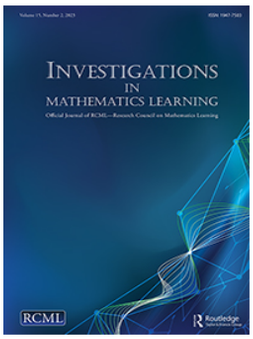


Routledge
Taylor & Francis Group

Investigations in Mathematics Learning

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uiml20

# Unpacking Response Process Issues Encountered When Developing a Mathematics Teachers' Pedagogical Content Knowledge (PCK) Assessment

Martha L. Epstein, Hamza Malik, Kun Wang & Chandra H. Orrill

**To cite this article:** Martha L. Epstein, Hamza Malik, Kun Wang & Chandra H. Orrill (2023): Unpacking Response Process Issues Encountered When Developing a Mathematics Teachers' Pedagogical Content Knowledge (PCK) Assessment, Investigations in Mathematics Learning, DOI: 10.1080/19477503.2023.2201115

**To link to this article:** https://doi.org/10.1080/19477503.2023.2201115

INVESTIGATIONS IN MATHEMATICS LEARNING
https://doi.org/10.1080/19477503.2023.2201115



# Unpacking Response Process Issues Encountered When Developing a Mathematics Teachers' Pedagogical Content Knowledge (PCK) Assessment

Martha L. Epstein, Hamza Malik, Kun Wang, and Chandra H. Orrill

STEM Education & Teacher Development, University of Massachusetts, Dartmouth, MA, USA

**ABSTRACT**

It is essential for items in assessments of mathematics' teacher knowledge to evoke the desired response processes – to be interpreted and responded to by teachers as intended by item developers. In this study, we sought to unpack evidence that middle school mathematics teachers were not consistently interacting as intended with constructed response (i.e. open-ended) items designed to assess their pedagogical content knowledge (PCK). We analyzed recent data derived from think-aloud interviews with 13 teachers involving 38 assessment items designed to tap PCK regarding proportional reasoning. Five key issues associated with undesired response processes were identified: (1) scenarios provided insufficient information, (2) content knowledge (CK) and PCK elements were confounded, (3) questions asked about the scenarios lacked specificity, (4) items contained distracting text and/or visual elements, and (5) differences between math education research and classroom teacher work cultures led to unanticipated interpretations of items. These issues were associated with teacher responses that were problematic (e.g. vague, off topic, etc.). In addition, we suggest that obtaining response process evidence is critical, and the way it is obtained may impact the average difficulty of the final pool of assessment items developed.

**KEYWORDS**

Assessment item development; construct-irrelevant responses; constructed response; pedagogical content knowledge; response process validity evidence

The mathematics education community has shown a growing interest in mathematics teacher knowledge assessment. Within the past 10 years, a *Journal of Research in Mathematics Education* monograph (Izsák et al., 2016) and a special issue of *The Mathematics Enthusiast* (Mosvold & Hoover, 2016) were dedicated to the topic. However, there is still much to do to understand what constitutes effective teacher knowledge assessment and how best to develop such instruments. Orrill and colleagues (Orrill et al., 2015) noted the challenges of developing assessment items extend from the broad theoretical (e.g., how to define teacher knowledge) to the elemental details (e.g., how to write items that assess what we intend them to assess). This paper focuses on the elemental details as we unpack factors that introduce construct-irrelevant responses and undermine desired response processes (RPs)—that is, whether assessment items are consistently interpreted and responded to by participants in the ways intended. Although the *Standards for Educational and Psychological Testing* (Standards; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) stress the importance of assessment items being interpreted as intended and not resulting in construct-irrelevant responses and there is also considerable expert guidance on the topic (e.g., Frey et al., 2005; Hogan & Murphy, 2007), there is little relevant supporting empirical research. The work in this paper extends the emerging response process evidence research by providing insights into factors that led to assessment items not performing as intended and, therefore, not measuring the intended construct. We were specifically interested in items written to

**CONTACT** Hamza Malik hmalik1@umassd.edu STEM Education & Teacher Development, University of Massachusetts, Dartmouth, MA, USA

measure teachers' pedagogical content knowledge (PCK) for teaching proportional reasoning. To help other assessment developers create more robust items, our purpose is to share empirical evidence of issues that led to construct-irrelevant responses and undermined desired response processes.

## Background

As discussed in the Standards (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), developing and implementing assessments in a way that supports the intended interpretation is complicated and multifaceted. A viable assessment should not only use theory to identify relevant constructs to be measured (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Hill et al., 2008; Orrill et al., 2015; Orrill & Cohen, 2016), but also, within those constructs, must differentiate across participants with different knowledge and skill (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Hill et al., 2008). Achieving meaningful differentiation across participants can be especially challenging for complicated constructs such as assessing mathematics teachers' PCK. Rowan et al. (2001) noted, "One difficulty we faced was developing items (and scenarios) that adequately tapped the full range of underlying "abilities" or "levels" of teachers' content and pedagogical knowledge . . . " (p. 16).

Orrill and Cohen (2016) noted that the first challenge in developing an assessment to measure mathematics teacher knowledge is identifying the constructs to measure. This should be specific; not overly broad. For example, Kim and Remillard (2011) identified four specific dimensions of curriculum-embedded knowledge of which mathematics teachers need to make sense. Even with this attention to detail in conceptualizing the domain, the researchers found it difficult to write items that adequately measured all four identified dimensions. In interviews with teachers, the researchers found certain items were simply confusing and other items were not always interpreted as the research team intended. The researchers gave the example that " . . . on a question asking to rank the order of difficulty in three word problems . . . , many teachers determined the order based on types of computation (i.e., addition, subtraction, etc.) while the intent was complexity of the part-whole models representing those word problems" (p. 25). Thus, identifying the specific constructs to measure is an important step but is not adequate to ensure the assessment measures those intended constructs.

