

## **Why is measuring Mathematical Knowledge for Teaching so hard? A struggle towards validation through student work analysis**

Dédé de Haan<sup>1,2</sup>, Gerrit Roorda<sup>3,2</sup>, Siebrich de Vries<sup>2,3</sup>, Paul Drijvers<sup>1</sup>

<sup>1</sup>*Utrecht University*, <sup>2</sup>*NHLStenden University of Applied Sciences*, <sup>3</sup>*University of Groningen*

Measuring Mathematical Knowledge for Teaching (MKT) through open-response items proves challenging, with studies reporting variability in reliability and unclear subdomain boundaries. We illustrate this through the development of a 19-item test assessing three MKT subdomains - Specialized Content Knowledge (SCK), Knowledge of Content and Students (KCS) and Knowledge of Content and Teaching (KCT) - for Dutch pre-service teachers. Despite careful validation, the instrument demonstrated low internal consistency across subdomains (standardized  $\alpha=.36-.42$ ). Examining two KCT items revealed two construct-dependent challenges. One item achieved high inter-rater reliability (ICC=.890) yet confounded SCK/KCT constructs, as indicated by qualitative analysis. The other item showed lower inter-rater reliability (ICC=.524); resolving this through strict conceptual scoring criteria revealed pre-service teachers predominantly provide procedural explanations, causing restricted variance and contributing to low internal consistency across the KCT subdomain. These findings suggest that achieving both theoretical purity and internal consistency may be difficult when participants demonstrate predominantly procedural knowledge.

**Keywords: pre-service mathematics teacher education, Knowledge of Content and Teaching, measuring, validation**

### **Introduction**

Measuring Mathematical Knowledge for Teaching (MKT) through written assessment proves difficult. Despite decades of instrument development following Ball et al.'s (2008) framework, studies consistently report variability in reliability and ambiguity in subdomain boundaries (e.g., Copur-Gencturk & Tolar, 2022). We experienced these challenges when developing an instrument measuring Dutch pre-service teachers' (PSTs') knowledge of algebraic reasoning across three of Ball et al.'s (2008) MKT subdomains: Specialized Content Knowledge (SCK), which refers to mathematical knowledge uniquely needed for teaching, such as analyzing why procedures work; Knowledge of Content and Students (KCS), which combines mathematical understanding with knowledge of how students typically think about, learn, or misunderstand particular content and Knowledge of Content and Teaching (KCT), which involves knowing how to design instruction and select representations or examples to support student learning of specific mathematical content. Despite a careful two-year validation process involving expert content validation (Phase 1), systematic scoring reliability analysis (Phase 2), and psychometric testing (Phase 3), the instrument showed low internal consistency across subdomains (standardized  $\alpha$  between .36 and .42). The aim of this paper is to document where and why these validation difficulties arose, illustrating construct-dependent challenges that may reflect fundamental tensions in measuring KCT rather than merely correctable

instrument flaws. Such detailed documentation remains rare (Flake et al., 2017) but can clarify whether measurement difficulties stem from correctable instrument flaws or from fundamental tensions in the MKT construct itself.

While Ball et al.'s (2008) framework distinguishes SCK, KCS and KCT as theoretically separate domains, empirically distinguishing these domains can be challenging. Explaining why an algebraic procedure works (SCK) is inseparable from evaluating how to teach it conceptually (KCT), making it difficult to isolate what KCT items actually measure. Another challenge involves normative judgments about procedural versus conceptual approaches. Skemp's (1976) distinction between instrumental understanding (knowing how) and relational understanding (knowing why) applies across MKT domains. Open-response formats make these differences visible (Fauskanger, 2015), but introduce dilemmas: should procedural teaching strategies be credited as pragmatically effective, or scored lower as conceptually limited? This tension is particularly acute with PSTs, who often demonstrate instrumental rather than relational understanding even in their final year (Maher & Muir, 2013).

