CYCLES OF EVIDENCE COLLECTION IN THE DEVELOPMENT OF A MEASURE OF TEACHER KNOWLEDGE

Laurie O. Cavey, Tatia Totorica, Ya Mo, Michele Carney

Boise State University

This study highlights some of the tensions that arise during measure development while attending to both Rasch measurement principles and mathematics education's focus on high quality operationalization of complex theoretical constructs. We situate our measure development work within the context of a larger design-based mathematics teacher preparation intervention project focused on improving teacher candidate attentiveness, and illustrate how these tensions have shaped our instrument and item development work over the last four years.

INTRODUCTION

Persistent global concerns regarding the quality and efficacy of mathematics instruction have long influenced mathematics education research agendas (e.g., Council for the Accreditation of Educator Preparation, 2022; Grossman, Hammerness, & McDonald, 2009) and have led to ongoing efforts to (a) articulate the range of constructs related to effective mathematics teaching (Ball, Thames & Phelps, 2008), (b) develop scaled instruments which reliably measure the skills and knowledge associated with each construct (e.g., Mathematical Knowledge for Teaching Measures from the Learning Mathematics for Teaching Project, 2005), and (c) design interventions with the potential to improve teachers' and prospective teachers' position on those scales (Hill, Rowan, & Ball, 2005). This study reports on some of the challenges found at the intersection of measure development, intervention design, and mathematics teacher education program implementation. The focus of this paper is on the development of a measure of teacher attentiveness, the Disciplinary Attentiveness to Student Ideas-Quantitative Reasoning Instrument (DASI-QRI), and items which feature evidence of student quantitative reasoning at the secondary level. Operating under the constraints of mathematics teacher preparation programs and the realities of intervention implementation while also adhering to the charge that "a series of interrelated investigations is required to understand the construct(s) that a measure assesses" (Clark & Watson, 2019, p.1413) has surfaced new complexities associated with measure development for mathematics teacher education. Through a focus on the iterative development of one item in our instrument, we illustrate how multiple cycles of evidence collection and analyses can be used to inform revisions, delineate how these multiple cycles may be necessary to surface a range of different issues, and highlight some of the tensions that must be navigated while designing scalable measures with the potential to yield meaningful data for mathematics education researchers, teacher educators, and professional development providers.

2022. In C. Fernández, S. Llinares, A. Gutiérrez, & N. Planas (Eds.), *Proceedings of the 45th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 115-122). PME.

BACKGROUND

The DASI-QRI was developed to measure *attentiveness* to students' quantitative reasoning. Attentiveness integrates components of mathematical knowledge for teaching (Ball, Thames, & Phelps, 2008; Shulman, 1987), professional noticing (Jacobs et al., 2010), progressive formalization (Freudenthal, 1973; Gravemeijer & van Galen, 2003; Treffers, 1987), and formative assessment (Black & Wiliam, 2009). It is defined as *the ability to analyze and respond to a particular student's mathematical ideas from a progressive formalization perspective* (Carney, Cavey, & Hughes, 2017). Previous work with construct map development for attentiveness (Carney, Totorica, Cavey & Lowenthal, 2019) informs item development for the DASI-QRI.

The instructional intervention associated with the development of the DASI-QRI is designed to increase attentiveness to students' quantitative reasoning and consists of a series of modules with both asynchronous and synchronous components. Each module centers upon a challenging, nontraditional task and features a sequenced collection of curated video and written artifacts of secondary students working on the task. The focus and development of module content has been described elsewhere (e.g., Cavey, Libberton, Totorica, Carney, & Lowenthal, 2020).

The Standards for Educational and Psychological Testing's (AERA, APA, NCME, 2014) argument-based approach to validity encourages conceptualizing development and validation as an ongoing, iterative process. However, certain constraints can lead to more iterations than might typically be expected. For the DASI-QRI, three interrelated, yet distinct factors led to numerous iterations. These factors were:

1. Test development within an instructional intervention development project,
2. Measuring and defining the components of the attentiveness construct, and
3. Use of the Rasch measurement model, which demands consideration of many different test and item indicators, yet also yields a high-quality product.

