Applied Measurement in Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hame20

Think Alouds: Informing Scholarship and Broadening Partnerships through Assessment

Jonathan David Bostic

To cite this article: Jonathan David Bostic (2021) Think Alouds: Informing Scholarship and Broadening Partnerships through Assessment, Applied Measurement in Education, 34:1, 1-9, DOI: 10.1080/08957347.2020.1835914

To link to this article: https://doi.org/10.1080/08957347.2020.1835914

Published online: 01 Feb 2021.

APPLIED MEASUREMENT IN EDUCATION
2021, VOL. 34, NO. 1, 1–9
https://doi.org/10.1080/08957347.2020.1835914

Routledge
Taylor & Francis Group

Think Alouds: Informing Scholarship and Broadening Partnerships through Assessment

Jonathan David Bostic

School of Teaching & Learning, Bowling Green State University College of Education and Human Development

ABSTRACT

Think alouds are valuable tools for academicians, test developers, and practitioners as they provide a unique window into a respondent's thinking during an assessment. The purpose of this special issue is to highlight novel ways to use think alouds as a means to gather evidence about respondents' thinking. An intended outcome from this special issue is that readers may better understand think alouds and feel better equipped to use them in practical and research settings.

This special issue focuses on ways to use think alouds as a means to gather evidence about respondents' thinking, including the ways they engage in problem-solving. Three articles presented in this special issue draw upon assessment research from science, technology, engineering, and mathematics (STEM) contexts. This special issue may inform academicians and practitioners' assessment work, especially within STEM education. Assessment is essentially *idea mining* and think alouds are useful tools to extract respondents' problem solving (Leighton, 2004, 2017). Thus, readers may deduce ideas from the manuscripts that help (a) develop, utilize, and revise instruments for educational research and (b) investigate respondents' thinking while problem-solving. Problem-solving is woven across academic standards in STEM education (Common Core State Standards Initiative [CCSSI], 2010; Computer Science Teacher Association, 2017; National Research Council, 2011; Next Generation Science Standards, 2013). Hence, researchers and practitioners need effective tools to investigate problem-solving within STEM education contexts. Thinks alouds help researchers and practitioners seek to investigate respondents' cognitive behaviors within settings that require problem-solving and critical thinking (Leighton, 2017). They also have the potential to support gathering response process validity evidence as part of assessment development or revision (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Evidence gathered from think alouds can inform developing validity arguments about appropriate uses of assessments and interpretations from their results. Well-formed validity arguments communicate necessary information about an instrument and how it ought to be used (Kane, 2016). The next section of this introduction begins with some foundational ideas about validity, response processes, and think alouds.

1. Key Ideas about Response Process Evidence and Think Alouds

Validity is central to conducting high-quality research (AERA et al., 2014; Hill & Shih, 2009; Kane, 2016). It can strengthen one's confidence in measuring a latent construct (Bostic, 2017; Nielsen & Bostic, 2020). The measurement process operationalizes the means by which a construct is connected to a particular theoretical framework (Cronbach & Meel, 1955), which is tied to the strength and

CONTACT Jonathan David Bostic bosticj@bgsu.edu School of Teaching & Learning, Bowling Green State University College of Education and Human Development, Bowling Green, OH 43403-0001

source of validity evidence (Kane, 2016). Robust validity evidence supports a validity argument, which characterizes how effectively and appropriately the measurement process may have accounted for contexts, settings, environmental factors, or potential sources of error (Kane, 2016; Lavery, Jong, Krupa, & Bostic, 2019). Thus, validity should be an everpresent tenet in the eyes of the assessment designer and user. The *Standards for Educational and Psychological Testing* (AERA et al., 2014) characterize five sources of validity evidence: test content, response process, relations to other variables, internal structure, and consequences from testing/bias. This special issue focuses on think alouds, which are commonly associated with gathering response process evidence.

Response processes are the cognitive actions that individuals employ to think about and answer assessment items (AERA et al., 2014; Hubley & Zumbo, 2017; Leighton, 2017). A goal of gathering response process data is to understand a respondent's engagement with a prompt. Those data help to draw conclusions about an individual's problem solving or an ability to produce a desired response (AERA et al., 2014; Padilla & Benítez, 2014). Leighton (2017) provides a useful description of think alouds, which I briefly summarize in the next paragraph.

