APPLIED MEASUREMENT IN EDUCATION
2021, VOL. 34, NO. 1, 10–26
https://doi.org/10.1080/08957347.2020.1835910

Routledge
Taylor & Francis Group

Using Think-Alouds for Response Process Evidence of Teacher Attentiveness

Ya Mo[a], Michele Carney[a], Laurie Cavey[b], and Tatia Totorica[c]

[a]Department of Curriculum, Instruction, and Foundational Studies, Boise State University, Boise, USA; [b]Department of Mathematics, Boise State University, Boise, USA; [c]IDoTeach, Boise State University, Boise, USA

ABSTRACT

There is a need for assessment items that assess complex constructs but can also be efficiently scored for evaluation of teacher education programs. In an effort to measure the construct of teacher attentiveness in an efficient and scalable manner, we are using exemplar responses elicited by constructed-response item prompts to develop selected-response assessment items. Through analyses of think-aloud interview data, this study examines the alignment between participant responses to, and scores arising from, the two item types. The interview protocol was administered to 12 mathematics teachers and teacher candidates who were first presented a constructed-response version of an item followed by the selected-response version of the same item stem. Our analyses focus on the alignment between responses and scores for eight item stems across the two item types and the identification of items in need of modification. The results have the potential to influence the way test developers generate and use response process evidence to support or refute the assumptions inherent in a particular score interpretation and use.

1. Introduction

Teacher education programs need ways to efficiently evaluate the effectiveness of program activities and innovations (Bell, Gitomer, Savage, & Mckenna, 2019). While there have been efforts to optimize the design of teacher preparation programs (e.g. Grossman, Hammerness, & McDonald, 2009), research at the program level is needed to support claims about outcomes (Bell et al., 2019). In particular, program leaders need access to instruments that efficiently and meaningfully assess complex constructs associated with effective teaching practices. Current guidelines for test development suggest an iterative process involving operationalization of the construct through the creation of items intended to elicit aspects of the intended construct, and repeated collection and examination of validity evidence to inform their work (American Educational Research Association (AERA), American Psychological Association (APA), & National Council of Measurement in Education (NCME), 2014).

Our project centers upon developing instruments to measure teacher attentiveness–*the ability to analyze and respond to a particular student's mathematical ideas from a progressive formalization perspective*. Our item development process (Carney, Cavey, & Hughes, 2017) involves the analysis of data elicited by constructed-response items and the curation of exemplar responses for use in selected-response items to target a particular content area and aspect of attentiveness (e.g., the inferences that can be made about a student's understanding based on their work on a mathematical task). In this

CONTACT Ya Mo yamo@boisestate.edu College of Education, Boise State University, 1910 University Drive, Boise, ID 83725-1745.

paper, we examine response process validity evidence related to the functioning of selected-response items developed as part of an instrument with a focus on quantitative reasoning. By drawing upon data from think-aloud interviews, we examine the alignment between respondents' scores on the constructed-response and selected-response versions for eight pairs of items.

2. Literature Review

2.1. Assessment of Attentiveness for Teacher Preparation

Attentiveness is a complex construct that combines aspects of professional noticing (Jacobs, Lamb, & Philipp, 2010), mathematical knowledge for teaching (Ball, Thames, & Phelps, 2008; Shulman, 1987), and progressive formalization, an instructional process that affords students' reinvention of formal mathematics by building upon their informal, intuitive, and concrete problem solving strategies (Freudenthal, 1973; Gravemeijer & van Galen, 2003; Treffers, 1987). Attentiveness contributes to the student-centered pedagogy called for by national stakeholders in mathematics education (National Council of Teachers of Mathematics (NCTM), 2014; Teaching Works, 2018). Though similar to professional noticing in that teachers must decide where to direct their attention and instructional efforts, the narrower focus of attentiveness bounds the construct to one-on-one interactions between a student, the evidence they produce, and the teacher. Similarly, attentiveness incorporates components of mathematical knowledge for teaching (MKT) through its reliance upon the teacher's knowledge of how students interact with and learn mathematical concepts. Lastly, attentiveness includes elements of pedagogical practice which align with progressive formalization and recent appeals for a more purposeful integration of students' mathematical thinking into classroom instruction (NCTM, 2014; Teaching Works, 2018). As defined here, the attentiveness construct provides two affordances for the mathematical research community. First, by weaving together three distinct theoretical strands found in the literature and limiting the construct to the individual student level, it helps to capture and describe the work effective teachers do when eliciting, interpreting, analyzing, and responding to evidence of student thinking. Second, it bounds attentiveness at a grain size suitable for operationalization and potential large-scale assessment.

The instrument development work described here is, in part, undertaken to assist national, state, and local efforts to hold teacher preparation and professional development programs accountable for the work they do (e.g., Council for the Accreditation of Educator Preparation, 2013; Grossman et al., 2009). For teacher preparation and professional development programs, high quality instruments designed for program evaluation and research provide key benefits for stakeholders. Through use of a meaningful and easily-implemented assessment of attentiveness, stakeholders can more easily identify program strengths and weaknesses and make more informed decisions about program content, activities, and goals.

2.2. Selected-response Items vs. Constructed-Response Items

Meaningful operationalization of complex constructs such as attentiveness is a critical consideration when constructing instruments, particularly when the provision of efficient methods of scoring and interpretation is a key desired outcome. According to the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014), constructed-response (CR) items encompass short-answer items, performance tasks, and portfolios. CR items range from simple to complex and can elicit very short to very long responses (Hogan, 2013); CR items can offer contextualized tasks, but are time-consuming to score and can suffer from poor rater reliability (Humphry & Heldsinger, 2014; McMillian, 2013). Selected-response (SR) items ask students to choose an answer from given alternatives, require short response times, and are easy to score, though the structure of the items may allow students with surface level knowledge to respond correctly through use of response option cues

(Popham, 2017). Thus, weighing the pros and cons of each and deciding which benefits and trade-offs to accept remain ongoing challenges for test developers.