We illustrate these measurement challenges through our validation process. The instrument combined multiple-choice and open-response items. We focus on two open-response KCT items because they most clearly illustrate two construct-dependent challenges described above: (1) subdomain overlap, where items blend SCK/KCT distinctions, and (2) normative judgment challenges, where raters disagree on whether procedural teaching strategies constitute adequate KCT. Open-response formats were essential because they distinguish procedural from conceptual understanding (Fauskanger, 2015), yet this very distinction creates scoring dilemmas about what counts as adequate pedagogical reasoning.

## Method

We developed an MKT instrument on algebraic reasoning for Dutch PSTs, consisting of 19 items: three multiple-choice items targeting SCK and sixteen open-response items (3 SCK, 6 KCS, 7 KCT). Participants were fourth-year PSTs from three teacher education institutes ( $N=65$ ). Validation followed a three-phase process.

- Phase 1, May 2022: Content validation. Four experts classified all items into MKT subdomains and collaboratively developed scoring rubrics.
- Phase 2, 2022-2023: Reliability scoring. The same four experts independently scored responses from six PSTs. We calculated inter-rater reliability (ICC, two-way random effects, absolute agreement) and then facilitated consensus discussions on disagreements. These deliberations were transcribed and analyzed to identify sources of scoring variability. Where experts agreed on what responses demonstrated but disagreed on scoring standards, explicit criteria were established.
- Phase 3, 2023-2024: Empirical piloting and psychometric testing. With the full sample ( $N=65$ ) and applying Phase 2 scoring agreements, we computed Cronbach's alpha per subdomain and analyzed score distributions. We selected two KCT items (Items 5 and 8) that proceeded through all three phases to illustrate the validation challenges identified.

## Results

Psychometric analysis of the full 19-item instrument showed consistently low internal consistency across all three MKT subdomains (standardized  $\alpha$ : SCK=.36, KCS=.42,

KCT=.37). Analysis of Phase 2 expert deliberations identified normative disagreement: experts agreed on what knowledge was demonstrated but disagreed on pedagogical adequacy. Phase 3 analyses uncovered a second construct-dependent challenge: subdomain overlap, where items blurred SCK/KCT distinctions despite intended classification.

We illustrate these challenges through two items. First, Item 5 displayed high inter-rater reliability (ICC=.890 in Phase 2). However, Phase 3 analysis revealed subdomain overlap through qualitative analysis of responses: PSTs defaulted to demonstrating solution methods (SCK) rather than instructional explanations (KCT). Second, Item 8 showed lower inter-rater reliability (ICC=.524 in Phase 2) due to normative disagreement; resolving this through strict scoring criteria improved inter-rater agreement but created floor effects when applied in Phase 3, with nearly half of PSTs scoring zero. These restricted score distributions were characteristic of the KCT subdomain and contributed to low internal consistency ( $\alpha=.37$ ).

### Pattern 1: Subdomain overlap (Item 5)

Item 5 (see Figure 1) asks PSTs to show two ways to explain solving  $(2x + 3)/(4x + 6) = 2$ , indicating advantages and disadvantages. This item was designed to assess KCT.

Figure 1. Item 5 from the 2023-2024 instrument, based on Arcavi (1994, p. 27)

**5.**  
The textbook contains the following exercise:

Solve:

$$\frac{2x + 3}{4x + 6} = 2$$

Show **two ways** to explain to your students how to solve the equation above. For each explanation, indicate what the potential advantages and disadvantages are.

PST 1 (Figure 2, left) warns against “instrumental understanding” yet fails to recognize  $x = -1.5$  makes the denominator zero. PST 2 (Figure 2, right) presents mathematically correct solutions with pedagogical commentary, but demonstrates how to solve, not how to explain to their students. High reliability (ICC=.890) resulted from shared scoring heuristics but the item prompted integrated responses that confounded subdomain boundaries.

