTS)
- Exclusion criteria:
  - Instrument not in English
  - Instrument not statistics-specific

Summary of Round 2

- Structured searches to identify articles that seemed to use or be about the instruments from Round 1
  - If new instruments were found, they were included (no year limit)
- Articles classified based on:
  - Using the instrument (or not)
    - Population of use recorded
  - Seems to contain validity evidence (or not)
    - A detailed examination of validity evidence in each article will be in Round 3, so in Round 2 we erred on the side of including sources
- Instruments classified based on:
  - Instrument type
  - Item type

Summary of Round 2

- Currently, we have identified 111 instruments
- Of these:
  - 50 relate to student attitudes, beliefs, or perceptions (SA)
  - 45 relate to student knowledge (SK)
  - 16 relate to teachers (TCH)

- Many of these are seldom used; a few are very popular

- *Note: there will be procedures for updating the database to include new instruments and sources for validity evidence*

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

Population of use	Student Attitudes (SA)	Student Knowledge (SK)	Teacher (TCH)	Total
Elementary/Primary/K-6 Students	3	5		8
Secondary/7-12 Students	13	13		26
Undergraduate Students	37	34		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		6	11
Other	3	3	1	7

Note. Some instruments were used with multiple populations.
0s omitted for readability.

Table 2. Number of instruments of each instrument type.

Instrument Type	Student Attitudes (SA)	Student Knowledge (SK)	Teacher (TCH)	Total
Likert/Rating Scale	47	3	11	61
Summative		36	2	38
Survey		3	4	7
Diagnostic		6		6
Formative		7		7
Observation		2		2
Missing	1			1

Note. Some instruments were classified as having multiple types.
0s omitted for readability.

Table 3. Number of instruments using different item types.

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

Note. Some instruments included multiple item types.
0s omitted for readability.

Example Instruments

- Detailed information will be presented for three instruments:
  - **SA group**: *Survey of Attitudes Toward Statistics* (SATS) family (Schau, 1992, 2003)
  - **SK group**: *Levels of Conceptual Understanding in Statistics* (LOCUS) family (Jacobbe et al., 2014; Whitaker et al., 2015)
  - **TCH group**: *Self-Efficacy for Teaching Statistics* (SETS) family (Harrell-Williams et al., 2014a, 2014b)

- These instruments were chosen because they typified instruments that had many sources that were examined in Round 2.

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		282		2		6
Total	110	432	7	13	10	8

Note. Some sources may have used more than one instrument.
0s omitted for readability.

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.

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***. 0s omitted for readability.

	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						
Secondary/7-12 Students	1	4	***3***	***3***	1	
Undergraduate Students	***81***	***120***	1	3		
Graduate Students	5	10				
PSTs (Undergrad/MAT/etc.)	4	4			***9***	***2***
Elementary/Primary/K-6 Teachers	1	1				
Secondary/7-12 Teachers		2	1	4	2	
Tertiary Instructors						
Other (write in column to the right)	6	4	4	5		
Missing	16	10				

Example Instruments

- 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!

- “Validation is the joint responsibility of the [instrument] developer and [instrument] user” (AERA et al., 2014, p. 13).

Observations: Problematic Pattern 1

- There is interest in using instruments with populations beyond the one originally intended
  - Especially using student instruments with teachers
- Of the studies using instruments with new populations, some provide validity evidence…
  - … but many do not

- *Takeaway: users of instruments need to engage in providing validity evidence, too*

Observations: Problematic Pattern 2

- Many instruments have been developed to measure the same or similar constructs (e.g., statistics attitudes)

- A few become widely adopted
- Many see very limited use

- Before developing a new instrument, there should be a clear reason why a new instrument is needed!

V-M^{2}Ed Project: Next Steps

- The Statistics Education synthesis group is currently in *Round 3*.
- We are documenting the validity evidence claims and evidence for each instrument.

- At the end of Round 3:
  - We will be able to quantify the problematic patterns that we have observed.
  - The results will be added to the searchable database that is being developed.

Validity: What can ***you*** do?

- Adopt best practices when…
  - Developing new instruments (and ensure a new instrument is *needed*)
  - Using existing instruments (document validity evidence, especially when using in new ways)
- Resources:
  - "Measurement Schmeasurement" (Flake & Fried, 2020)
    - Provides an accessible overview of what they term *questionable measurement practices* (QMPs)
  - Special issue in *Applied Measurement in Education* (Vol. 32, No. 1)
    - Begins with an overview of different validity frameworks (Krupa et al., 2019)
  - V-M^{2}Ed books focus on examples of projects that seek to provide rigorous validity evidence (Bostic et al., 2019a, 2019b)

Validity: What can ***you*** do?