The equivalence of SR and CR measures (i.e., whether SR and CR assess the same trait, knowledge, or level of knowledge) has been explored using predominantly three kinds of procedures – (1) direct correlation between SR and CR measures including factor analytic methods, (2) correlation between SR and a superordinate criterion compared to correlation between CR and the same criterion, and (3) an experimental design to study the treatment effect or instructional sensitivity of an intervention comparing students' performance on SR and CR measures (Hogan, 1981, 2013). Past research utilizing these three methods has shown a lack of measurement differences between SR and CR tests, suggesting that students score similarly regardless of the test format (Hogan, 1981; Rodriguez, 2003; Troub, 1993). Hogan (1981) reviewed 35 studies that investigated the relationship between SR and CR tests and found them to be equivalent or nearly equivalent in most correlational studies. Student scores were highly correlated on tests of SR and CR when the measurement errors of the tests were taken into consideration; in cases where they differed and when there was a superordinate criterion, the SR test was often more correlated to the external criterion. Troub (1993) reviewed nine studies that investigated SR-CR equivalence in four domains – writing, word knowledge, reading comprehension, and quantitative ability. Findings included SR-CR construct equivalence for reading and quantitative ability, a lack of equivalence for writing, and a contradictory finding between the two studies on word knowledge. Despite the mixed findings, Troub (1993) concluded that based on the explained score variance, the differences between the two measures, if they existed, were likely to be small for any domain. Rodriguez's (2003) meta-analysis of 29 studies suggested that when SR and CR tests were created purposefully to measure the same cognitive behavior, SR and CR tests' equivalence can be achieved, especially with same stem (stem-equivalent) items. It is worth noting, though, that only studies that assessed examinees' knowledge instead of affect or personality were reviewed; Hogan's (1981) review also excluded studies that examined the communication skills of reading, writing, and speaking. Hogan (2013) discussed possible reasons for the lack of differences between SR and CR items: (1) the SR and CR instruments often measured constructs at a coarse-grain level, so more fine-grain differences may not be picked up by the instruments; (2) the measurement errors associated with the SR and CR instruments are often underestimated and can be sizable compared to the effect of test formats; (3) the differences in cognitive skills assessed by SR and CR measures are overestimated, meaning that the items on each type of assessment require similar skill or ability to complete; and (4) individual differences in cognitive skills of test-takers are underestimated and mask the minor differences caused by test formats.

Despite the equivalence of SR and CR tests' effects on students' outcomes, studies have shown that test format may interact with student characteristics such as race (Longstreth, 1978), ability level (Snow, 1993), test anxiety and self-confidence (Martinez, 1999), and gender (DeMars, 2000), suggesting that students with certain characteristics may perform better on one test format than another. Thus, awareness of potential test item effects should inform the item construction process. Because SR and CR measures have been shown to assess constructs equivalently, the more efficiently scored and interpreted SR items have a clear advantage for use in assessing teacher attentiveness. Purposeful construction of SR items to measure the same level of construct complexity that CR items measure is key. As Rodriguez and Haladyna (2013) note, "Can SR and CR items measure the same construct? They can when we write them to do so" (p.295).

2.3. Validation of Response Processes Using Think-Aloud Interviews

Validation is an ongoing process that involves providing evidence to support or refute the assumptions inherent in a particular score interpretation and use (AERA, APA, &, NCME, 2014). Validation arguments involve clearly stating the proposed score interpretation and use, articulating the claims and assumptions that underlie the score interpretation and use, followed by gathering evidence to

support or refute the claims (Kane, 2013). Oftentimes, the most critical or suspect claims are prioritized for investigation.

Test interpretations often involve articulating where, along a construct's continuum of development, a particular test score lies and where the test respondent might be placed along that continuum. Making these types of test interpretations involves making claims about the relationship between the cognitive activity of the respondent as they respond to test items and the construct of interest. Therefore, in situations where such claims are made, test developers must provide response process validity evidence (Padilla & Benítez, 2014) or " . . . evidence concerning the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers" (p.15).

Leighton (2017) articulates the distinction between two common approaches to gathering response process validity evidence: think-aloud interviews and cognitive interviews. Both involve one-on-one interviews between a researcher and a participant. For the purposes of assessment design and validation, think-aloud interviews typically focus on understanding the participant's problem-solving process, while cognitive interviews focus on understanding how the participant is comprehending a particular assessment item. This distinction informs both how interviews are conducted, and data are analyzed.

When CR items are converted to SR items to ease the scoring burden, evidence of alignment between response processes is necessary to support the claim that conversion will not compromise score interpretation. Analysis of think-aloud interview data, where respondents articulate their problem-solving process for both a CR item prompt and its paired SR item prompt can be used to examine the alignment across item types (Troub, 1993).

3. Study Objectives and Questions

This study examines data from think-aloud interviews to determine if responses to item types that can be efficiently scored (e.g., SR items) align with responses generated by open-ended items (e.g., CR items) for the construct of teacher attentiveness. Unlike previous studies that examined the equivalence of two test formats (e.g., Lukhele, Thissen, & Wainer, 1994; Thissen, Wainer, & Wang, 1994), this study examines students' responses at the item level and provides insights into students' think-aloud processes on the construct of attentiveness and the affordances and limitations of using SR items to target the same level of complexity as CR items. More specifically, there are two research questions:

(1) Do scores from the CR items align with scores from the paired SR items?
(2) When scores from the two formats do not align, what features of the assessment items contribute to the misalignment and could be adjusted?

4. Methods

4.1. Participants

Selecting participants to interview for validation studies is less about quantity and more about purpose (Padilla & Leighton, 2017). Our specific purpose was to examine evidence in relation to the use and score interpretation of assessment items created to measure attentiveness for candidates in mathematics teacher preparation programs and for mathematics teachers in professional development settings. Because representatives of these populations also generated responses to the CR items from which the SR items were built, choosing interviewees with the potential to exhibit the full range of low, medium, and high levels of attentiveness across both CR and SR items was a critical goal. Consequently, our recruitment efforts focused on interviewing a representative sample of in-service secondary mathematics teachers with a range of classroom experience and of preservice secondary mathematics teacher candidates from both the undergraduate and graduate levels. For ease of

communication, we refer to the group of individuals who participated in this study as "participant teachers" and use the abbreviation PT with a number (e.g. PT3) to refer to a particular individual.

We interviewed twelve participant teachers. Five participant teachers (PT1, PT2, PT3, PT9, and PT12) were undergraduates enrolled in two different mathematics teacher education programs at large state universities who had not yet completed their student teaching and were at various points in the program. PT7 and PT11 were recent graduates of an undergraduate teacher preparation program who had yet to be hired as classroom teachers. PT8 was a graduate student enrolled in the first semester of a three-semester teacher preparation program. PT6 had just completed the same graduate program and was seeking employment as a mathematics teacher. Three additional participants (PT4, PT5 and PT10) were secondary mathematics instructors with classroom teaching experience ranging from five to 33 years.

4.2. Instrument Development – Selected-Response Format and Constructed-Response Format

The Disciplinary Attentiveness to Student Ideas (DASI) item development process has been detailed in previous publications (Atkins Elliott, Totorica, Carney, & Hagenah, 2020; Carney et al., 2017; Carney, Totorica, Cavey, & Lowenthal, 2019). We provide a brief overview here to help the reader better understand the overall item structure and the CR and SR item types. Generally, and in the context of mathematics, the DASI item development process makes use of authentic student responses to K-12 tasks to examine teachers' attentiveness to student thinking. More specifically, the DASI item development process makes use of the framework from the professional noticing literature (Jacobs et al., 2010) of identifying the mathematical intent of a task, followed by examining student work samples and describing (a) the student approach, (b) what the approach indicates about student understanding, and (c) an appropriate pedagogical response to the student based on the work presented and the disciplinary intent of the task.