Further elaborating on designing assessments for measuring teacher knowledge, Schilling and Hill (2007) stressed the need to identify elemental assumptions embodied in the assessment and measure those as part of the effort to validate assessments. They noted, for example, that in the *Learning Mathematics for Teaching assessment* (LMT) the ability to make relevant inferences about teachers' knowledge is based on the elemental assumption that "items reflect teachers' mathematical knowledge for teaching, and not extraneous factors such as test taking strategies or idiosyncratic aspects of the items (e.g., flaws in items)" (p. 79). This was operationalized in Hill et al. (2007) when they used clinical interviews to evaluate their elemental assumption that the LMT measured teachers' mathematical knowledge for teaching rather than other factors such as test-taking skills. Elemental assumptions are critical for assessment development, and testing those assumptions is essential for validating an instrument.

Schilling and Hill (2007) also argued for making structural assumptions explicit and measuring those. In Hill et al. (2007), the researchers tested the assumption that mathematical knowledge for teaching is a measurable construct. Their interview study confirmed this as they saw that teachers performed differently on items than mathematicians and other adults.

In this study, we tackle an important type of evidence that allows us to make inferences about the validity of an assessment instrument: response process evidence (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Response process evidence concerns the ways in which respondents interact with the items (e.g., Bostic et al., 2021) and whether or not, that is consistent with the item developers' intent. While

creating items that interact with participants as intended is an intuitive idea, achieving it is often challenging. Many researchers have noted that their assessment items generated unintended interpretations and reactions among participants. For example, Hill and her colleagues (Hill et al., 2008) noticed in developing the LMT, distractors that were blatantly "wrong" may not have worked as intended, as they were dismissed as "absurd" (p. 396). In addition, some items may have engaged teachers in a way that teachers "learned" from the test, which undermined such items' ability to tap teachers' extant knowledge. *En route* to developing the *Curriculum Embedded Mathematics Assessment*, other researchers found that in their piloted multiple-choice items "teachers saw things differently and used their own interpretations and reasoning to support their answers" (Kim & Remillard, 2011, p. 24).

Assessment design literature includes guidance for avoiding unintended interpretations and construct-irrelevant responses for multiple choice/selected response items (e.g., Haladyna & Rodriguez, 2013; Haladyna et al., 2002; Hogan & Murphy, 2007). Dolan et al. (2013) recommended items should be clear, instructions should convey the scope and intent of a task, all necessary steps should be noted, and the context should be clearly defined. Mullis and Martin (2013) reported on both selected and constructed response (i.e., open-ended) item-writing for content assessments, including TIMSS and PISA. They echoed items should be clear, include key details, and specify the level of detail expected in a response. They suggested avoiding overly-complicated contexts and distractions and stressed labeling graphics clearly.

While the above recommendations are an excellent starting point, their guidance for developing items with desired RPs stays relatively broad and high level. Moreover, most advice generally reflects conventional practice or expert opinion, not empirical findings about ways of achieving desired RPs (Frey et al., 2005; Hogan & Murphy, 2007). In fact, Frey et al. (2005) noted of their 40 recommendations on writing assessment items, only four were supported by research. This suggests a considerable opportunity for empirical evidence to inform guidelines for effective item writing.

### Response Process Research

Garnering RP evidence has gained recent attention in STEM education (e.g., Deng et al., 2021; Padilla & Benítez, 2014). Hill and colleagues (Hill et al., 2008; Schilling & Hill, 2007; Schilling et al., 2007) stressed the importance of using qualitative methods, such as think-alouds and retrospective cognitive interviews, to investigate individual assessment item performance to determine if items are performing as planned. Likewise, using methods such as think-alouds (Bostic, 2021) and whole-class think alouds (Bostic et al., 2021), other researchers have sought to determine the match between test developer intent and test taker interpretation and to garner insights into any mismatches.

Examples of recent RP efforts include Bonner et al. (2021) use of interviews to determine RPs evoked in an assessment of computational thinking and self-regulatory processes. They highlighted that using interviews helped illuminate how and why participants selected particular responses and led developers to conclude their tasks measured the intended constructs. Mo et al. (2021) used think-alouds to determine the alignment between parallel selected and constructed response items, finding when teachers were confused about an item, selected and constructed responses tended to diverge. Most relevant to our study, Zhai et al. (2021) used a think-aloud methodology to investigate teacher responses to video clips. They found teachers do not necessarily focus on the intended video clip elements and concluded assessment developers should be thoughtful about what scenario elements are, in fact, included in such PCK assessments. Building from the extant research noted above, our work in this paper focuses on the need for empirically based findings regarding factors that negatively impact desired RPs of constructed response items designed to measure mathematics teacher PCK. We were guided by the following research question: What are common characteristics of piloted teacher mathematics PCK assessment items that exhibited unintended RPs? In investigating this, we hope to provide empirical evidence for some of the item-development conventions. We also will share "what went wrong" details for the benefit of other assessment developers.

## Methodology

### Context

The data analyzed for this project were pilot data collected as part of a larger assessment development and validation process. The assessment comprised items to measure mathematics teachers' content knowledge (CK) or PCK (Shulman, 1986) about proportional reasoning. The constructs to be included in the CK and PCK middle school teacher assessment were informed by extant research on elements of CK and PCK as distinguishable constructs (e.g., Baumert et al., 2010). The findings reported here are based on a qualitative analysis of middle school teachers' responses to PCK items. The PCK assessment items were all open-ended or had open-ended components. All involved asking teachers questions about realistic classroom scenarios and typically comprised two components: a scenario involving a classroom discussion or student work (scenario) and one or more questions teachers were asked about that scenario (question). The decision to use open-ended items was driven by the research team's desire to measure teachers' understandings of student work and classroom situations and based on Kersting and her colleagues' (N. B. Kersting et al., 2012; N. Kersting, 2008) successful work with similar items. Consistent with the Standards (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) (i.e., Standard 4.8), prior to being piloted with middle school mathematics teachers, assessment items were refined using expert feedback from a mathematician and a mathematics educator on the study's advisory board.

### Participants

Data were collected from a convenience sample of 13 middle school mathematics teachers (nine female, four male). Of teachers who provided ethnicity information, nine identified as White; one as Asian or Pacific Islander; one as Black; and one as Hispanic. Six teachers had taught math for eight or more years, six teachers for three to four years, and one teacher for one year. Nine teachers had graduate-level education degrees (e.g., MAT), and three of the 13 teachers majored in mathematics as undergraduates. All names reported below are pseudonyms.