A think aloud is a one-on-one interview between an assessor and respondent in a quiet setting, typically outside of a classroom setting (Leighton, 2004, 2017; Schindler & Lilienthal, 2019). Tasks that require problem-solving or critical thinking are usually involved in think alouds (Leighton, 2017). An assessor asks the respondent to share any ideas that come to mind during the interview and usually encourages the respondent to also show written work. The interviews are video and audio recorded and data may be analyzed quantitatively and/or qualitatively. Responses elicited during a think aloud may indicate what stimuli in the task an respondent notices, and how the respondent attends to those stimuli (Padilla & Benítez, 2014). Think alouds with problem-solving tasks, like those used within studies found in this special issue, can reveal the degree to which the prompt, within the flow of all prompts in a think-aloud protocol, connects the respondent's behavior(s) with an intended outcome (Leighton, 2017; Padilla & Benítez, 2014). Problem-solving requires critical, complex thinking and is multi-faceted (Mayer & Wittrock, 2006). Therefore, a tool like think alouds that gives an assessor a window into a respondent's cognition is central to research on problem-solving. Think alouds – as used in this special issue – can be paired with quantitative and qualitative analyses as a way to conduct comparisons across scales that focus on a desired construct. Much of what is revealed in studies within this special issue can be applied to support other methods investigating response processes and more broadly, respondents' problem-solving.

2. Two Studies Leveraging Think-Aloud Data within STEM Education

Think alouds are tools for individuals seeking to better understand a respondent's thinking in relation to a complex task, like problem-solving in STEM contexts. Bonner and colleagues (this issue) use think alouds as a means to understand respondents' problem-solving processes on a computer science-based formative assessment of computational thinking. The five computational thinking practices are abstraction, sequencing, iteration, testing, and debugging (Shute, Sun, & Asbell-Clarke, 2017). Eleven students enrolled in an Advanced Placement course focused on Principles of Computer Science completed well-developed formative assessment tasks as part of the present study. These tasks were connected to Zimmerman's (2000) three-phase, self-regulated learning framework: (a) engagement in forethought, (b) task completion, and (c) reflection. Think alouds were implemented during the formative assessment to understand how participants drew upon computational thinking practices while problem-solving. Quantitative and qualitative evidence from those think alouds were intended to illuminate aspects of self-regulated learning during their problem-solving tasks. A finding from their study was that think alouds, used in conjunction with qualitative artifact coding, highlighted a link between students' cognition, self-regulated learning, and task performance. Prior work using a single methodological approach had not connected findings as Bonner and colleagues' work did. Their research provides substantive methodological evidence for academicians and practitioners to consider when examining how think alouds might be used to further explore self-regulated learning

and computational thinking. An implication from their study is that think alouds used in conjunction with other investigative tools help construct a more robust narrative about an individual's problem solving as compared to a single methodological approach. Notably, drawing upon multiple data collection approaches, including implementing different forms of quantitative or qualitative data collection as well as utilizing quantitative and qualitative tools, better grounds evidence from an operationalized latent construct to a chosen theoretical perspective (Howe, 2012).

In the second article of this special issue, Mo and coauthors use think alouds as tools to compare data from constructed response (CR) and selected response (SR) items addressing mathematics teacher attentiveness. To accomplish this comparison, the authors used CR items developed as part of a prior study and engaged in think alouds with respondents who were presented with SR items constructed to be parallel in nature. CR items from the Disciplinary Attentiveness to Student Ideas instrument (Carney, Cavey, & Hughes, 2017) were revised so that similar SR items were created. The researchers purposefully chose 11 teachers in total (preservice and inservice) who represented various years of experience as well as enrollment in undergraduate or graduate-level coursework. Think-aloud invstigations provided a window into participants' problem-solving processes on both CR and SR items. Data were analyzed both quantitatively. An important finding from this study is that SR items, when carefully developed from robust CR items as done in their study, led to analogous results. An implication from this study was that CR and SR items focused on mathematics teacher attentiveness can be constructed to be similar in nature. Their work fills a needed gap in literature about item development. The authors hypothesize that utilizing a CR to SR question development process as done in their study with other STEM-related contexts has the potential to generate findings that are much the same; however, further study is warranted.