The DASI-Quantitative Reasoning Inventory (DASI-QRI) was developed for use within the National Science Foundation funded Video Case Analysis of Student Thinking (VCAST) project. VCAST's purpose is to develop instructional materials with the potential to increase secondary mathematics teacher candidates' ability to analyze and respond to student thinking in quantitative reasoning contexts, and use of the DASI-QRI as a pre-post measure helps us examine the impact of the VCAST intervention on teacher candidates' attentiveness.

To develop the DASI-QRI assessment, we first administered quantitative reasoning tasks (see *Tasks given to secondary students* in Figure 1) to secondary students. Next, we analyzed the secondary student data in order to select work samples with the potential to elicit a range of teacher interpretation and pedagogical response (see *Secondary student response to task in* Figure 1). For the CR items, these secondary student responses to the quantitative reasoning tasks are presented and respondents are prompted to (a) describe the student approach, (b) make inferences about student understanding, or (c) describe their hypothetical response to the student (see *Constructed Response (CR) task prompt* example in Figure 1). In addition, some items ask respondents to identify the mathematical intent of the task itself. Though stopping our item development process at this point would support many of our assessment development goals, the scoring burden and scalability limitations of a CR instrument led us to the final stage of item development. Following analysis of the CR item response data, we selected one relatively common response from each of the three hierarchical coding categories to use in the paired SR item (see *Selected Response (SR) task prompt* example in Figure 1).

4.3. Think-Aloud Interview Procedures

The purpose of conducting the think-aloud interviews was to examine the alignment of scores between paired CR and SR items (RQ1) and to identify potential SR item modifications when issues of misalignment occurred (RQ2). We utilized a think-aloud interview methodology due to (a) our

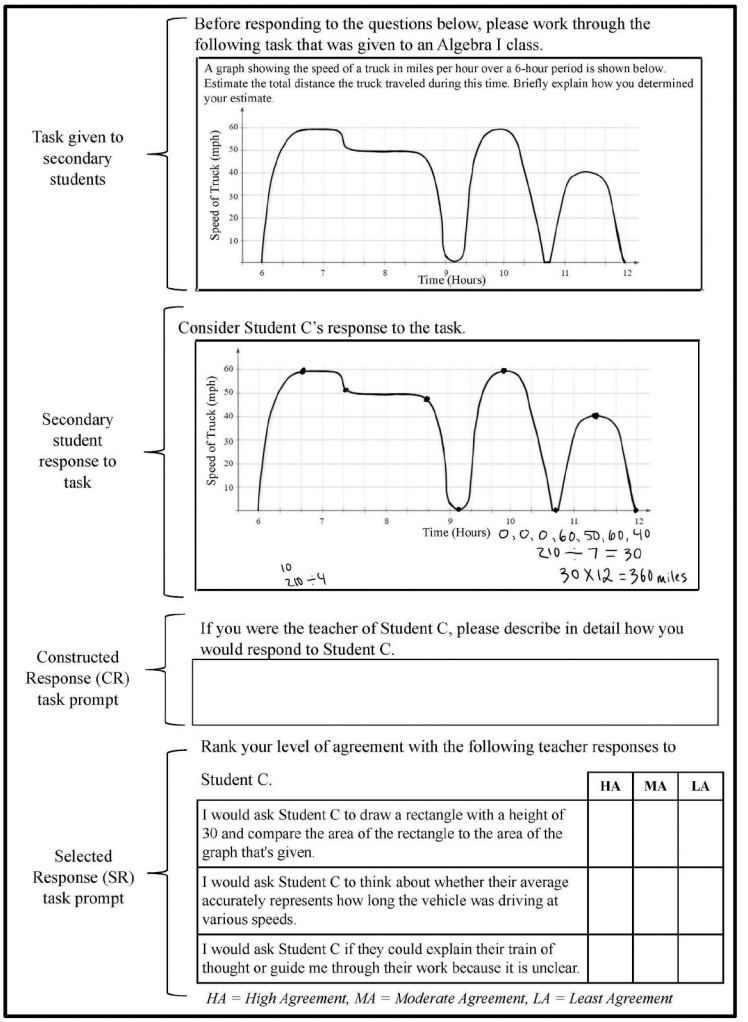

Figure 1. DASI-QRI example item (truck student C response) with CR and SR task prompts.

focus on understanding respondents' thinking while completing the CR items (Leighton, 2017), and (b) our use of a previously-developed cognitive model for coding the CR items.

Our think-aloud interview protocol incorporated the eight mathematical tasks (see *Task given to secondary students* for one example) in the DASI-QRI. For each task, the CR version of an item was presented first and then followed immediately by the SR version. Each version was presented individually on separate pieces of paper. As recommended by Leighton (2017), we focused on obtaining Type 2 verbalization for the CR items through requests for concurrent verbalization while participants engaged in analyzing and responding to the task.

Three of the authors conducted the think-aloud interviews, which were video recorded using an iPad. For the majority of the interviews, two researchers were present: one to interview and the other to provide technical support related to the video recording. Transcription of the video recordings occurred after all interviews were completed. Each interview transcription was embedded within a copy of the interview protocol so the item and transcription could be viewed simultaneously.

4.4. Scoring Process of Constructed-Response

Analysis of the think-aloud interviews involved protocol analysis (Leighton, 2017) and use of a previously-developed cognitive model for the coding of verbal reports. The general cognitive model for DASI items involves a hierarchical coding system of disparate (score of 0), general (score of 1), and specific (score of 2) response categories (detailed in Table 2 of Carney et al., 2017). When developing an SR item, this general cognitive model is used to select exemplar responses from the CR item data for use in the SR item creation. Once an SR item is generated, the item itself can be used as a cognitive model that encompasses the more general model but is specific to the content of the task and its associated secondary student response.

Responses to the SR items were coded based on the scoring system already established for these items (that is, the built-in hierarchy in the selected response options). Our scoring of the verbal reports for each CR item consisted of two phases. The first phase involved identifying whether the verbal report itself aligned with any of the response options provided by the SR item. The second phase involved the assignment of scores to the CR item response and a calculation of the difference between the CR item and SR item scores. If the verbal report for a CR item:

- Aligned with one of the options provided by the SR item, the CR item response was scored using the associated SR score. The difference for the item was calculated using this resultant CR item score and the respondent's actual SR item score.
- Aligned with more than one of the options provided by the SR item, the CR item response was scored based upon the aligned SR options' highest score. This was based on the assumption that when presented with all SR options, the respondent would rank the highest-scoring SR option evident in their verbal report as "high agreement."
- Did not align with any of the three SR options, we coded and scored the response based upon the general DASI-QRI hierarchical coding categories and then computed the difference between scores.