Figure 2. Responses on item 5 from PST 1 (left) and PST 2 (right)

**UTITLEG 1**  
 $\frac{2x+3}{4x+6} = 2 \Rightarrow 2x+3 = 2(4x+6)$   
since want  $\frac{2}{4} = \frac{1}{2} \Rightarrow 6 = 2 \cdot 3$

$$\begin{array}{r} 2x+3 = 8x+12 \\ -8x \quad -8x \\ -6x+3=12 \\ \quad \quad \quad -3 \\ -6x = 9 \\ \quad \quad \quad \cdot \frac{1}{-6} \\ x = -\frac{3}{2} = -1\frac{1}{2} \end{array}$$

**UTITLEG 2**  
 $\frac{2x+3}{4x+6} = 2$   
eerst kijken, en zien dat teller helfte is van noemer

$$\begin{array}{r} \frac{2x+3}{2(2x+3)} = 2 \\ 2x+3 = 0 \\ 2x = -3 \\ x = -1\frac{1}{2} \end{array}$$

**VOOR- EN NADELEN**  
Bekende, veilige methode.  
Opassen dat het dan geen instrumenteel begrip wordt.

Familier, save method. Watch out that it doesn't turn into instrumental understanding

**VOOR- EN NADELEN**  
Sneller, maar daar moet je eerst opletten voor naarde vergelijking kijken

Faster, but you first need to look at the equation observantly

**UTITLEG 1**  
you can write as  $\frac{2x+3}{2(2x+3)} = 2$  dan heb je then you have

$$\frac{2x+3}{2(2x+3)} = 2$$

Above and below the division bar you divide by  $(2x+3)$   
Nu heb je  $\frac{1}{2} = 2$   
Actual eigen oplossingen voor het probleem.

Advantage: This way is fast  
This method is efficient  
Disadvantage: This method is not always possible, because you can't always divide the numerator and the denominator by the same factor

**UTITLEG 2**  
 $\frac{2x+3}{4x+6} = 2$   
 $\Rightarrow 2x+3 = 2(4x+6)$   
 $\Rightarrow 2x+3 = 8x+12$   
 $\Rightarrow -6x = 9$   
 $\Rightarrow x = -1.5$

Check Confuse  $x = -1.5$  fill in  $\frac{2(-1.5)+3}{4(-1.5)+6} = 0$  geen oplossing

**VOOR- EN NADELEN**  
Deze manier is overzichtelijk  
Deze manier werkt altijd

Advantage: This way is well-organized  
This method always works

Disadvantage: Solutions are not always checked, as a result of which you don't encourage the problem of 0/0

### Pattern 2: Normative judgment challenges (Item 8)

Item 8 (Figure 3) presents a classroom dialogue about simplifying  $3e + 7e$  where the teacher uses counting (“How many  $e$ ’s?”) and the student responds “10  $e$  squared.” PSTs are asked to explain how they would teach simplifying this. This item was designed to isolate KCT by providing a teaching scenario requiring pedagogical response.

Figure 3. Item 8 from the 2023–2024 instrument, based on <https://www.timssvideo.com/nl3-surface-area> (from 4:42 – 5:08)

8.

3a Write the formula  $p = 3e + 7e$  in shorter form

A student in grade 2HV (Year 8 / lower secondary, second year) asks for help with the above exercise. The following dialogue then takes place:

Teacher: "You're just doing algebra. So for example in exercise number three:  $p = 3e + 7e$ "  
 Student: "That gives you when you combine them..."  
 (simultaneously) Teacher: "How many  $e$ 's do you have?"  
 Student: "10  $e$ . 10  $e$  squared."  
 Teacher: "No... why squared?"  
 Student: "Yes, because you have 2  $e$ 's"  
 Teacher: "How much is 3  $e$ 's plus 7  $e$ 's?"  
 Student: "10  $e$ 's..."  
 Teacher: "Period. As soon as you multiply them together, then you get a square."  
 Student (surprised): "Oh!"