Collectively, these studies convey new information about how think alouds may be used in educational research. They highlight means to (a) gather response process evidence related to an intended construct and (b) illuminate phenomenological connections that may not have been readily observed through a single methodological approach. Researchers in these studies drew logical inferences from think alouds with small samples of purposefully selected participants that might be further taken up in studies with larger samples. Think alouds, as used in these studies, offer scholars and practitioners a valuable opportunity to observe a live phenomenon through a participant's verbal and written statements and in turn, make inferences about respondents' cognition. Thus, think alouds have potential as tools for researchers seeking to scale up or generalize research findings.

3. Think Alouds: Novel Uses and New Techniques

While the previous two studies in this special issue showcase ways to use think-aloud data in novel ways, Bostic and colleagues (this issue) describe a new think-aloud technique called whole-class think alouds (WCTAs). They suggest that K-12 teachers might work alongside researchers through WCTAs and both benefit from the partnership. One goal of the present study was to understand third-grade students' mathematical problem-solving. Items were selected from a problem-solving measure series connected to grade-level content standards (see Bostic, Matney, Sondergeld, & Stone, 2020; Bostic & Sondergeld, 2015; Bostic, Sondergeld, Folger, & Kruse, 2017 for more information). The second goal of this project was to develop a different approach to gathering response process data. The research team wondered if gathering data differently might provide confirmatory or discriminant validity evidence about students' mathematical problem-solving. This wondering led to the development of WCTAs, which happened as part of normal, everyday instruction as a formative assessment.

Teachers placed third-grade students into pairs based upon mathematics ability level and past experience working together. Videorecorders were placed strategically to capture student discourse in much the same way they are used during 1–1 think alouds. WCTA data were compared to data from 1–1 think alouds with a group of similar respondents, which had been collected during a prior study. Data were initially coded to compare students' performance and strategy use across 1–1 think alouds and WCTAs. A secondary analysis was performed to examine students' perceptions of WCTAs. One

finding from this study was that there were similar results about students' performance and strategy use when 1–1 think aloud and WCTA data were compared. A second finding was that students felt comfortable in WCTAs because they were engaged by a teacher and researchers. This level of comfort had not been previously reported by students who participated in 1–1 think alouds. Teachers reported that students' work during WCTAs helped to inform future mathematics instruction.

An implication from this research is that WCTAs are a new technique for gathering response process data that do not require removing K-12 students from their instructional environment. A second implication is that those working in K-12 settings may benefit from using WCTAs as part of their formative assessment tools, like the teachers in this study. Think alouds, as formative assessments, can elucidate data about students' thinking on complex problem-solving tasks for K-12 teachers or teacher educators (Fennell, Kobett, & Wray, 2017; Nielsen & Bostic, 2020). WCTAs, as used in the present study, included practitioners in conversations about ways to connect assessment developers and assessment users. Such collaboration between academicians and practitioners fosters positive school–university partnerships as well as improvements in instrument development and administration for K-12 settings.

4. Connecting Validity Evidence to a Validity Argument

The *Standards for Educational and Psychological Testing* (AERA et al., 2014) communicate that response process evidence is important evidence as part of a broader validity argument associated with a test. However, response process *implications* often focus on only one aspect drawn from response process *evidence*. As an example of, "the construct of interest is only concerned with whether the problem was solved correctly. … there may be multiple possible routes to obtaining the correct solution to a mathematical problem" (AERA et al., 2014, p. 16). Those engaged in STEM education research may be interested in students' strategy use while problem-solving as well as their problem-solving performance. Varied research agendas drive data collection decisions. These decisions about data collection inform what and how to collect response process data through tools like think alouds. Ultimately, these data collection choices about what tools to use and how to use them to collect validity data influence the framing of the validity arguments.