There were three coders for the verbal report data. The coders were all well versed in the general cognitive model, the DASI item development process, and the mathematical ideas associated with the items. To calibrate coding, each coder independently coded the alignment between the CR item verbal report data and its associated SR item options for one third of the participants. The coders then met to discuss, compare and reach consensus on the alignment codes. When raters differed on codes, a discussion was held until agreement was reached across all three coders. This process was conducted two additional times for the remaining participants' verbal report data. Table 1 provides an example of the resulting data for each respondent for the Truck C Response – SR and CR items.

4.5. Statistical Analyses

Our statistical analyses focused on the alignment between responses and scores for eight item stems across the two item types. Three methods were used to measure the alignment between scores. The first method examined the Goodman and Kruskal's gamma correlation. It is calculated as

Table 1. Truck C response CR and SR item scores for all think-aloud interview participants.

Item		PT1	PT2	PT3	PT4	PT5	PT6	PT7	PT8	PT9	PT10	PT11	PT12
Truck C Response	SR	2	2	1	2	1	2	0	2	1	1	1	0
	CR	2	2	1	1	2	2	2	2	1	1	1	2
Difference		0	0	0	1	−1	0	−2	0	0	0	0	−2

$$G = \frac{N_c - N_d}{N_c + N_d}$$

where N_c is the total number of pairs that rank the same (concordant pairs); N_d is the total number of pairs that don't rank the same. The gamma correlation's (i.e., G's) possible range is [−1, 1] with 1 indicating a perfect positive relationship, −1 indicating a perfect negative relationship, and 0 indicating no association between variables. Gamma correlation was run to determine the association between SR and CR amongst 12 participants for the eight items using SPSS v.25.

The second method used the average difference measure to gauge the magnitude of the difference between SR and CR item scores across a sample of 12 subjects. There are two variations. The average absolute difference measure is calculated as the average of the absolute differences between each pair of SR and CR item scores across 12 subjects. Its possible range is [0,2] for seven items except for item "LOBF Student N Understanding" whose range is [0,3].

$$d_i = \frac{\sum_{j=1}^{n} \left| y_{sij} - y_{fij} \right|}{n}$$

where d_i is the average difference measure for item *i;* n is the number of the subjects in the sample (n = 12 in our study); y_{sij} is the selected-response score of the subject j on item *i;* y_{fij} is the constructed-response score of the subject j on item i.

Another average difference measure is calculated as the square root of the average squared difference. Both variations of the average difference measure remove the signs of the differences between selected-response scores and constructed-response scores, but the squaring (i.e., the second variation) makes the large differences more important. Its possible range is [0,2] for seven items except for item "LOBF Student N Understanding" whose range is [0,3].

$$d_i = \sqrt{\frac{\sum_{j=1}^{n} \left(y_{sij} - y_{fij} \right)^2}{n}}$$

where d_i is the average difference measure for item *i;* n is the number of the subjects in the sample (n = 12 in our study); y_{sij} is the selected-response score of the subject j on item *i;* y_{fij} is the constructed-response score of the subject j on item i. Both average absolute difference measure and the square root of the average squared difference were calculated for eight items.

The third method used a related-samples Wilcoxon signed-rank test to check whether the differences of the formats were statistically significant and to determine the effect sizes. Wilcoxon signed-rank test ranks the differences between scores in the two item types for each item, sums the ranks that come from a positive difference between the two item types, and sums the ranks that come from a negative difference between the two item types. When the difference is 0, the scores are excluded from the ranking. The sum of positive ranks is the test statistics (T), which is compared to the mean rank ($\overline{T}$) in the unit of the standard error of the mean rank ($SE_{\overline{T}}$) to calculate the significance of the test statistics using the formula below.

$$z = \frac{T - \overline{T}}{SE_{\overline{T}}}$$

The Wilcoxon signed-rank test was conducted on the eight items using SPSS v.25. Due to the small sample size and the number of zero score differences as a result of the same score on both the SR item type and CR item type for each item, significance testing results should be interpreted with caution. Nevertheless, the effect sizes can shed light on the direction and magnitude of the differences.

4.6. Qualitative Follow-Up Analyses

To determine which items warranted further investigation, we agreed upon a set of criteria based on quantitative alignment results. Specifically, we identified items for additional investigation that met one or more of the following conditions: (1) Gamma Correlation below.5, (2) average absolute difference above .5, or (3) average difference based on squaring above .9. Once an item was identified for further investigation, we examined the item's data analysis tables to identify score differences greater than one (e.g., a score of 0 on the constructed response and a score of 2 on the selected response version of the item) and the participants who produced those differences. Finally, for each of these participants, we examined their item transcripts and video data. Three researchers compared notes on each instance and collectively developed a written summary. After all summaries were written, we searched for potential patterns in the results.

5. Results

5.1. Statistical Analyses Results

5.1.1. Goodman and Kruskal's Gamma Correlation Results

Goodman and Kruskal's gamma correlation was run to determine the association between SR and CR scores amongst 12 participants for the eight items. There was a strong, positive correlation between SR and CR scores for the Truck Intent item (G = .538, p = .474), the LOBF Intent item (G = .667, p = .107), and the LOBF Student N Understanding item (G = .722, p = .026), which was the only item with a statistically significant correlation. There were positive, but not strong, correlations between SR and CR scores for the Truck Student C Response item (G = .185, p = .663) and the Truck Student D Approach item (G = .394, p = .348). SR and CR scores for the Hexagon Student M Approach item had a perfect negative correlation that was not statistically significant (G = −1.000, p = .355). No gamma statistic was reported for the Hexagon Student L Approach item because the score on the CR item type was a constant value of 2 or for the Hexagon Student M Response item because the score on the SR item type was a constant value of 2.

5.1.2. Average Difference Measure Results

The average absolute difference and the average difference calculated as the square root of the average squared difference between SR and CR scores was conducted for each of the eight items. A less than 0.1 point average absolute difference was found for the Hexagon Student L Approach item and the Hexagon Student M Response item (in both cases, d = .083) between the SR and CR item types with an average difference based on squaring d = .289. Truck Intent (d = .417), LOBF Student N Understanding (d = .250), and Hexagon Student M Approach (d = .417) had an average absolute difference of more than 0.1 point but less than 0.5 points with an average difference based on squaring of .764, .645, and .866, respectively. There were three items – Truck Student C Response (d = .500), Truck Student D Approach (d = .500), and LOBF Intent (d = .583) – of which the average absolute difference was equal or more than 0.5 with an average difference based on squaring of .913, .913, and .866, respectively. The fact that the Truck Student C Response item and the Truck Student D Approach item had a smaller absolute difference (.500 vs. .583) but a bigger average difference based on squaring (.913 vs. .866) than the LOBF Intent item suggested that these two items (i.e., Truck Student C Response and Truck Student D Approach) had larger score discrepancies (e.g., from 0–2 instead of from 0–1 or 1–2) between pairs of students' scores on the SR item type and the CR item type than the LOBF Intent items did.