### Data Collection

To accommodate the large number of original items (53), items were spread across five forms, with each form containing between eight and 13 assessment items of varying lengths and designed to be completed in under an hour. Four forms were completed by 10 teachers, and one form was completed by nine teachers. As seen in Table 1, most teachers (seven out of 13) completed all five forms, with three teachers completing three forms, two teachers completing two forms, and one teacher completing one form. After completing each assessment form, teachers were invited to complete additional assessments. Given time commitments, however, not all teachers were able to complete all forms. Thirty-eight of the items were selected for in-depth qualitative analysis. Those not selected for further analysis were parallel versions of items already selected for further analysis. On average, for an individual teacher who participated in this project, responses to 29 items were qualitatively explored in depth, with the number of items explored in depth per individual teacher ranging from seven to 38.

Teachers were provided pilot items via an online Qualtrics survey. Teachers' typed responses were immediately available and reviewed by interviewers prior to the teacher's follow-up interview which was typically scheduled for within 24 hours of survey completion. These interviews were recorded, transcribed using Zoom, and manually edited for transcription errors. The purpose of the follow-up interviews was to identify items exhibiting construct-irrelevant responses, investigate the likely reasons for such responses, and propose relevant changes. The interviews used a pre-established protocol regarding teachers' responses and areas of concern or confusion. Examples of questions include: "Was it clear what was happening in the video?" "Is the question clear? If not, how could it be

**Table 1.** Number of forms completed and items qualitatively analyzed per participant.

| | Number of forms teacher completed | Total items analyzed per teacher |
|---|---|---|
| Participants (pseudonyms) | | |
| Christie | 5 | 38 |
| Emma | 5 | 38 |
| Jane | 5 | 38 |
| Harry | 5 | 38 |
| Liz | 5 | 38 |
| Shareen | 5 | 38 |
| Kate | 5 | 38 |
| Britney | 3 | 26 |
| Elton | 3 | 26 |
| Lydia | 3 | 26 |
| James | 2 | 12 |
| Ruth | 2 | 12 |
| Morgan | 1 | 7 |
| Average number of forms completed per teacher | 3.8 | |
| Average number of items analyzed per teacher | | 28.8 |

improved?" During the Zoom interview, their Qualtrics survey was available to teachers, and classroom videos that were part of the original survey were offered for re-review at a participant's request.

## *Data Analysis*

The analysis plan first relied on deductive coding concerning participants' types, depth, and correctness of answers. For example, one survey item involved teachers reviewing a video involving a lesson on proportions. In the video, students jumped for a short period of time, and their teacher instructed them to project the number of jumps to a longer time period. Participants were asked a question similar to, "How well did the task support students' understanding of equivalent ratios?" The deductive codes called for determining if a teacher had discussed whether the jumping rate would be consistent over different time spans and if the discussion referenced scaling, unit rate, or something else. Coders quickly realized this coding structure did not align with teachers' responses and did not capture important aspects of teachers' answers. This triggered a reevaluation of how best to capture relevant information in teachers' responses. A subset of items was reviewed inductively to determine the nature of teacher responses, and this led to the development of an inductively derived coding rubric. Five researchers, including the projects' two principal investigators were involved in developing and refining the inductive rubric. All data from all participants was then analyzed with this rubric.

The new rubric focused on response process evidence and assessed if the participant understood the question as the research team intended. If a teacher did not appear to have understood the question as it was intended, coders used cues within a teacher's response to infer what the question was that a teacher seemed to be answering and to describe what it was about the teacher's response that suggested the teacher was not responding to the item as intended (e.g., an answer was overly vague, off topic, characterized solely by paraphrasing of provided information, etc.). If a teacher interpreted the question as intended, then answers were reviewed to determine the quality of the response, including correctness, where relevant, and the type of mathematical understanding evident in the answer.

Three researchers independently reviewed each item to make each of the above determinations. In addition, interviews and surveys were reviewed to find instances in which teachers explicitly noted confusion or concern about some aspect of the item or in which teacher responses or feedback suggested confusion or concern. Within a given survey form, data were reviewed for each teacher. As the primary goal of this stage of the research was to as expeditiously as possible identify potentially problematic items, a statistical analysis of interrater reliability was not performed. Instead, after each researcher independently analyzed a teacher's survey and follow-up interview, the researchers met to discuss their item analysis for that teacher. Divergent views triggered a re-review of a teacher's data until the research team reached 100% consensus regarding how to record the work for that teacher and

what RP issues to flag to the broader research team. After coding and recording was complete, the researchers focused on how each item worked across teachers. Items that were interpreted as intended by all teachers were considered to not have RP issues. Items for which at least three teachers did not interpret the item as intended or for which teachers explicitly noted the item was confusing or distracting were analyzed for recurring patterns that might illuminate reasons items might not be interpreted as intended. We identified reasons for RP issues that occurred across most of the participants and identified evidence that best showcased these issues.

## Results

Items with RP issues tended to have one or more of the following characteristics:

(1) Scenarios provided insufficient information – some scenarios did not provide enough information to result in consistent item interpretation or in-depth PCK-informed answers.
(2) CK and PCK elements were confounded – some PCK items depended on a foundational CK component, but if teachers did not invoke the intended CK component, the assessment item could not measure the intended PCK construct.
(3) Questions lacked specificity – some questions were not specific enough to be interpreted in a consistent way across teachers.
(4) Items contained text and/or visual elements that distracted teachers from the item's intent.
(5) Differences between math education research and classroom teacher work cultures resulted in the unintended interpretation of items.

These issues and key characteristics of participants' responses that flagged these issues are summarized in Table 2 and discussed in greater detail below. As can be seen in Table 3, most participants (nine out of 13) encountered each of these issues at least once. Perhaps not coincidentally, the four teachers (Lydia, James, Ruth, and Morgan) who did not encounter all of these issues completed only between one and three of the five assessment forms, and so had fewer opportunities to encounter problematic items. Because the assessment discussed here is still in development, actual items are not yet released. To provide examples of the issues encountered, parallel items are provided that are similar to the actual items used in the pilot testing, and relevant modifications were made to examples of excerpts from participants.