High-quality quantitative research can only be driven by rigorously developed instrumentation, which involves reflecting on validity and validity data collection (Hill & Shih, 2009). Recently, some authors have offered advice for creating validity arguments and how developers and users might leverage response process evidence and think alouds as part of the assessment design and use process (e.g., Lavery et al., 2019; Leighton, 2017). Typically, an argument has a claim, evidence, and justification (Kane, 2006, 2016; Krupa, Carney, & Bostic, 2019). This claim is supported with confirming evidence and delimited with divergent (discriminant) evidence. The ways in which an assessment developer connects the evidence and claim through justification implicate the coherence of the validity argument. There are many perspectives for a validity argument as evident by the wide range of authors who have contributed on this topic (e.g., Kane, 2006, 2016; Mislevy, Almond, & Lukas, 2003; Pellegrino, DiBello, & Goldman, 2016; Schilling & Hill, 2007; Wilson, 2005; Wilson & Wilmot, 2019). In addition, some assessment developers use the *Standards* (AERA et al., 2014) as a framework by which claims, evidence, and justification are connected (e.g., Bostic, Matney, & Sondergeld, 2019; Bostic et al., 2017; Gleason, Livers, & Zelkowski, 2017; Walkowiak, Adams, & Berry, 2019; Wilhelm & Berebitsky, 2019). Validity arguments communicate the connections between the claims and evidence in a way that includes a "coherent series of reasons, statements, or facts intended to establish a point of view" (Merriam-Webster, 2018). Assessment experts are necessary when working alongside content experts to develop an instrument and construct its validity argument. Combined expertise brings a valuable perspective to investigations into understanding phenomena through think alouds and interpreting those think-aloud data.

5. The Value of Multiple Perspectives and Collaboration

As seen in studies found in this special issue, assessment experts worked in tandem with content experts to analyze think-aloud data and draw inferences about respondents' thinking. This sort of collaborative work connects scholars across various knowledge areas in the aim of having intellectual merit and broader impact. This section describes how expertise from assessment and content experts can lead to meaningful work, and this collaboration may influence the approaches and analyses of think-aloud data.

Leighton's commentary for this special issue effectively notes that scholars bring unique perspectives and expertise to their research projects. Expertise can be both a strength and a challenge. One example from mathematics education research helps to illustrate this point. Mathematical expertise is often viewed as using symbolic representations while problem-solving effectively because it is precise, concise, effective, and efficient (Herman, 2007). Symbolic strategies include using abstract representations like expressions, equations, and inequalities (Goldin & Kaput, 1996). However, using a symbolic strategy is many times, less efficient than an approach with nonsymbolic representations such as tables or pictures (Santos-Trigo, 1996; Yee & Bostic, 2014). Research with think alouds has indicated that K-12 students with mathematical problem-solving expertise demonstrated flexibility with representations while problem-solving (Yee & Bostic, 2014). Expertise in mathematical problem solving was reified through strategic flexibility more so than facility with a single representation during problem-solving. Similarly, scholars designing or modifying instruments bring to bear their own perspectives on learning. Fluidity with multiple perspectives, as opposed to a single view, has the potential to draw upon expertise and influence think aloud interpretations. Multiple studies on learning have used think alouds, and future research which might be strengthened by greater inclusion of multiple theoretical perspectives or knowledge bases with think aloud data collection.

Bringing more than one perspective to bear upon validity and validation work has the potential to influence future scholarship. These perspectives have the power to illuminate ideas about response process evidence. Additionally, drawing upon numerous perspectives has capacity to bring more individuals into assessment design and use scholarship and practices. As an example, two National Science Foundation-funded projects have sought to convene scholars in assessment, psychometrics, mathematics and statistics education, and education policy (Validity and Measurement in Mathematics Education; NSF #1644314; #192019 and #1920621). It was clear during the 2017 face-to-face meeting how individuals from different scholarly backgrounds brought unique perspectives to conversations around validity and validation. One discussion point among the group of assessment and psychometrics experts and the group of mathematics and statistics educators was that the former group had robust methodological expertise while the latter group had disciplinary expertise. Neither group expressed having the same level of expertise in both areas so a shared conclusion from both groups was a need for collaboration on projects like assessment validation and research with think alouds. Think alouds were one specifically named area such that individuals in both groups felt their combined expertise was valuable to understand a phenomenon. A second named area was an agreed-upon aim to improve assessments and measurement practices. That is, working collaboratively on a common idea strengthened, not limited, scholarship. Several conference participants expressed having conducted think alouds in past work but few had actually paused to consider studying think alouds as tools. An outcome from this NSF project was a burgeoning community of individuals from various backgrounds who sought to collaborate purposefully around think alouds and validation. A follow-up National Science Foundation-funded meeting occurred in February 2020 and led to further discussions with some new members as well previous attendees with a goal of enhancing scholarship around measurement of mathematics education phenomena. The 2017 and 2020 meetings, as well as National Council on Measurement in Education (NCME) meetings on classroom assessment all point to ways that collaboration can strengthen educational research. This special issue is a direct product of shared work at one NCME classroom assessment session focused on think alouds, and each article includes a collaboration between assessment and content experts.