5.1.3. Related-samples Wilcoxon Signed-rank Test Results

The results of the Related-Samples Wilcoxon Signed-Rank Test show that there is not a statistically significant difference between students' scores on the SR item type and the CR item type on any of the eight items:

- Truck Intent (z = −1.134, p = .257, r = −.231)
- Truck Student C Response (z = 1.300, p = .194, r = .265)
- Truck Student D Approach (z = −1.300, p = .194, r = −.265)
- Line of Best Fit Intent (z = 1.000, p = .317, r = .204)
- Line of Best Fit Student N Understanding (z = 1.342, p = .180, r = .274)
- Hexagon Student L Approach (z = 1.000, p = .317, r = .204)
- Hexagon Student M Approach (z = .272, p = .785, r = .056)
- Hexagon Student M Response (z = −1.000, p = .317, r = −.204)

According to Cohen (1988), an effect size of r = .100, .300, or .500 represents a small effect, medium effect, and large effect, respectively. The absolute values of the effect sizes for all the items were in the range of .204-.274, indicating small to medium effect sizes, except for the Hexagon Student M Approach item. For the Hexagon Student M Approach item, an effect size of .056 indicates a minimal change in students' performance on selected-response and constructed-response formats; nine subjects out of 12 have the same score on the two formats. A positive effect size for an item indicates that students' CR item scores tends to be higher than SR item performance; a negative effect size of an item indicates the exact opposite, meaning that students' CR performance tends to be lower than SR performance.

5.1.4. A Synthesis of Statistical Analyses Results

The summary statistics are reported for the three methods in Table 2.

Though gamma statistics were not reported for the Hexagon Student L Approach and Hexagon Student M Response items due to constant values in either their SR scores or CR scores, the reported gamma correlations were positive between SR scores and CR scores for all other items except Hexagon Student M Approach. The positive gamma correlation suggests that, for most of the items, the better the participants' performance on the SR item, the better their performance on the paired CR item. The perfect negative gamma correlation for the Hexagon Student M Approach item may come from the fact that one participant who scored 2 on the SR item scored 0 on the CR item while another participant who scored 0 on the SR item scored 2 on the CR item. With a small sample of 12 participants, individual performances have a strong influence on results. Nevertheless, this item warrants further investigation to ensure the construct assessed in the SR and the CR items are comparable. For the Hexagon Student L Approach and Hexagon Student M Response items with no gamma correlation statistics reported, the average differences between participants' SR and CR scores were less than .3 points for both items. The average differences between the paired SR and CR scores are less than 1 for all of the items. Two items, Truck Student C Response and Truck Student D Approach, have relatively larger average differences (d = .913 in both cases) as well as smaller gamma correlation coefficients (G = .185 and G = .394, respectively) than other items. For these two items, participants tend to produce relatively more different scores on the SR format and the CR format. The

Table 2. Summary statistics for all CR and SR item pairs.

		Average Difference		Wilcoxon Signed-Rank Test	
Item Name	Gamma Correlation	Absolute	Based on Squaring	p-value	Effect Size
1. Truck Intent	.538	.417	.764	.257	−.231
2. Truck Student C Response	.185	.5	.913	.194	.265
3. Truck Student D Approach	.394	.5	.913	.194	−.265
4. Line of Best Fit Intent	.667	.583	.866	.317	.204
5. Line of Best Fit Student N Understanding	.722*	.25	.645	.18	.274
6. Hexagon Student L Approach	N/A[a]	.083	.289	.317	.204
7. Hexagon Student M Approach	−1	.417	.866	.785	.056
8. Hexagon Student M Response	N/A[b]	.083	.289	.317	−.204

*significant at α =.05 level. [a] There was no gamma statistic reported because the constructed response is a constant with a value of 2. [b] There was no gamma statistic reported because the selected response is a constant with a value of 2.

results of the Related-Samples Wilcoxon Signed-Rank Test show that for three items – Truck Intent, Truck Student D Approach, and Hexagon Student M Response, when there is a difference, participants' CR performance tends to be lower than their SR performance. For four items – Truck Student C Response, LOBF Intent, LOBF Student N Understanding, and Hexagon Student L Approach, when there is a difference, participants' CR performance tends to be higher. The results manifest that the formats yield equivalent data on balance.

Ideally, a high gamma correlation, a small average absolute difference, a small average difference based on squaring, and a statistically insignificant result for the Wilcoxon Signed-Rank Test with a small effect size is most desirable for each item. However, different methods highlight different aspects of the results and yield complementary information that other methods do not provide. Goodman and Kruskal's gamma correlation reveals the number of concordant pairs in relation to the number of discordant pairs and highlights unusual characteristics of data such as a constant score for an item type or a strong influence of certain cases. Still, it does not differentiate between the magnitude of the score difference for item types. The average absolute difference and the average difference based on squaring afford intuitive interpretations because the results are still on the original score scale. Also, the average difference based on squaring makes the larger differences more important. However, neither average difference measure shows us the direction of the difference. The non-parametric Wilcoxon Signed-Rank Test shows us the direction and magnitude of the differences with effect sizes. However, while we can categorize the effect sizes into small, medium, and large according to a benchmark, the actual numerical values of the effect sizes are calculated based on the ranks of the data; thus, they are not a direct representation of the original scores. With the small sample sizes often used in think-aloud interviews and unique data characteristics such as a constant score on an item across participants or frequent occurrences of zero score differences between item types, a combination of statistical analysis methods is both helpful and necessary. The three methods used in our study are examples of techniques that can be used for exploration and investigation. A synthesis of the results provide both evidence of the alignment between scores received on the CR and SR items and highlight items that need further investigation and development.

5.2. Items that Need Further Investigation and Development

Using our criteria for less-than-satisfactory quantitative alignment results, four items were identified for additional investigation: (1) Truck Student C Response, (2) Truck Student D Approach, (3) Line of Best Fit Intent, and (4) Hexagon Student M Approach. Across the four items, six participants were identified as having CR and SR scores that differed by 2. PT12 had a score difference of 2 on two items and was the only participant whose scores differed by 2 on more than one item. Table 3 shows the participants with score differences of 2, the items for which the score differences occurred, and the scores earned on each version of the item.