### *Scenarios Provided Insufficient Information*

#### *Overview of the Issue*

Participants noted some scenarios contained too little information about students or classroom situations. At times, teacher-participants wanted to know more about what a student knew before

**Table 2.** Characteristics of items with unintended RPs And characteristics of teachers' answers to them.

| | Characteristics of items with response-process issues | | | | |
|---|---|---|---|---|---|
| Teacher answer characteristics | Scenarios provided insufficient Information | CK and PCK were confounded | Questions lacked specificity | Items contained distracting text/visuals | Difference between cultures |
| Vague | ✓ | | ✓ | ✓ | |
| Needed to make assumptions to answer | ✓ | | ✓ | | |
| Focused on unintended topics | | ✓ | ✓ | ✓ | |
| Paraphrased information provided | | | ✓ | | ✓ |

**Table 3.** Characteristics of items with unintended RPs encountered by participants.

| | | Characteristics of items with response-process issues | | | | |
|---|---|---|---|---|---|---|
| Participants (pseudonyms) | Number of forms completed | Scenarios provided insufficient information | CK and PCK were confounded | Questions lacked specificity | Items contained distracting text/visuals | Difference between cultures |
| Christie | 5 | ✓ | ✓ | ✓ | ✓ | ✓ |
| Emma | 5 | ✓ | ✓ | ✓ | ✓ | ✓ |
| Jane | 5 | ✓ | ✓ | ✓ | ✓ | ✓ |
| Harry | 5 | ✓ | ✓ | ✓ | ✓ | ✓ |
| Liz | 5 | ✓ | ✓ | ✓ | ✓ | ✓ |
| Shareen | 5 | ✓ | ✓ | ✓ | ✓ | ✓ |
| Kate | 5 | ✓ | ✓ | ✓ | ✓ | ✓ |
| Britney | 3 | ✓ | ✓ | ✓ | ✓ | ✓ |
| Elton | 3 | ✓ | ✓ | ✓ | ✓ | ✓ |
| Lydia | 3 | ✓ | ✓ | ✓ | - | ✓ |
| James | 2 | ✓ | - | ✓ | ✓ | - |
| Ruth | 2 | ✓ | ✓ | ✓ | ✓ | - |
| Morgan | 1 | ✓ | ✓ | ✓ | - | - |

*Note*. ✓ indicates this RP issue was encountered by the participant.

starting work on a problem; the grade level of a classroom; or where in the unit a lesson occurred. Item scenarios that lacked enough relevant information often resulted in participants providing vague or overly general answers or making assumptions to allow them to provide more precise answers. Yet, if different teacher-participants make different assumptions, that essentially changes the question being asked from participant to participant.

### *Example of Scenario with Insufficient Student Information*

A scenario similar to that shown in Figure 1 was identified as providing insufficient student information. In this item, participants viewed a short video of a seventh-grade class discussing the data provided. In the video, a student made a comment that Martina's data did not show a proportional relationship because "None of them can be divided by the same number." The participants were asked to comment on that student's understanding based on the remark.

7th grade students are working to determine whether or not the number of laps Demonte and Martina run is proportional to their elapsed time. In this video, the teacher has asked students to explain their thinking.

| **Demonte's Time (s)** | 138 | 207 | 345 |
|---|---|---|---|
| **Laps** | 2 | 3 | 5 |

| **Martina's Time (s)** | 160 | 255 | 450 |
|---|---|---|---|
| **Laps** | 2 | 3 | 5 |

*Please view the ENTIRE video clip and then answer the question below. You may watch the video more than once.*

**Q: Comment on the student's understanding based on what he saidabout Martina's situation, "None of them can be divided by the same number."** ***Be as specific as possible.***

**Figure 1.** Example of scenario with insufficient student information.

In this video, the 7th grade teacher just solved the following problem with the class, and they are discussing what proportional means.

The bakery sells donuts for $3 each. Find the cost of 1, 2, 3, 6, and 12 donuts.

| **# of donuts** | 1 | 2 | 3 | 6 | 12 |
|---|---|---|---|---|---|
| **Cost** | 3 | 6 | 9 | 18 | 36 |
| **Unit rate** | 3/1 | 3/1 | 3/1 | 3/1 | 3/1 |

*Please view the ENTIRE video clip and then answer the questions below.*
*You may watch the video more than once.*

**Q. Would you have led the class discussion in this video differently to support students' understanding of proportional relationships?**

**Figure 2.** Example of scenario with insufficient classroom information.

Five out of 10 participants mentioned that this scenario contained insufficient information about the student. For example, Shareen noted, " . . . . [the student's] answer did not indicate whether [the student] understood that Martina's time per lap was not proportional . . . I would have asked the appropriate questions so that I could further understand . . . " Similarly, Lydia said, "I think it was difficult to understand what [the student] truly understood. . . . Did [the student] understand, yes or no? . . . I don't think you can answer that question." Lydia went on to say that the student used unit rate reasoning and responded to the question from that perspective, but commented that she "made an assumption about what he was sort of thinking . . . " We do not know the degree to which other teachers who answered this question based their answers on an assumption as Lydia indicated she did.

### *Example of Scenario with Insufficient Classroom Information*

Three of 10 participants explained that a scenario like the one in Figure 2 did not provide enough information about the classroom to answer the question. For this item, participants watched a short video of a seventh-grade classroom discussing proportions. Then they were asked, "Would you have led the class discussion in this video clip differently to support the student's understanding of proportional relationships?" Some teachers, like Jane, wanted more contextual information. She wondered, " . . . how much time would they have to work with the class . . . What is the objective of the lesson? . . . you don't know what the class has already done on the topic for that day or the unit." Perhaps responding to the insufficient contextual classroom information provided, six of 10 teachers' responses to this item were vague.

### *Why Insufficient Information is Problematic*

Scenarios with insufficient information about classroom situations or students resulted in vague teacher responses. Such responses provide no insight into the PCK the assessment was designed to measure. In addition, when teachers make assumptions about a situation or student to compensate for missing information, it potentially changes the assessment item for each teacher. For these reasons, items that featured scenarios with insufficient classroom and/or student information tended to have RP issues – they were not interpreted and responded to as item developers intended.

## *Confounding CK and PCK*

### *Overview of the Issue*

When designing items, intentional steps were taken to separate CK from PCK. However, some piloted items were found to rely on teachers first invoking a correct understanding of a CK component to

answer the PCK component of interest for that item. In these cases, if teachers did not notice or misunderstood the actual CK embedded in an item, their PCK response would not support inferences about the underlying PCK construct the item purported to measure.