CYCLES OF EVIDENCE COLLECTION AND ANALYSIS

Annual administrations, each consisting of multiple cycles of evidence collection, has informed the development and revision of DASI-QRI items. The type of evidence collected depended on the status of development for both the instrument and individual items. For example, administration of the DASI-QRI in Year 1 did not include Rasch analysis because the number of participants was limited. Additionally, each selected-response (SR) item was initially developed through analysis of responses to the constructed-response (CR) version and identification of exemplar responses for use in the SR version (Carney, Cavey, & Hughes, 2017). Response process analysis of cognitive interview data examined the degree of match between participant responses to the CR and SR versions of the item and informed SR item revision (Mo, Carney, Cavey, & Totorica, 2021). Once item development/revisions were completed, the DASI-QRI was administered as a pre/post measure in courses using the associated

intervention. Rasch analysis techniques were used to examine both individual item functioning and overall item and person statistics for the instrument as a whole. Results of these analyses then prompted additional evidence collection and revision. See Table 1 for a brief overview of the ways in which cycles of evidence collection have impacted the development of the DASI-QRI and one item, in particular, the Truck Intent Item, provided in Figure 1.

		Items			
	n	CR	SR	DASI-QRI Changes	Truck Intent Item Changes
Year 1 Pre	35	15	0	N/A	Development of SR version
Year 2 Pre	89	0	12	added 2 SR items	Revised SR version
Year 3 Pre	127	0	14	added 6 CR items	none
Year 4 Pre	129	6	14	none	none
Year 4 Post	116	6	14	TBD	TBD

Table 1. Cycles of Evidence Collection and Impact

The Truck Intent Item is the first of three questions related to the Algebra I task pictured in Figure 1 (Note: Algebra I refers to the standard first course in algebra for ages 13-14 in the U.S.). Subsequent items include images of secondary student work on the task and prompt candidates to indicate their level of agreement with SR options related to the student's approach and potential teacher responses.