5.2.1. Score of 0 on the CR and 2 on the SR

Two participants (PT8, PT9) scored 0 on the CR version and 2 on the SR version of the Truck Student D Approach item. After reexamining the interview data, it appears that both participants had difficulty determining the same critical aspect of Student D's approach as they worked through the CR version and resolved that difficulty after seeing the SR version. Consider PT8's response as they examined

Table 3. Items identified for additional investigation and participants with score differences of 2.

	Score Difference of Two	
Item	0 on CR and 2 on SR	2 on CR and 0 on SR
Truck Student C Response		PT7, PT12
Truck Student D Approach	PT8, PT9	
Line of Best Fit Intent		PT1
Hexagon Student M Approach	PT12	PT2

Student D's approach, "I'm not sure how they got some of their numbers for their graph . . . maybe they were trying to make some sort of estimate of how long the, yeah, I'm not sure where they got the numbers that they averaged." Similarly, PT9 expressed confusion about how Student D selected the points on their graph, noting two points in particular that were unclear, "this 10 I'm confused by . . . I'm just not sure why that isn't a zero and why that's not a 40 though." However, after reading the options in the SR version of the Truck Student D Approach item, PT9 stated, "Oh, okay. That makes sense. It's definitely. That is definitely the best one. They lined it up with the hour marks. Now that's [pause] okay [pause] I'm seeing it now." After reading, "Student D found the instantaneous speed at each hour mark" from one of the options, PT8 remarked, "Oh, yeah, they did do that." In terms of scoring, their responses to the CR version fit best with the lowest level of attentiveness as represented by the SR options, but when presented with the SR version of the task, both participants indicated their highest level of agreement with the highest-scoring SR option.

A similar situation occurred with PT12, who scored 0 on the CR version and 2 on the SR version of the Hexagon Student M Approach item. This participant's response to the CR version focused on the last, incorrect segment of the student's work and aligned best with the lowest-scoring option "Student M wrote the wrong equation in Step c." However, with the SR version, PT12 agreed most with the highest-scoring option: "So, I mean it's a true statement, but it, it's the wrong equation. But I would rather highlight the things that she did right."

Overall, when participants had score differences of two by scoring 0 on the CR version and 2 on the SR version of the item, there is evidence that exposure to the SR options can improve the score (from 0 on CR to 2 on SR). From the perspective of our goal of measuring attentiveness, it appears the SR options for these items provide some further opportunity (when compared to the CR versions) to demonstrate attentiveness. In the cases we identified, this appears to happen when a participant experiences some minor confusion regarding the student's approach or when the participant focuses on a particular error in student thinking. In either case, the participant is able to move to a stronger pedagogical position when they see the SR options.

At the same time, in the case of the Truck Student D Approach item, evidence suggests that the range of participants' analyses of the provided secondary student's problem-solving approach is not adequately represented in the SR version of this item. Because of the mathematical complexity of the task, the SR options include a variety of mathematical ideas (e.g., identifying how points were selected, recognition of an averaging strategy, and possible treatment of the speed values at each hour as distance traveled in the hour) in different combinations. For participants who cannot initially figure out the significance of Student D's point selection, providing point selection strategies in the SR options may contribute to a higher SR score. One potential revision is to eliminate the point selection strategy from the SR options altogether, thus narrowing the range of mathematical ideas represented in the item.

5.2.2. Score of 2 on the CR and 0 on the SR

Two participants (PT7, PT12) scored 2 on the CR version and 0 on the SR version of the Truck Student C Response item. When reexamining the interview data for these participants, we noted that each participant expressed multiple ideas on the attentiveness continuum when responding to the CR version and then prioritized more formalized mathematical ideas when responding to the SR version. For instance, when PT7 shared their thinking in response to the CR version, they initially commented on Student C's work, saying, "It feels to me like that's something to do with area, like they tried to do some area under the curve to [breaks off]" and a moment later remarked that they would want to, "sit down with them and, and kinda ask them what they, how they thought about the problem." As PT7 continued, they appeared to change their mind about the reasoning behind Student C's work, saying, "Okay, I see what they did," then stating, "what I would respond to this student with is, um, first that, that that's an interesting insight. Um, and that I would push them to consider what's happening between these points." In this instance, PT7's response provides evidence of all three levels of attentiveness represented in the SR version of this item. Thus, following our scoring protocol for

instances where the PT's response addressed multiple levels of attentiveness, we assigned the highest score of 2. In the case of PT12, their observations were very similar to PT7's with PT12 also noting that Student C had used an incorrect number for the amount of time lapsed. In both cases, the participants selected "I would ask Student C to draw a rectangle with a height of 30 and compare the area of the rectangle to the area of the graph that's given" as the response with which they had the highest agreement for the SR version of the task, resulting in a score of 0 on the SR version.

One participant, PT1, scored 2 on the CR version and 0 on the SR version of the Line of Best Fit Intent item. Our analysis of the interview data revealed this participant expressed informal notions associated with the task in response to the CR version of the item but indicated highest agreement with the SR option that used more formal vocabulary. In response to the CR version, PT1 explained that the task involves, "understanding that like [pause] being able to look at this and be like, oh it's not linear, but there, we can put a line on it so that there's the least amount of data that are outliers." In this way, PT1 appeared to make a case that the mathematical task involves modeling the data with a linear function in a way that minimizes the variance, which is associated with the highest attentiveness score on this item. However, when presented with the SR version of this item, PT1 indicated highest agreement with the lowest level option, "The task targets how the shape of a graph can be used to determine the relationship between an independent and dependent variable," a selection that did not align well with their CR response.

A similar situation occurred with PT2, who scored 2 on the CR and 0 on the SR for the Hexagon Student M Approach item. PT2's response to the CR version was extremely brief, and then they selected an option from the SR version that was not mentioned in their CR response. During the CR version, PT2 stated, "I think what the student is doing is, um, in this one they're trying to, uh, they're taking more of a subtractive approach. They're trying to say, okay, I know hexagons have six sides always, and then trying to figure out how many sides we're not counting at a time." Based on our scoring protocol, PT2's response to the CR version aligned with the highest level of attentiveness. However, when presented with the SR version, PT2 indicated they agreed most with the lowest level option, "Student M wrote the wrong equation in Step c," an idea which was not expressed during their think-aloud response to the CR version of the item.