Explain below **how** you would teach students to write this in shorter form, and **why** you would use that approach:

During Phase 2, Item 8 achieved moderate inter-rater reliability (ICC=.524). Raters agreed on what responses demonstrated, but disagreed whether procedural explanations constituted adequate KCT. This reflected Skemp’s (1976) debate: should procedural teaching strategies receive credit, or only relational and conceptual explanations? To resolve this disagreement, experts established explicit scoring criteria: procedural explanations would receive no points, while conceptual explanations emphasizing the variable concept would receive full credit. When these scoring agreements were applied in Phase 3 ( $N=65$ ), improving inter-rater reliability, a new pattern emerged. Figure 4 shows two contrasting responses: PST 1 (at the left) combining notational and conceptual approaches with explicit attention to the meaning of variables, receiving full credit; PST 2 (at the right) focusing on notational conventions with only implicit conceptual attention, receiving minimal credit under the Phase 2 agreements. However, most Phase 3 responses resembled PST 2, providing largely procedural explanations with limited conceptual emphasis.

Figure 4. Responses on item 8 (how and why) from PST 1 (left) and from PST 2 (right)

<p>As introduction:  <math>3e + 7e =</math>  <math>e+e+e + e+e+e+e+e+e</math>                  together 10e.</p> <p><math>e</math> is a variable, take numerical example  <math>e = 4</math>  <math>3 \cdot 4 + 7 \cdot 4</math>  <math>12 + 28 = 40</math>                  or <math>(3+7) \cdot 4 = 10 \cdot 4</math></p> <p>and not suddenly <math>10 \cdot 4^2 = 160</math></p>	<p>Give a clear understanding of what a term like '3e' means. Also attention for <math>e</math> as a variable and not an object like 'apple' or something.</p>
<p>What does <math>3e</math> mean?  <math>e + e + e</math>                  What does <math>7e</math> mean?  <math>e + e + e + e + e + e + e</math>                  How many <math>e</math>'s do we have in total then?</p> <hr/> <p>Or what does <math>e^2</math> mean?                  "e times e" / "e times itself"                  Does my exercise show e times e?</p>	<p>Back to the definition to retrieve the concept formation, form of recognition.                  My opening explanation also always starts with how can I write <math>3+3+3+3+3+3+3</math> more easily?                  "Oh yes! <math>7 \cdot 3</math>"                  Give the student understanding of what a square give [sic], that for that you have to multiply something.</p>

Assigning 0 points to procedural explanations directly caused floor effects, with nearly half of PSTs scoring zero. Achieving adequate variance would require either less strict scoring criteria or items better aligned with PSTs’ knowledge levels.

## Discussion

Documenting the validation process unveils two construct-dependent challenges in assessing MKT: subdomain overlap (Item 5) and normative judgment challenges (Item 8). These may not be merely correctable instrument deficiencies (Hoover et al., 2016) but appear to reflect fundamental tensions in how KCT is conceptualized and assessed when participants demonstrate predominantly instrumental rather than relational understanding.

Our findings indicate that subdomain boundaries blur in PSTs' responses. Item 5 achieved high inter-rater reliability yet revealed subdomain confusion: responses focused on solution methods (SCK) rather than instructional explanation (KCT). Without explicit pedagogical context, such as specifying student level, prior knowledge or instructional moment, the item design led to this blurring. Moreover, the substantial floor effects observed across KCT items point to deeper measurement challenges: when mathematical understanding is largely instrumental, pedagogical reasoning cannot emerge clearly. As Item 5 illustrates, PSTs demonstrated their own procedural solution methods, thereby revealing instrumental understanding, rather than conceptual pedagogical explanations.

Item 8's explicit pedagogical scenario more successfully isolated KCT by clearly signaling what type of knowledge was required. Yet Item 8 demonstrates how resolving one assessment challenge can expose another. Phase 2 identified normative disagreement: experts agreed on what responses demonstrated but disagreed on whether procedural strategies constituted adequate KCT, reflecting tensions between pragmatic views and normative standards requiring full conceptual justification (Scheiner et al., 2024). Establishing strict scoring criteria resolved rater disagreement and improved inter-rater reliability. However, when applied at scale in Phase 3, these criteria revealed that most PSTs provide primarily procedural teaching explanations. This is consistent with Maher and Muir's (2013) finding that final-year PSTs demonstrate instrumental rather than relational understanding. With nearly half of the PSTs scoring zero, this resulted in restricted variance, a pattern observed across multiple KCT items.