### *Example*

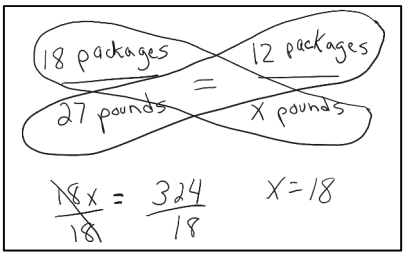
Figure 3 shows an item that problematically conflated CK and PCK. It was developed to measure PCK strategies relevant to the Grade-6 Common Core State Standards for Mathematics (CCSSM) (National Governors Association Center for Best Practices (NGA) & Council of Chief State School Officers (CCSSO), 2010) standard involving differentiating proportional (i.e., $y = kx$) from non-proportional situations. The task presented an inverse proportion (i.e., $y = k/x$), and the student's use of cross multiplication – a strategy appropriate for solving proportional problems, but not inverse proportional problems – was therefore problematic. The item's goal was not to assess teachers' personal CK involving recognizing and differentiating between proportional and inversely proportional situations. Instead, the item's goal was to invoke teachers' PCK strategies for supporting a student who demonstrated a common incorrect interpretation of the situation (de Bock et al., 2002). Although the item's wording did not explicitly note the CK issue at play in the student's work – assuming the situation was proportional when it was not – teachers needed to be aware of this specific issue for the item to solicit the PCK strategies of interest.

Only three of 10 teachers, however, interpreted the entire item as intended. This subgroup of participants identified the student's work was incorrect, as the situation was not a direct proportion. These participants invoked relevant PCK strategies to redirect the student. In contrast, seven out of 10 teachers either interpreted or assumed the student's work was correct and answered the question based on that interpretation or assumption. For example, Liz responded by discussing additional ways to think about proportional problems, "I would [use] something more visual like a double number line to represent this question. One number line representing # of [packages] and the other representing the

At the end of a proportional reasoning unit in which students learned about situations that are proportional or nonproportional, the following problem was used to assess students' understanding:

> A company is considering new packaging for its cocoa. The original packaging involved equally splitting 27 pounds of cocoa across 18 packets. If 27 pounds of cocoa were equally split among 12 packets, how many pounds of cocoa will each packet contain?

One student's work is shown below. Use it to answer the following question.




**Q: If you were the teacher, what strategy would you use next to support the student's understanding of proportions? *Be as specific as possible.***

**Figure 3.** Example of confounded content knowledge (CK) and pedagogical content knowledge (PCK).

pounds of [cocoa]. I would establish the proportionality constant . . . .” Responses based on the view the student work was correct did not support the goal of the item – to understand how a teacher would redirect a student who has made the mistake featured. Responses to this item also did not facilitate valid CK inferences, as it was not possible to differentiate between teachers who had assumed the work was correct and those who actively concluded the task was solved correctly.

When analyzing teachers’ responses to and feedback about this item, we realized underlying assumptions may be at play. Teachers may assume test writers would make it clear if the teacher is being asked to determine the correctness of student work. If not, teachers may assume the student work featured in an item is correct unless it is explicitly flagged as incorrect. As evidence of this orientation, one teacher who recognized the student’s work was mathematically incorrect seemed flummoxed and sought verification for her assessment of the student’s work. She said, “So, the student work is wrong. Right? . . . I had to kind of be like, Wait! What? And like kind of be like, am *I* doing this right?” The many teachers who may have assumed the student work was correct may not have even assessed the appropriateness of the student’s response as part of their process in answering, and this led to unintended RP issues.

#### *Why Confounding CK and PCK is Problematic*

When assessing PCK depended on teachers first recognizing the correctness of students’ work, if a teacher did not correctly ascertain the correctness of the student’s work, the item could not measure the intended PCK. Using the example above, once a teacher accepted the student’s work as correct, the assessment item fundamentally changed from measuring PCK strategies for redirecting a student with a misunderstanding about proportions to measuring PCK about how to broaden a student’s understanding of ways to solve a direct proportion. These interpretations are different; thus the item has RP issues, and any data resulting from the item would likely be a confounded mixture of two different understandings of the item.

### *Questions Lacking Specificity*

#### *Overview of the Issue*

A challenge with open-ended items is to balance specificity and generality. Once presented with a scenario, questions about the scenario that are too specific may telegraph a response, resulting in teachers giving predictable and nearly identical answers. Yet, questions that are not specific enough may not yield answers that reliably address the intended issue. In our study, we found some questions were so vague that teachers reported they were unsure what question was, in fact, being asked. In other cases, teachers did not explicitly express confusion about the question, but teachers’ responses suggested many interpretations of the same question.

#### *Example of Uncertainty About What Question Was Being Asked*

One type of question that lacked sufficient specificity was when teachers were asked “what can you say about” or to “comment on” a student’s understanding. Several teachers flagged this wording as problematic. Jane noted, “ . . . when you just say ‘comment,’ you know . . . It’s kind of like we’re casting a broad net. . .without getting to the scope of what you’re trying to study . . .” Harry said, “I don’t exactly know what they’re [the item developers] going for here.” Lydia told us that even though she was unclear how to interpret “comment on” a student’s understanding, she would still try to “ . . . figure a way to answer it,” suggesting she was making assumptions to do so. In addition to making assumptions, such questions often resulted in vague answers or simply paraphrasing what the student did with no additional analysis.