Overall, when participants had score differences of two by scoring 2 on the CR version and 0 on the SR version of the item, evidence suggests that in some cases, participants tend to focus on multiple aspects of the mathematics in the CR item but then default to the formal perspective of the mathematics when presented with the SR options. In other cases, evidence suggests that participants sometimes select options in the SR versions of items that do not directly correspond to the ideas expressed during their think-aloud CR response. In the case of PT2 and based upon the ideas shared during their response to the SR version of the Hexagon Student M Approach item, they seemed to recognize the equation was incorrect but it is difficult to determine if PT2 held this idea prior to seeing the SR options since it was not previously articulated. In most of these cases, the participants appeared to default to a more formal perspective on the mathematical ideas involved when viewing the SR options. This may be due to confusion on the part of the participants about the intent of the mathematical task associated with the item, or perhaps because the participants cast themselves in the role of "student", rather than considering the mathematical sophistication of students for whom the task is intended. It is an open question whether or not these four participants would have selected the highest scoring option on the related items had the participant not already talked through the CR.

6. Discussion

Teacher preparation and professional development programs are charged with equipping participants with the student-centered pedagogies needed for effective mathematics instruction (NCTM, 2014; Teaching Works, 2018). Yet the challenges inherent to measuring complex constructs, like attentiveness, being targeted for development make meeting this charge elusive. Due to a lack of efficient, easily-scored instruments which measure such complex constructs, evaluating program outcomes and

accurately identifying program strengths and weaknesses continue to present challenges for stakeholders. The use of response process evidence to inform assessment development as described here has the potential to ameliorate these challenges, whether by affording meaningful program evaluation or by informing decisions related to future improvements and program design.

Our study examined students' responses at the item level for item revision. We used nonparametric statistical procedures to study the alignment between SR-CR items. When the SR items were carefully constructed by using exemplar responses elicited by constructed-response item prompts as alternatives, evidence of alignment between scores received on the CR and SR items was observed. Our review of empirical research (e.g., Hogan, 1981, 2013) suggests the equivalence of SR-CR measures and this study provides additional support for that claim. Our statistical analyses also highlighted items that need further investigation and development. Troub (1993) called for the next generation of studies to use the think-aloud procedure to provide evidence for the cognitive processes examinees use while answering SR and CR items. This study answered that call by utilizing think-aloud interviews to investigate the affordances and limitations of using SR items to target the same level of complexity as CR items. Past research has examined the equivalence of SR-CR measures for knowledge (Hogan, 1981). The current study extends this research to a more complex construct – teacher attentiveness. Think-aloud interviews unveiled the problem-solving processes for the CR items and their paired SR items. Though four items were identified for further investigation in this study, our analysis of the think-aloud interviews revealed potential revisions that can be made to improve alignments between the responses and scores of the items across the two item types. As Rodriquez and Haladyna (2013) suggested, this study supports the claim that careful construction allows SR items to target the same level of complexity as CR items.

Implications

This work has implications for test development in three areas: (1) how think-aloud interviews can be used to investigate the problem-solving processes elicited by alternative item types used to assess complex constructs such as attentiveness; (2) how a combination of statistical methods can be used to investigate the quantitative data produced from a limited sample size and number of items in think-aloud interviews; and (3) how quantitative data can complement qualitative data to provide response process evidence for validation argument claims specific to item construction and score interpretation.

Think-aloud interviews are often used to investigate human information processing and to determine whether test items elicit the desired problem-solving processes (Padilla & Benítez, 2014; Padilla & Leighton, 2017). Attentiveness, a complex construct that blends aspects of professional noticing (Jacobs et al., 2010), mathematical knowledge for teaching (Ball et al., 2008; Shulman, 1987), and progressive formalization (Freudenthal, 1973; Gravemeijer & van Galen, 2003; Treffers, 1987), requires higher-level processing skills including problem-solving in the context of teaching and learning; thus, think-aloud interviews can be a useful tool for investigating item functioning for assessments intended to measure attentiveness. For the purpose of operationalizing the measurement of attentiveness with efficient methods of scoring and interpreting, think-aloud interviews can be further used to (a) compare response processes elicited by different item types, (b) ensure the desired response processes are elicited by specific item types, and (c) determine whether items are well constructed for their assessment purpose.

This work has also shown that a synthesis of results arising from a combination of statistical methods is helpful in minimizing the sample size required for think-aloud interviews. Quantitative analysis of participant scores on CR and SR assessment item pairs that have been carefully constructed to be comparable provide valuable information on whether scores across item types align as intended. In think-aloud interviews, a small number of participants are typically recruited for an in-depth study; for exploration purposes, a limited number of items are also used in those interviews. Small sample sizes, small numbers of items, and unique data characteristics such as a constant score on an item across participants or frequent occurrences of zero score differences between item types present

challenges for quantitative data analyses. This study illustrates how analyses using a combination of statistical methods followed by a synthesis of those analyses' results is both helpful and necessary.

We utilized think-aloud methods in combination with protocol analysis (Leighton, 2017) to examine our interview data with the addition to quantitative analysis elements. Our study is an example of how qualitative data (i.e., participants' verbal reports from the think-aloud interviews) can complement quantitative data (i.e., participants' scores on SR and CR items) to provide response process evidence (AERA, APA, &, NCME, 2014) for validation argument claims specific to item construction and score interpretation. This mixed-methods approach can generate validity evidence to support or refute claims of alignment between response processes across item types, as well as illuminate some of the affordances and challenges of using different item types. In particular, this work provides evidence of score alignment across CR and SR item pairs when the SR items are created using analysis, categorization, and selection of exemplar CR item responses, thus providing a plausible test development practice. Items highlighted for further investigation and development also shed light on the aspects of item construction and score interpretation that warrant caution.

A validation argument approach recognizes the cyclical nature of test development and validation. In the case of investigating response processes, the validation aspect of test development starts the process but then further informs test development and revision. This cyclical process continues until there is sufficient evidence to support the assumption that a respondent's response process as they engage with test items aligns with the construct of interest. In the case of ensuring alignment between CR and SR items, our results indicate the importance of recognizing the cyclical nature of the process as opposed to one designed to demonstrate "validity".

The use of SR items also has implications for assessment administrators and users of assessments. By administering the SR items, assessment administrators will have an easier-to-grade format. With the same amount of time, examinees can respond to more SR items instead of CR items, thus possibly increasing the reliability of the assessment. However, Rodriguez's (2003) review of empirical studies also cautioned that the clueing provided by the multiple-choice options could introduce random error and subsequently reduce reliability. With the SR items, examinees can be asked to distinguish subtle differences between several alternatives in terms of their relative correctness (Popham, 2017).

6.1. Limitations

Because the study examines students' responses at the item level for item revision, the measurement errors associated with the test formats are not accounted for using common procedures like the correction for attenuation/unreliability (Lord & Novick, 1968). Also, due to the study's exploratory nature with a small number of items and participants, psychometric modeling using item response theory (Lord, 1980) cannot be applied to study the characteristics of the items and participants. Instead, nonparametric statistical procedures were used to measure the alignment between the SR and CR items to highlight items for further investigation.