These findings converge on a central insight: both challenges stem from PSTs demonstrating predominantly instrumental instead of relational understanding (Skemp, 1976), that is, procedural rather than conceptual knowledge. Item 5 shows this through blurred SCK/KCT boundaries and extensive floor effects: PSTs' responses did not demonstrate mathematical or pedagogical reasoning at the required conceptual level. Item 8 shows this through mostly procedural teaching explanations; even when explicitly asked for pedagogical reasoning, PSTs provide procedural approaches. This pattern resulted in restricted variance, producing low internal consistency ( $\alpha = .37$ ). The low alpha may not indicate poor item quality but rather accurately reflect PSTs' actual knowledge profile of predominantly instrumental understanding that characterizes both their mathematical reasoning and pedagogical explanations.

This analysis suggests that achieving both theoretical purity (measuring KCT as conceptually grounded pedagogical reasoning) and acceptable internal consistency may be inherently difficult when participants demonstrate predominantly procedural knowledge. Items targeting this conceptual standard primarily demonstrate what PSTs cannot yet do, generating floor effects that undermine psychometric quality. However, crediting instrumental understanding to improve score variance does not resolve the fundamental issue: items still risk measuring an undifferentiated blend of knowledge domains. When PSTs' own mathematical understanding remains instrumental, distinguishing their pedagogical reasoning from their mathematical competence in

written assessments becomes extremely difficult. This creates an assessment dilemma with no straightforward solution: theoretical purity conflicts with psychometric adequacy when the construct demands conceptual reasoning that most PSTs have not yet developed. That measurement challenges with MKT vary depending on how content knowledge is conceptualized (Copur-Gencturk & Tolar, 2022) indicates these difficulties reflect fundamental tensions in the construct itself, which may be particularly acute when participants demonstrate predominantly instrumental understanding.

## Acknowledgements

This research is funded by the Dutch Research Council (NWO), Grant Number: 023.016.040

## References

- Arcavi, A. (1994). Symbol Sense: Informal Sense-making in Formal Mathematics. *For the Learning of Mathematics*, 14(3), 24–35.
- Ball, D., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, 59(5), 389–407. <https://doi.org/10.1177/0022487108324554>
- Copur-Gencturk, Y., & Tolar, T. (2022). Mathematics teaching expertise: A study of the dimensionality of content knowledge, pedagogical content knowledge, and content-specific noticing skills. *Teaching and Teacher Education*, 114, Article 103696. <https://doi.org/10.1016/j.tate.2022.103696>
- Fauskanger, J. (2015). Challenges in measuring teachers' knowledge. *Educational Studies in Mathematics*, 90(1), 57–73. <https://doi.org/10.1007/s10649-015-9612-4>
- Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research: Current Practice and Recommendations. *Social Psychological and Personality Science*, 8(4), 370–378. <https://doi.org/10.1177/1948550617693063>
- Hoover, M., Mosvold, R., Ball, D. L., & Lai, Y. (2016). Making progress on mathematical knowledge for teaching. *Mathematics Enthusiast*, 13(1–2), 3–34. <https://doi.org/10.54870/1551-3440.1363>
- Maher, N., & Muir, T. (2013). "I know you have to put down a zero, but I'm not sure why": Exploring the link between pre-service teachers' content and pedagogical content knowledge. *Mathematics Teacher Education and Development*, 15(1), 72–87.
- Scheiner, T., Buchholtz, N., & Kaiser, G. (2024). Mathematical knowledge for teaching and mathematics didactic knowledge: a comparative study. *Journal of Mathematics Teacher Education*, 27(6), 1083–1104. <https://doi.org/10.1007/s10857-023-09598-z>
- Skemp, R. R. (1976). Relational Understanding and Instrumental Understanding. *Mathematics Teaching*, 77(1), 20–26.