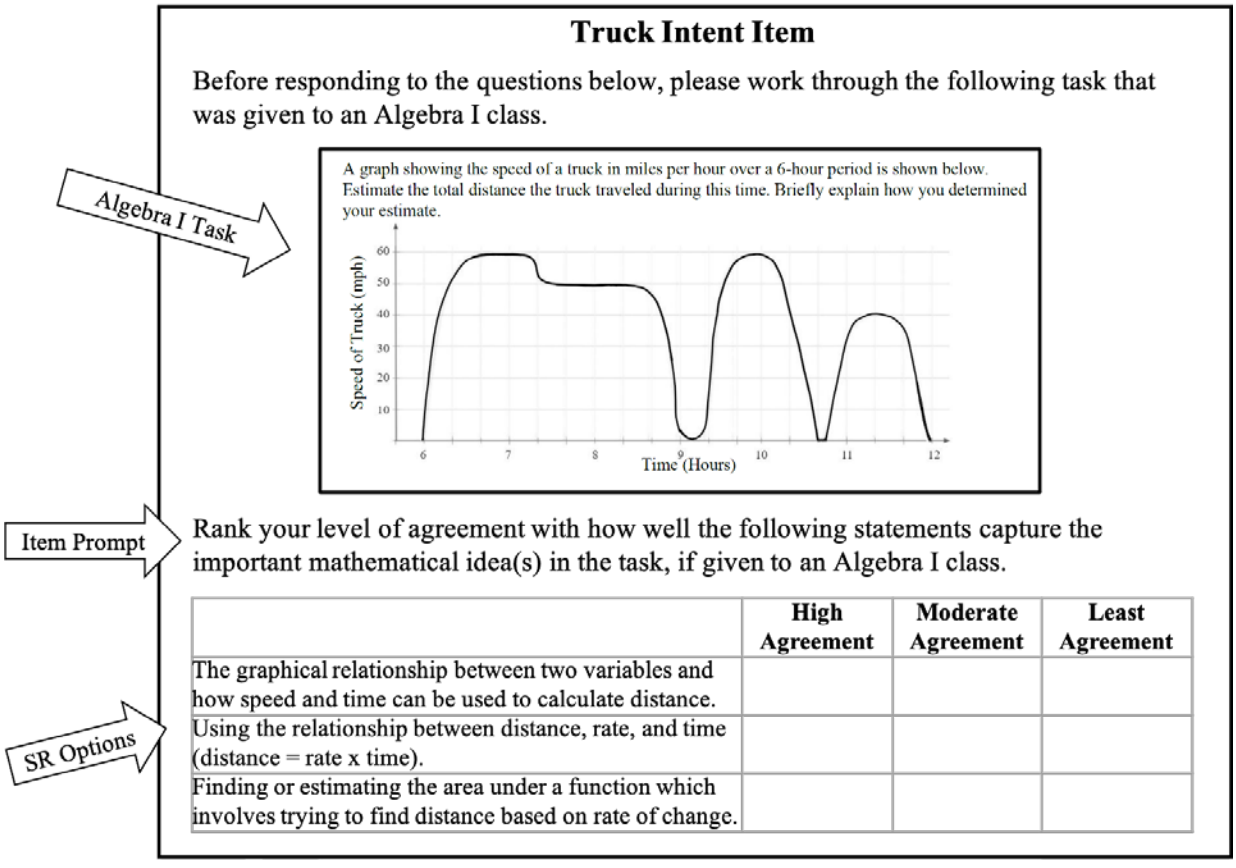

Figure 1. Truck Intent Item

The intended response for the ranked item, listed from high {H} to least {L} agreement, is: {H} The graphical relationship between two variables and how speed and time can be used to calculate distance, {M} Using the relationship between distance, rate, and time (distance = rate × time), and {L} Finding or estimating the area under a function which involves trying to find distance based on rate of change. Given that the stated context is Algebra I, finding the area under a function is not a generally appropriate mathematical focus. Option {H} situates {M} in the context of graphical reasoning, and is thus the more complete and appropriate description of the mathematical focus. The Truck Intent Item is scored based on correctly ranking the {H} (1 point) and the {L} (1 point) (see Table 2).

Ordering of SR Options	Score
{HML}	2
{HLM} or {MHL}	1
{LHM} or {MLH} or {LMH}	0

Table 2. Scoring Scheme for the Truck Intent Item

We focus our results on the Year 4 administration of the DASI-QRI and the Truck Intent Item to illustrate how repeated cycles of evidence collection and analysis are necessary to uncover potential issues. Participants in Year 4 were enrolled in a mathematics course across 13 U.S. universities in which the course instructor implemented the project's intervention. Year 4 analyses of the 20-item instrument, to date, have included Rasch analysis on the pre measure for 129 participants, on the post measure for 116 participants, and on the pre-post paired data for the 62 candidates who appeared to meaningfully engage with the intervention. Qualitative analyses of individual item response trends for the 62 participants and previously collected cognitive interview data for 13 participants were also completed.

With respect to the Rasch analyses, two persons with extreme scores of 0 were dropped from the post measure responses of 116 participants across 20 items. There were no extreme scores on the pre. No extreme scoring categories were dropped from the analyses for either the pre or the post. There were four groups of items based upon the format (SR versus CR) and the number of ranking options for the SR (2, 3, and 4 options). The items within the same grouping share the same partial credit response structure. The JMLE estimation process converged when the maximum logit change was .0041 (.0033 pre).

RESULTS

Rasch Analysis

Overall, the item "test" reliability is .96(.97 pre), which is very high, with a separation index of 5.15(5.81 pre). The sample size allows the item difficulties to be estimated precisely and confirms the item difficulty hierarchy (e.g., high, medium, low item difficulties) of the instrument. The person "test" reliability is .72(.44 pre); the person

separation index is 1.62(.89 pre). Thus, the instrument may not be sensitive enough to distinguish between high and low performers or more performance levels in the sample. The raw variance explained by the Rasch measure was 24.09%(24.7% pre). The point-measure correlations (PTMEASUR) were all positive, suggesting that all the items were pointing in the same direction. Except for the Truck Intent Item on the post, the mean-squares (MNSQ) were not excessive, so the misfit was acceptable; the standardized statistics (ZSTD) for both INFIT and OUTFIT were not extreme; thus, we failed to reject the null hypothesis that these data fit the Rasch model. However, the Truck Intent Item had an OUTFIT MNSQ of 1.69 and a ZSTD of 4.64 due to some unexpected responses on the post. This indicated additional analysis may be needed.

For the Truck Intent Item, item category frequency analysis indicates that the average measures advanced with the score categories for the pre but do not advance with the categories on the post; 22 people with a score of 2 had an average measure of -.12, less than the average measure of -.02 of 35 people with a score of 1 (see Table 3).