6. Think Alouds Informing Theory: Case of Cognitively Guided Instruction

Think alouds are one resource available to assessment scholars, content or equity experts, as well as qualitative and mixed-methods researchers seeking to understand phenomena. In addition, school district personnel might be in a position to conduct and analyze data in conjunction with a research-based assessment team. Cognitively guided instruction (CGI) is a historical example from mathematics education literature that illustrates work across scholars with a shared aim, as well as ways that think alouds might be used to inform, that is, develop and later shape, theory.

Early work (e.g., Carpenter & Moser, 1984) and follow-up research (e.g., Carpenter, Fennema, & Franke, 1996) used think alouds as a way to investigate how students develop knowledge of mathematics operations. Results from using think alouds in the early research were to develop a theory about young children's mathematics development. A goal of CGI is to understand children's thinking and addressing that goal has the power to inform instruction and research (Carpenter & Franke, 2000). Think alouds in CGI research were conducted with a goal to impact both theory and practice. CGI as a theory began to gain greater influence within research literature and teachers' instruction during the 1980s. Subsequent studies began to reexamine the limitations, inferences, and possible delimitations of CGI theory. Thus, think alouds became tools to use after the initial theory development research took place. One example of reexamination that led to shaping theory included conducting think alouds with participants who were English Language Learners, which had not been done before. A result from this work was that CGI results did not necessarily translate to Spanish-speaking students. The nuances of English and Spanish languages played a pivotal role in students' mathematical development (Carpenter & Franke, 2000). Thus, CGI results were reexamined in light of the new evidence and new studies were conducted to further shape CGI theory. This example highlights the need for think alouds to be conducted as *a priori* tools of investigation as well as *a posteriori* tools. That is, think alouds can inform initial theory development as well as subsequent theory revisions. A second implication from this example is that think-aloud techniques, when paired with various participant sampling techniques, can lead to different conclusions. Participation in think alouds impacts the degree to which equity and bias – as sources of validity – are taken up critically. It is important that researchers carefully consider who participates in think alouds and the reason for their selection (Leighton, 2004, 2017).

7. Final Thoughts

This special issue includes three unique studies about think alouds as well as a commentary on these studies. Its purpose is to focus on ways to use think alouds to gather evidence about respondents' thinking, including the ways respondents engage in problem-solving and critical thinking. An outcome from this special issue is to provide ideas about ways think alouds might be used as research tools, and how their results inform scholarship across education contexts, including assessment and measurement. Think alouds, as seen in these studies, were used in conjunction with other data collection techniques to support multi-modal arguments. Such cross-collaborative methodological studies inform general knowledge bases.

Think-aloud findings, as discussed in this special issue, may perturb scholars to consider questions about their results about other phenomena. A few questions are provided as a way to jumpstart further work with think alouds, particularly those working in STEM education contexts: What STEM phenomena needs greater exploration with think alouds? How might think-aloud findings that informed a priori hypotheses, influence studies with posteriori hypotheses? How might findings from think alouds with a different sampling technique, like WCTAs, inform theory development revision, or practical applications of theory? Scholars and practitioners drawing upon multiple theoretical perspectives and using their combined skill sets are in a position to use think alouds as part of robust education research agendas.

The National Research Council (2002) has described six related principles of inquiry for scientific research in education. One of those principles is to "use methods that permit direct investigation of the question" (p. 52). Think alouds offer a rare moment through which an observer might see and hear a respondent's thinking in relation to an intended construct, like problem-solving or critical thinking. Concomitantly, "psychological and educational tests influence who gets what in society, fresh challenges follows shifts in social power or social philosophy. *So validation [work] is never finished* (emphasis from original)" (Cronbach, 1988, p. 5). Think alouds are frequently part of assessment development or validation work. Hence, work with think alouds offers academicians, test developers, and practitioners multiple opportunities to engage in assessment research because new societal issues influence how test results are used. When those assessments are used to inform research or make decisions that impact individuals, especially those that have high-stakes outcomes, then the power of and need for think alouds as tools to understand phenomena becomes even more critical.