#### *Example of a Lack of Question Specificity That Triggered Multiple Interpretations*

Questions that lacked specificity often led to multiple interpretations. For example, Figure 4 shows an item in which teachers watched a video of a classroom discussion and were asked “ . . . how

7th grade students are working to determine whether or not the number of laps Demonte and Martina run is proportional to their elapsed time. In this video, the teacher has asked students to explain their thinking.

| **Demonte's Time (s)** | 138 | 207 | 345 |
|---|---|---|---|
| **Laps** | 2 | 3 | 5 |

| **Martina's Time (s)** | 160 | 255 | 450 |
|---|---|---|---|
| **Laps** | 2 | 3 | 5 |

*Please view the ENTIRE video clip and then answer the question below. You may watch the video more than once.*

**Q: If you were to take over the class at the end of the video, how would you conclude this discussion to help students further develop their understanding of proportional and nonproportional situations? *Explain your choice of strategies.***

**Figure 4.** Example of an item that lacked specificity and resulted in multiple interpretations.

would you conclude the discussion to help students further develop their understanding of proportional and nonproportional situations … ?" The item's objective was to measure PCK regarding what mathematical concepts a teacher foregrounded. Yet, only four of nine teachers exclusively interpreted the question this way. One of these teachers, Liz, noted, "I would summarize by explaining how [the unit rates] represent different constants, and that is vital to a proportional relationship." Two teachers, including Kate, answered in a way that mentioned both math and relevant pedagogical choices. She noted she would highlight the unit rates for Martina, stressing that they are not equivalent (e.g., 80/1 is different from 85/1), then "I would turn the discussion back to the students … " Yet, three teachers answered in a way that focused only on pedagogical choices for soliciting relevant mathematical concepts, without mentioning what those mathematical concepts were. Specifically, James answered he would repeat and rephrase a correct student's answer from earlier and ask a student who previously answered incorrectly to revisit the question. Similarly, Harry offered, "I would have them work in groups." While all the answers above are understandable interpretations of the question posed, they do not all align with the item developers' intent to solicit feedback about what mathematical concepts a teacher would highlight.

### *Why Questions That Lack Specificity are Problematic*

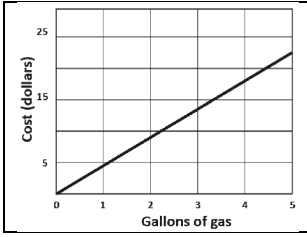
As with vague item scenarios, items with questions that lacked specificity also tended to have RP issues. Based on their answers, it was clear that teachers did not interpret such items as we had intended. Some teachers offered off-topic or vague answers that did not illuminate the PCK at issue, some teachers simply paraphrased the information provided, and others based their answers on assumptions, resulting in different teachers essentially answering different questions.

## *Items That Contained Distractions*

### *Overview of the Issue*

Some items may inadvertently include elements that distract teachers from responding to the item as intended by item developers. For items to elicit the intended RPs, it is important to avoid potentially

A middle school class learned to use graphs and ratio tables to represent proportional relationships. Some students did not understand how these two representations are connected and how proportionality characteristics can be seen in both representations.




| Gas (gallons) | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Cost (dollars) | 4.45 | 8.90 | 13.35 | 17.80 | 22.25 |

**Q: How could you help students see how the characteristics of proportions are shown in both representations?** ***Be as specific as possible.***

**Figure 5.** Example of an item with distraction linked to absent/suboptimal information.

distracting elements so that teachers' attention is not diverted away from how item developers intended the item to be interpreted. Our review revealed instances when incidental elements of items were associated with unintended distractions that may have resulted in unintended RPs. Some distractions were linked to absent or suboptimal information while others were linked to included, but distracting, information.

### *Example of Distraction Linked to Absent/Suboptimal Information*

In one example, similar to Figure 5, participants seemed distracted by suboptimal information in the problem. The graph and table in this item featured a proportional relationship between gallons of gas and cost. This item's goal was to solicit PCK regarding helping students build a conceptual understanding of proportions. However, some teachers focused on the graph's suboptimal labeling and scaling. Four of nine teachers discussed these suboptimal qualities. For example, Harry noted, " . . . I would pick a graph that had . . . more specific number line marks or access marks so that kids could make that relation a little better . . . " Among the four teachers who voiced concern about the graph, Harry was the only one who also answered in a way that was aligned with the item's intent – to elicit PCK regarding developing a relevant conceptual understanding of proportions. In addition to these four teachers who showed explicit evidence of being distracted by suboptimal elements of the graph, two additional teachers responded with vague answers, and it was unclear how these teachers had interpreted the item.

### *Example of Included Distracting Information*

The unintuitive, unfamiliar numeric values used in the item shown in Figure 6 seemed to distract some participants from eliciting PCK regarding supporting students' conceptual understanding of proportions. Highlighting this distracting element, Liz, noted, " . . . there is no [5 inches by 6 ¼ inches] . . . So, why did you choose that?. . . Why did you choose to use weird numbers?" Four of 10 teachers gave vague answers to this item.

### *Why Distracting Information is Problematic*

Based on our interviews, once a participant was distracted, there was no guarantee they would answer in a way that was aligned with the item's intent. Distracting information likely diverted the teachers in this study's attention away from the content of interest to the item developers. Hence distraction tended to be associated with items eliciting unintended RPs.

Mrs. Benson asked her middle school students to solve the following problem:

Shelby wants to shrink a photo to add to a collage. The original photo was 8 inches by 10 inches. Which of the following dimensions maintains the same proportions as the original photo?

A) 5 inches by 7 inches B) 5 inches by 6 ¼ inches

**Q: If your students were unsure about this problem, how could you help them understand which picture is proportional to the original? *Be as specific as possible.***

**Figure 6.** Example of an item that included distracting information.

## *Difference Between Math Education Research and Classroom Teacher Work Culture*

### *Overview of the Issue*

We found an interesting issue in our item analysis that we believe reflects differences between the work cultures of mathematics education research and mathematics teaching. Specifically, items that were intended to explore teachers' understandings of students in general, or certain groups of students, were interpreted as being questions about the specific, individual student described in the scenario. Because we did not account for the ways the research community uses single examples as launching points for generalizable discussion, we did not anticipate that teachers would not see those items the same way. Our questions did not specify that we wanted teachers to use their broader PCK about students to anticipate what students similar to the one described *likely* understood. Hence, such items triggered unintended RPs, as teachers interpreted such items in a way we had not anticipated or intended.