- Adopt best practices when…
  - Developing new instruments (and ensure a new instrument is *needed*)
  - Using existing instruments (document validity evidence, especially when using in new ways)
- Resources:
  - "Measurement Schmeasurement" (Flake & Fried, 2020)
    - Provides an accessible overview of what they term *questionable measurement practices* (QMPs)
  - Special issue in *Applied Measurement in Education* (Vol. 32, No. 1)
    - Begins with an overview of different validity frameworks (Krupa et al., 2019)
  - V-M^{2}Ed books focus on examples of projects that seek to provide rigorous validity evidence (Bostic et al., 2019a, 2019b)

References

- Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist*, *63*(1), 32–50. https://doi.org/10.1037/0003-066X.63.1.32
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association. https://www.testingstandards.net/open-access-files.html
- Anunciação, L., & Portugal, A. C. (2020). A Case Study on Strengthening the Link Between Psychometrics, Assessment, and Intervention in Autism Spectrum Disorder (ASD). In A. Singh, M. Viner, & C. J. Yeh (Eds.), *Special Education Design and Development Tools for School Rehabilitation Professionals:* IGI Global. https://doi.org/10.4018/978-1-7998-1431-3
- 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.
- Childs, R. A., & Eyde, L. D. (2002). Assessment Training in Clinical Psychology Doctoral Programs: What Should We Teach? What Do We Teach? *Journal of Personality Assessment*, *78*(1), 130–144. https://doi.org/10.1207/S15327752JPA7801_08
- 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., 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. *Sustainability in Statistics Education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA*. http://icots.info/icots/9/proceedings/pdfs/ICOTS9_C270_HARRELLWILLIAMS.pdf
- 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
- Harrell-Williams, L., & Whitaker, D. (2019, May). *Evaluating validity evidence for instruments in statistics education*. United States Conference on Teaching Statistics 2019, State College, PA.
- 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.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA*. International Statistical Institute. http://icots.net/9/proceedings/pdfs/ICOTS9_7C2_JACOBBE.pdf
- 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. https://doi.org/10.1080/08957347.2018.1544139
- 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/
- Whitaker, D., Foti, S., & Jacobbe, T. (2015). The Levels of Conceptual Understanding in Statistics project: Results of the pilot study. *Numeracy*, *8*(2), Article 4. https://scholarcommons.usf.edu/numeracy/vol8/iss2/art3/

Selected Contact Information

Name	Institution	Email
Presenter		
Douglas Whitaker	Mount Saint Vincent University	douglas.whitaker@msvu.ca
Statistics Education K-20 Synthesis Group Leads		
Stephanie Casey	Eastern Michigan University	scasey1@emich.edu
Leigh Harrell-Williams	The University of Memphis	Leigh.Williams@memphis.edu
V-M^{2}Ed Project Leads		
Jonathan Bostic	Bowling Green State University	bosticj@bgsu.edu
Erin Krupa	North Carolina State University	eekrupa@ncsu.edu

The search for validity evidence for instruments in statistics education: preliminary findings

IASE Satellite Conference

2 September 2021

Questions?

Thank you!

Douglas Whitaker
Mount Saint Vincent University

Charlotte Bolch
Midwestern University

Leigh Harrell-Williams
University of Memphis

Stephanie Casey
Eastern Michigan University

Corinne Huggins-Manley
University of Florida

Christopher Engledowl
New Mexico State University

Hartono Tjoe
Pennsylvania State University

Supplemental Slides

Abstract

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 both 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.

USCOTS 2021 Poster Survey

- At our poster at USCOTS 2021, we made a survey available to attendees asking about their background with validity as part of the participant engagement focus of the conference.
  - The link to the survey was shared in the conference Slack channel prior to the live poster session.
- The next few slides present selected results from this survey.
- Note that that the data is from a *convenience sample of people who came to a presentation about validity*.
  - The results are certainly not broadly generalizable...
  - ... but the results may still be of interest given that the respondents showed an *interest in validity*.
    - (We suspect people without an interest in validity would have less of a background in validity.)
- (This is similar to the survey results on Slide 3 of this presentation from the USCOTS 2019 breakout session.)

USCOTS 2021 Poster Survey

- **Q4 - Have you ever taken a course that addressed reliability and validity regarding tests/instruments?**

Answer	Count	%
Yes	7	58%
No	5	42%
Total	12	100%

USCOTS 2021 Poster Survey

- **Q5 - How have you learned about validity evidence? (select all that apply)**

Answer	Count	%
Graduate coursework	6	50%
Sessions at conferences	4	33%
Professional development workshop	1	8%
Reading measurement books/articles/etc.	7	58%
Other: (explain)	1	8%

"Working with colleagues in a research group."

USCOTS 2021 Poster Survey

- **Q7 - How do you decide if an instrument is appropriate for use in your work? (select all that apply)**

Answer	Count	%
Other people that I am citing have used the instrument.	5	42%
I read the instrument development and validation paper(s).	10	83%
I consult with other Stat Ed educators/researchers.	8	67%
I consult with the instrument developer.	5	42%
Other: (explain)	1	8%

"aligns with my purpose"