Acknowledgments

The author is grateful to the *Applied Measurement in Education* editorial team for their feedback on this manuscript, especially Kristen Huff for her encouragement related to this manuscript. Finally, this manuscript benefitted from critical feedback provided by Jim Middleton, Jeff Shih, Finbarr Sloane, and Toni Sondergeld.

Disclosure Statement

No potential conflict of interest was reported by the author.

Funding

Ideas in this manuscript stem from grant-funded research by the National Science Foundation (NSF #1644214, #1720646; #1720661; #1920619, and #1920621). Any opinions, findings, conclusions, or recommendations expressed by the authors do not necessarily reflect the views of the National Science Foundation.

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.

Argument. (2018). *Merriam-webster.com*. Retrieved from https://www.merriam-webster.com/dictionary/argument

Bostic, J. (2017). Moving forward: Instruments and opportunities for aligning current practices with testing standards. *Investigations in Mathematics Learning, 9*(3), 109–110.

Bostic, J., Sondergeld, T., Folger, T., & Kruse, L. (2017). PSM7 and PSM8: Validating two problem-solving measures. *Journal of Applied Measurement, 18*(2), 151–162.

Bostic, J., & Sondergeld, T. (2015). Measuring sixth-grade students' problem solving: Validating an instrument addressing the mathematics Common Core. *School Science and Mathematics Journal, 115*, 281–291.

Bostic, J., Matney, G., & Sondergeld, T. (2019). A lens on teachers' promotion of the standards for mathematical practice. *Investigations in Mathematics Learning, 11*(1), 69–82. doi:10.1080/19477503.2017.1379894

Bostic, J., Matney, G., Sondergeld, T., & Stone, G. (2020, March). Measuring what we intend: A validation argument for the grade 5 problem-solving measure (PSM5). Validation: A Burgeoning Methodology for Mathematics Education Scholarship. In J. Cribbs & H. Marchionda (Eds.), *Proceedings of the 47th annual meeting of the research council on mathematics learning* (pp. 59–66). Las Vegas, NV.

Carney, M., Cavey, L., & Hughes, G. (2017). Assessing teacher attentiveness: Validity claims and evidence. *Elementary School Journal*, *118*(2), 281–309. doi:10.1086/694269

Carpenter, T., & Franke, M. (2000). Cognitively guided instruction: Challenging the core of educational practice. In T. Glennan jr, S. Bodilly, J. Galegher, & K. Kerr (Eds.), *Perspectives from leaders in the scale up of educational interventions* (pp. 41–77). Santa Monica, CA: RAND.

Carpenter, T., Fennema, E., & Franke, M. (1996). Cognitively guided instruction: A knowledge base for reform in primary mathematics instruction. *The Elementary School Journal*, *97*(1), 3–20. doi:10.1086/461846

Carpenter, T., & Moser, J. (1984). The acquisition of addition and subtraction concepts in grades one through three. *Journal for Research in Mathematics Education*, *15*(3), 179–202. doi:10.5951/jresematheduc.15.3.0179