### *Example*

An example of an item that encapsulated the above issue involved teachers viewing a video clip in which a student explained why he thought two ratios (e.g., 14/18 and 28/36) were equivalent. Teachers were then asked to "Comment on [the student's] understanding based on his method, 'If you divide down, you'll get the same answer.'" Only one of 10 teachers interpreted the item in such a way that clearly invoked PCK to describe what else the student *likely* understood. This teacher, Harry, explained, "I'm making some predictions on my own because of having numerous kids like that and asking very similar questions over multiple years of teaching." Two teachers stopped at noting their suspicions based on the information provided. Specifically, Lydia noted, " . . . the way he phrased it, my assumption would be that . . . he memorized the method and not necessarily why or what it meant. But . . . I can't say he doesn't understand based on one [example]." Most teachers (seven out of 10) simply paraphrased the information we provided to them. For example, Christie paraphrased the student's answer, noting that the student, " . . . understands that you can check for equivalent fractions by simplifying them." These teachers seemed to adopt a view of the hypothetical student as a unique individual, whose other knowledge remained unknown rather than using their PCK to predict what a student who made such a statement was also *likely* to know about proportions.

### *Why Not Recognizing Differences Between Math Education Research and Classroom Teacher Work Culture Was Problematic*

By not being sensitive to the different cultures of mathematics education research and mathematics teaching, we inadvertently phrased questions in a way that resulted in unintended RPs. Specifically, the

way we worked these questions did not communicate that we wanted teachers to use their PCK to project what other knowledge a featured student was also likely to have. This led to teachers simply paraphrasing the information that was provided in the prompt.

## Discussion

For an assessment to measure the construct of interest, many pieces must fall into place including fine-grained details of individual item development. What might seem a trivial detail in wording or supporting visuals can undermine item validity by confusing, sidetracking, or misdirecting a participant. Ideally, researchers, item developers, and teachers involved in education assessment would appreciate the subtlety and complexity of writing items that elicit the desired RPs and interact with participants as intended. Yet, while the Standards (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) stress evaluating and revising items to identify and address item-specific issues, such issues often go unaddressed (Schilling et al., 2007), resulting in items that do not elicit desired RPs. This study has been our attempt to highlight challenges we faced in writing items that participants interpreted and responded to as item developers intended. By analyzing and sharing "what went wrong" in our piloted items we hope to share assessment-item writing challenges and to offer key issues to keep in mind to write items that interact with participants as intended.

Our empirically based findings lend support to the expert opinions about writing good assessment items and are likely to apply across a variety of settings. Below is a summary of our research-based conclusions and recommendations for developing items that elicit desired RPs.

(1) Item scenarios should provide sufficient information, and item questions need to be clear and specific. This echoes Dolan et al.'s (2013) recommendation for context to be clearly identified and task boundaries well defined.
(2) Anything that distracts from the intended RPs, such as a poorly labeled graph, should be avoided. Our item that accidentally included a poorly labeled graph diverted teachers' attention from the item's purpose, compromised the intended RPs, and led to construct-irrelevant responses. This is consistent with Mullis and Martin's (2013) recommendations to label graphics properly to avoid distractions. Distraction can also be caused by using real-life scenarios. If, as Zhai and colleagues (2021) also encountered, real-life scenarios contain distracting, off-task elements, the scenarios risk being realistic, but unsuccessful. Worth noting, it should not be assumed that eliminating a distraction will result in the intended RPs, as the distraction may be masking other issues, and consistent with the Standards (AERA et al., 2014), this should be evaluated by retesting the item.
(3) Avoid structuring items such that one part of the item is conflated with another part of the item. This was the root of the issue we uncovered regarding CK-PCK conflation. Items structured this way risk unintended RPs since a test developer cannot know if someone answering the item attended to each part of the item as needed.
(4) Keep the perspective of the test taker in mind (in our case, middle school mathematics teachers) when developing items. Consistent with other researchers, our study concluded that it is best not to ask teachers to "infer what a hypothetical teacher or student knew or was thinking" (Hill et al., 2017, p. 92). Among the reasons this may be problematic is the different cultural norms likely leading to the mathematics education research and mathematics teacher communities interpreting such questions differently.

Ultimately, to provide insights into how participants make sense of items, interviewing people in the target population via think-aloud interviews or other relevant methodologies such as those used by Bostic and colleagues (Bostic et al., 2021; Bostic, 2021) as well as Hill and colleagues (e.g., Hill et al., 2007) is critical. We suggest how a series of pilot interviews are scheduled and used

matters. For example, if a research team relied solely on feedback reviewed after a single round of pilot interviews with the target population was completed, researchers would be likely to accept only those assessment items with strong initial evidence of having the desired RPs. And items that seem to require little, if any modifications in order to elicit the desired RPs are likely to be associated with simpler, easy-to-communicate concepts. Consistent with this, Kim and Remillard (2011) observed in their assessment efforts that not all constructs are equally easy to communicate. Hence, if items are only subjected to one round of vetting, items involving easy-to-communicate, simple ideas will tend to be accepted at a much higher rate than items involving challenging-to-communicate, more complex ideas. This has implications for the pool of items' ability to discriminate well across teachers since most teachers would be able to answer these easier items correctly.

Rather than relying on a one-shot approach to identifying assessment item issues among the target population, we assert that piloting items should be an iterative process. Ideally, this process would allow time for multiple iterations of items. This approach would allow items measuring more complicated ideas to be honed, rather than being discarded, because of a lack of time or funding available to revise and retest them.

We also suggest honing items internally prior to formal vetting by the target population. Knowledgeable others not involved in item development may help flag many issues such as ambiguities, distractions, and multiple interpretations, allowing items piloted with the target population (in our case, teachers) to contain fewer RP issues. In retrospect, while early feedback from mathematics and mathematics education research experts was very helpful, it may have been useful to obtain input from classroom teachers earlier. A challenge we noted at this and other stages of obtaining feedback is to not dismiss identified issues as trivial (e.g., "The question isn't about the graph, so it doesn't really matter how it is labeled."), but to take all noted areas of possible confusion, misdirection, and miscommunication seriously.

The implications of our work for researchers and test developers are discussed above, but its applicability for teachers is less clear. Certainly, teachers can attend to our guidance on item clarity and specificity (both scenarios and questions), avoiding conflated two-part items, avoiding distractions, and keeping the perspective of the test taker in mind. While we believe that, ideally, teachers should care about RP evidence, we recognize that they are not in a position to test RPs because of the demands on their time. It seems worth thinking about as a community how to support teachers in writing items that elicit the desired RPs.