The responses, including the SR and CR answers and the think-aloud interviews, were from 12 participant teachers. Though we purposefully selected the participant teachers from candidates in mathematics teacher preparation programs and mathematics teachers in professional development settings to be representative of the full range of low, medium, and high levels of attentiveness, their responses, especially their thinking processes, are likely to exhibit their individuality.

We investigated using think-aloud interviews for response process evidence for developing SR items from CR items to assess a complex construct – teacher attentiveness. Though we found a combined use of statistical analyses and think-aloud interviews helpful in our item revision, whether this practice can be generalizable to item development for communication skills such as writing or other problem-solving skills requires further investigation (Hogan, 1981).

6.2. Future Directions

We see two potential future directions for our work. First, it is important to examine whether the response process evidence presented is similar or different across subgroups of the intended population. More specifically, is the alignment or misalignment between the item types when used to assess teacher attentiveness related to teachers' characteristics? For example, were score differences of two between CR and SR items associated with participant teachers with fewer years of classroom experience? Second, our current focus is at the item level and examines whether the problem-solving processes elicited by the SR item type and CR item type in a given item pair are similar. With more items assembled into tests, a natural next step is to explore whether CR and SR tests purposefully constructed to measure the same level of attentiveness can yield equivalent scores and generate validity evidence for their respective use and score interpretations.

In the context of this special issue of AME, our aim is to contribute to the ways researchers and test developers think about generating and using response process evidence to provide evidence for or against stated assumptions inherent in a particular score interpretation and use. More work is needed to better understand the issues created by having participants respond to the SR version of an item immediately after responding to the CR version, but our results show this is a promising approach. Further, when we began our quantitative analyses of the CR-SR score differences, we were hard-pressed to find examples of how other researchers have done this. By offering our approach to doing so, we hope to inform the broader conversation and work related to using response process evidence to inform validation arguments, specifically with respect to item development.

Acknowledgments

This material is based on work supported by the National Science Foundation under grant #1726543. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Funding

This research was supported in part by grant #1726543 Preparing Secondary Mathematics Teachers with Video Cases of Students' Functional Reasoning from the National Science Foundation..

References

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

Atkins Elliott, L., Totorica, T., Carney, M., & Hagenah, S. (2020). Developing an assessment of attentiveness for program evaluation. In J. Goodell & S. Koc (Eds.), *Preparing STEM teachers: A replication model* (pp. 311–323).Charlotte, NC: Information Age Publishing.

Ball, D., Thames, M., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*(5), 389–407. doi:10.1177/0022487108324554

Bell, C., Gitomer, D., Savage, C., & Mckenna, A. H. (2019). *A synthesis of research on and measurement of STEM teacher preparation*. American Association for the Advancement of Science.

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

Carney, M. B., Totorica, T., Cavey, L. O., & Lowenthal, P. R. (2019). Developing a construct map for teacher candidate attentiveness. In J. D. Bostic, E. E. Krupa, & J. C. Shih (Eds.), *Quantitative measures of mathematical knowledge: Researching instruments and perspectives* (pp. 152–178). New York, NY: Routledge.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Academic Press.

Council for the Accreditation of Educator Preparation. (2013). *Council for the accreditation of educator preparation report to the public, the states, the policymakers, and the education profession*. Washington, DC: Author.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*(1), 55–77. doi:10.1207/s15324818ame1301_3

Freudenthal, H. (1973). *Mathematics as an educational task*. Dordrecht, The Netherlands: Reidel.
Gravemeijer, K., & van Galen, F. (2003). Facts and algorithms as products of students' own mathematical activity. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 114–122). Reston, VA: National Council of Teachers of Mathematics.
Grossman, P., Hammerness, K., & McDonald, M. (2009). Redefining teaching, re-imagining teacher education. *Teachers and Teaching: Theory and Practice*, *15*(2), 273–289. doi:10.1080/13540600902875340
Hogan, T. P. (1981). *Relationship between free-response and choice type tests of achievement: A review of the literature*. Paper prepared under contract for Education Commission of the States. Princeton, NJ: ERIC Clearinghouse on Tests & Measurements.
Hogan, T. P. (2013). Constructed-response approaches for classroom assessment. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 275–292). Thousand Oaks, CA: Sage.
Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher*, *43*(5), 253–263. doi:10.3102/0013189X14542154
Jacobs, V. R., Lamb, L. L. C., & Philipp, R. A. (2010). Professional noticing of children's mathematical thinking. *Journal for Research in Mathematics Education*, *41*(2), 169–202. doi:10.5951/jresematheduc.41.2.0169
Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, *42*(4), 448–457. doi:10.1080/02796015.2013.12087465
Leighton, J. P. (2017). *Using think-aloud interviews and cognitive labs in educational research*. New York, NY: Oxford University Press.
Longstreth, L. (1978). Level I-Level II abilities as they affect performance of 3 races in the college classroom. *Journal of Educational Psychology*, *70*(3), 289–297. doi:10.1037/0022-0663.70.3.289
Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York, NY: Routledge.
Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement*, *31*(3), 234–250. doi:10.1111/j.1745-3984.1994.tb00445.x
Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*(4), 207–218. doi:10.1207/s15326985ep3404_2
McMillian, J. H. (2013). *Classroom assessment: Principles and practice for effective standards-based instruction* (*6th* ed.). Boston, MA: Allyn and Bacon.
National Council of Teachers of Mathematics. (2014). *Principles to actions: Ensuring mathematical success for all*. Reston, VA: Author.
Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136–144.
Padilla, J. L., & Leighton, J. P. (2017). Cognitive interviewing and think aloud methods. In B. D. Zumbo, & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 211–228). New York, NY: Springer.
Popham, W. J. (2017). *Classroom assessment: What teachers need to know*. Boston, MA: Pearson.
Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, *40*(2), 163–184. doi:10.1111/j.1745-3984.2003.tb01102.x
Rodriguez, M. C., & Haladyna, T. H. (2013). Writing selected-response items for classroom assessment. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 293–311). Thousand Oaks, CA: Sage.
Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1–22. doi:10.17763/haer.57.1.j463w79r56455411
Snow, R. E. (1993). Construct validity and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 45–60). Hillsdale, NJ: Lawrence Erlbaum.
Teaching Works. (2018). Retrieved from http://www.teachingworks.org.
Thissen, D., Wainer, H., & Wang, X. B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement*, *31*(2), 113–123. doi:10.1111/j.1745-3984.1994.tb00437.x
Treffers, A. (1987). *Three dimensions: A model of goal and theory description in mathematics instruction – The Wiskobas project*. Dordrecht, The Netherlands: Reidel.
Troub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & C. W. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 29–44). Hillsdale, NJ: Lawrence Erlbaum.