Common Core State Standards Initiative. (2010). *Common core standards for mathematics*. Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf
Computer Science Teacher Association. (2017). *CS standards*. CSTA K-12 Computer Science Standards. Retrieved from https://www.csteachers.org/page/standards
Cronbach, L. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–18). Hillsdale, NJ: Erlbaum.
Cronbach, L., & Meel, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. doi:10.1037/h0040957
Fennell, F., Kobett, B., & Wray, J. (2017). *The formative 5: Everyday assessment techniques for every math classroom*. Thousand Oaks, CA: Corwin.
Gleason, J., Livers, S., & Zelkowski, J. (2017). Mathematics Classroom Observation Protocol for Practices (MCOP2): A validation study. *Investigations in Mathematics Learning*, *9*(3), 111–129. doi:10.1080/19477503.2017.1308697
Goldin, G., & Kaput, J. (1996). A joint perspective on the idea of representation in learning and doing mathematics. In L. Steffe & P. Nesher (Eds.), *Theories of mathematical learning* (pp. 397–431). Mahwah, NJ: Erlbaum.
Herman, M. (2007). What students choose to do and have to say about multiple representations in college algebra. *Journal of Computers in Mathematics and Science Teaching*, *26*(1), 27–54.
Hill, H., & Shih, J. (2009). Examining the quality of statistical mathematics education research. *Journal of Research in Mathematics Education*, *40*(3), 241–250.
Howe, L. R. (2012). Mixed methods, triangulation, and causal explanation. *Journal of Mixed Methods*, *6*(2), 89–96.
Hubley, A. M., & Zumbo, B. D. (2017). Response processes in the context of validity: Setting the stage. In B. Zumbo & A. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 1–12). Cham, Switzerlands: Springer.
Kane, M. T. (2006). Validation. In R. L. Brennan & National Council on Measurement in Education & American Council on Education (Ed.), *Educational measurement* (pp. 17–64). Westport, CT: Praeger Publishers.
Kane, M. T. (2016). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In S. Lane, M. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development*. (Vol. 2nd ed., pp. 64–80). New York, NY: Routledge.
Krupa, E., Carney, M., & Bostic, J. (2019). Approaches to instrument validation. *Applied Measurement in Education, 32* (1), 1–9.
Lavery, M., Jong, C., Krupa, E., & Bostic, J. (2019). Developing an assessment with validity in mind. In J. Bostic, E. Krupa, & J. Shih (Eds.), *Assessment in mathematics education contexts: Theoretical frameworks and new directions* (pp. 12–39). New York, NY: Routledge.
Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, *23*(4), 6–15. doi:10.1111/j.1745-3992.2004.tb00164.x
Leighton, J. P. (2017). *Using think-aloud interviews and cognitive labs in educational research*. New York, NY: Oxford University Press.
Mayer, R., & Wittrock, M. (2006). Problem solving. In P. Alexander & P. Winne (Eds.), *Handbook of educational psychology* (pp. pp. 287–303). Mahwah, NJ: Erlbaum.
Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i–29. doi:10.1002/j.2333-8504.2003.tb01908.x
National Research Council. (2002). *Scientific research in education*. Washington, DC: National Academies Press. doi:10.17226/9853
National Research Council. (2011). *Successful K-12 STEM education: Identifying effective approaches in science, technology, engineering, and mathematics*. Washington, DC: National Academies Press.
Next Generation Science Standards Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: National Academies Press.
Nielsen, M., & Bostic, J. (2020). Informing programmatic-level conversations on mathematics preservice teachers' problem-solving performance and experiences. *Mathematics Teacher Education and Development, 22*(1), 33–47.
Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136–144.
Pellegrino, J., DiBello, L., & Goldman, S. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist*, *51*(1), 59–81. doi:10.1080/00461520.2016.1145550
Santos-Trigo, M. (1996). An exploration of strategies used by students to solve problems with multiple ways of solution. *Journal of Mathematical Behavior*, *15*, 263–284.
Schilling, S. G., & Hill, H. C. (2007). Assessing measures of mathematical knowledge for teaching: A validity argument approach. *Measurement: Interdisciplinary Research and Perspectives*, *5*(2–3), 70–80.
Schindler, M., & Lilienthal, A. J. (2019). Domain-specific interpretation of eye tracking data: Towards a refined use of the eye-mind hypothesis for the field of geometry. *Educational Studies in Mathematics*, *101*, 123–139. doi:10.1007/s10649-019-9878-z
Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review*, *22*, 142–158. doi:10.1016/j.edurev.2017.09.003

Walkowiak, T. A., Adams, E. L., & Berry, R. Q. (2019). *Quantitative measures of mathematical knowledge: Researching instruments and perspectives* (E. K. Bostic, & J. Shih, Eds., pp. 90–119). New York, NY: Routledge.
Wilhelm, A. G., & Berebitsky, D. (2019). Validation of the mathematics teachers' sense of efficacy scale. *Investigations in Mathematics Learning*, *11*(1), 29–43. doi:10.1080/19477503.2017.1375359
Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Routledge.
Wilson, M., & Wilmot, D. B. (2019). *Quantitative measures of mathematical knowledge: Researching instruments and perspectives* (J. Bostic, E. Krupa, & J. Shihpp, Eds., pp. 63–89). New York, NY: Routledge.
Yee, S., & Bostic, J. (2014). Developing a contextualization of students' mathematical problem solving. *Journal for Mathematical Behavior, 36*, 1–19.
Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, *25*(1), 82–91. https://doi.org/10.1006/ceps.1999.1016