We encourage others to take item-level assessment issues seriously and to share findings of characteristics of items with RP issues. By doing so, we can build our empirically informed knowledge about factors that impact assessment developers ability to elicit the desired RPs and measure the intended constructs. This should be a helpful additional perspective to testing experts' recommendations currently available in textbooks and elsewhere.

## Acknowledgments

The researchers would like to acknowledge Yasemin Copur-Gencturk for her support with this research.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

The work was supported by the National Science Foundation under Grant Number 1813760.

## References


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

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Tsai, Y. M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*(1), 133–180. https://doi.org/10.3102/0002831209345157

Bonner, S., Chen, P., Jones, K., & Milonovich, B. (2021). Formative assessment of computational thinking: Cognitive and metacognitive processes. *Applied Measurement in Education*, *34*(1), 27–45. https://doi.org/10.1080/08957347.2020.1835912

Bostic, J. D. (2021). Think alouds: Informing scholarship and broadening partnerships through assessment. *Applied Measurement in Education*, *34*(1), 1–9. https://doi.org/10.1080/08957347.2020.1835914

Bostic, J. D., Sondergeld, T. A., Matney, G., Stone, G., & Hicks, T. (2021). Gathering response process data for a problem-solving measure through whole-class think alouds. *Applied Measurement in Education*, *14*(1), 46–60. https://doi.org/10.1080/08957347.2020.1835913

de Bock, D., Van Dooren, W., Janssens, D., & Verschaffel, L. (2002). Improper use of linear reasoning: An in-depth study of the nature and the irresistibility of secondary school students' errors. *Educational Studies in Mathematics*, *50*(3), 311–334. https://doi.org/10.1023/A:1021205413749

Deng, J. M., Streja, N., & Flynn, A. B. (2021). Response process validity evidence in chemistry education research. *Journal of Chemical Education*, *98*(12), 3656–3666. https://doi.org/10.1021/acs.jchemed.1c00749

Dolan, R. P., Burling, K., Harms, M., Strain-Seymour, E., Way, W. D., & Rose, D. H. (2013). *A universal design for learning-based framework for designing accessible technology-enhanced assessments*. Pearson Assessment Research Report. http://images.pearsonclinical.com/images/tmrs/DolanUDL-TEAFramework_final3.pdf

Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education*, *21*(4), 357–364. https://doi.org/10.1016/j.tate.2005.01.008

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education*, *39*(4), 372–400. https://doi.org/10.5951/jresematheduc.39.4.0372

Hill, H. C., Dean, C., & Goffney, I. M. (2007). Assessing elemental and structural validity: Data from teachers, non-teachers, and mathematicians. *Measurement*, *5*(2–3), 81–92. https://doi.org/10.1080/15366360701486999

Hogan, T. P., & Murphy, G. (2007). Recommendations for preparing and scoring constructed-response items: What the experts say. *Applied Measurement in Education*, *20*(4), 427–441. https://doi.org/10.1080/08957340701580736

Izsák, A., Remillard, J. T., Templin, J. (Eds.). (2016). Psychometric methods in mathematics education: Opportunities, challenges, and interdisciplinary collaborations. *Journal for Research in Mathematics Education monograph series*. National Council of Teachers of Mathematics.

Kersting, N. (2008). Using video clips of mathematics classroom instruction as item prompts to measure teachers' knowledge of teaching mathematics. *Educational and Psychological Measurement*, *68*(5), 845–861. https://doi.org/10.1177/0013164407313369

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal*, *49*(3), 568–589. https://doi.org/10.3102/0002831212437853

Kim, O. K., & Remillard, J. T. (2011). Conceptualizing and assessing curriculum embedded mathematics knowledge. In *Annual meeting of the American educational research association*. LA. https://icubit.gse.upenn.edu/sites/default/files/CEMA.pdf

Mo, Y., Carney, M., Cavey, L., & Totorica, T. (2021). Using think-alouds for response process evidence of teacher attentiveness. *Applied Measurement in Education*, *34*(1), 10–26. https://doi.org/10.1080/08957347.2020.1835910

Mosvold, R., & Hoover, M. (2016). *The Mathematics Enthusiast*, *13*(1–2). https://doi.org/10.54870/1551-3440.1362

Mullis, I. V. S., & Martin, M. O. (2013). TIMSS 2015 item writing guidelines. In *International Association for the Evaluation of Educational Achievement*. Lynch School of Education.

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common core state standards for mathematics: Grade 6 introduction*. http://www.corestandards.org/Math/Content/6/introduction/

Orrill, C. H., & Cohen, A. 2016. Purpose and conceptualization: Examining assessment development questions through analysis of measures of teacher knowledge. In: A. Izsák, J. T. Remillard & J. Templin (Eds.), *Psychometric methods in mathematics education: Opportunities, challenges, and interdisciplinary collaborations. Journal for Research in Mathematics Education Monograph Series No. 15* (pp. 139–153). Reston, VA: National Council of Teachers of Mathematics.

Orrill, C. H., Kim, O. -K., Peters, S. A., Lischka, A. E., Jong, C., Sanchez, W. B., & Eli, J. A. (2015). Challenges and strategies for assessing specialised knowledge for teaching. *Mathematics Teacher Education & Development*, *17*(1), 12–29.

Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136–144. https://doi.org/10.7334/psicothema2013.259

Rowan, B., Schilling, S. G., Ball, D. L., Miller, R., Atkins-Burnett, S., & Camburn, E. (2001). Measuring teachers' pedagogical content knowledge in surveys: An exploratory study. *Consortium for Policy Research in Education*.

Schilling, S. G., Blunk, M., & Hill, H. C. (2007). Test validation and the MKT measures: Generalizations and conclusions. *Measurement*, *5*(2–3), 118–128. https://doi.org/10.1080/15366360701487146

Schilling, S. G., & Hill, H. C. (2007). Assessing measures of mathematical knowledge for teaching: A validity argument approach. *Measurement*, *5*(2–3), 70–80. https://doi.org/10.1080/15366360701486965

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14. https://doi.org/10.3102/0013189X015002004

Zhai, X., Haudek, K. C., Wilson, C., & Stuhlsatz, M. (2021). A framework of construct-irrelevant variance for contextualized constructed response assessment. *Frontiers in Education*, 6. https://doi.org/10.3389/feduc.2021.751283