		Pre		Post	
		Frequency	Mean ability	Frequency	Mean ability
Item Score	0	42 (33%)	-.26	59 (51%)	-.35
	1	44 (34%)	-.15	35 (30%)	-.02
	2	43 (33%)	.07	22 (19%)	-.12*

Table 3. Item Score Frequencies and Mean Ability

The Item Characteristic Curves (ICC) in Figure 2 (pre on the left and post on the right) show the empirical ICC (blue line) of Truck Intent with an unexpected behavior (outside the 95% confidence bands [grey line] around the expected Item Characteristic Curves [red line]). For the pre, there is an unexpected drop in the highest end of the latent variable (i.e., measure), and on the post, there is an unexpected rise at -2.8 and -1.2 and an unexpected drop at 0.5 in the latent variable relative to item difficulty scale.

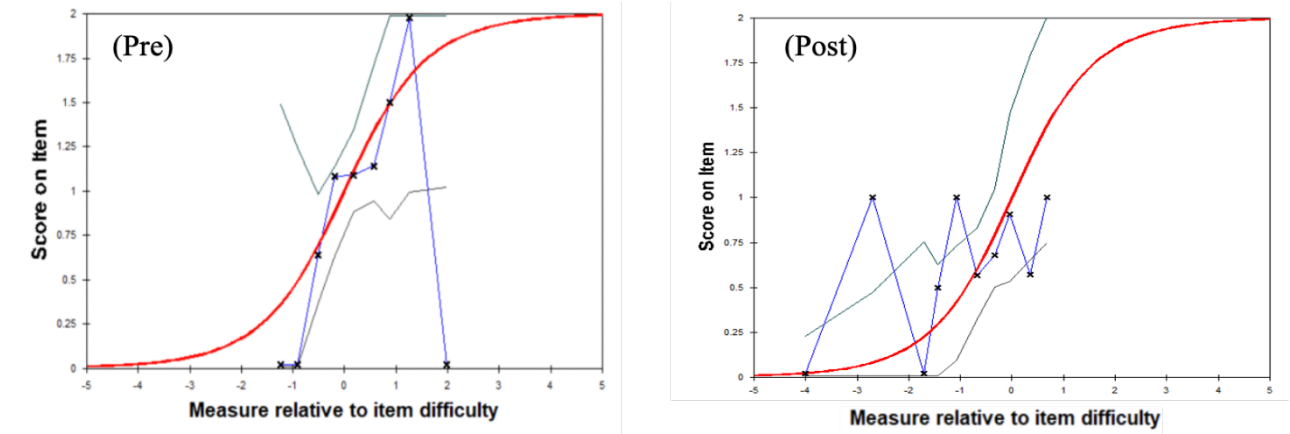

Figure 2. Item Characteristic Curves (ICC) of Truck Intent Item for Pre and Post

Truck Intent Response Analysis

Examination of the raw response data for the subset of 62 participants revealed that 21 (~34%) scored lower on the Truck Intent Item on the post compared to the pre. The highlighted portion of Table 4 shows how participant rankings changed, which includes 12 of the 16 participants (75%) who originally scored a 2 and 10 of the 23 (~43%) participants who originally scored a 1. While there were also participants who scored higher on Truck Intent Item on the post, it was the large percentage of participants with decreased scores that prompted further examination of the data.

		Post-Test Ordering						
		{HML}	{HLM}	{MHL}	{LHM}	{MLH}	{LMH}	Total
Pre-Test Ordering	{HML}	4	1	6	1	1	3	16
	{HLM}	2	1	2	0	0	3	8
	{MHL}	3	1	5	1	2	3	15
	{LHM}	1	0	1	0	1	4	7
	{MLH}	2	3	0	1	4	2	12
	{LMH}	0	0	1	1	1	1	4
	Total	12	6	15	4	9	16	**62**

Table 4. Pre- and Post-Test Rank Ordering of SR Options for the Truck Intent Item

Re-examination of the cognitive interview data, collected in an effort to better understand response process, revealed that of the 10 participants interviewed who mentioned the Algebra 1 context of the Truck Intent Item, all 10 ranked {H} with the highest level of agreement, and 8 responded with the intended ranking {HML}. Interestingly, it did not seem to matter whether or not a participant noticed the potential connection to calculus or the graphical reasoning aspect, though data are limited as there were only 4 different ranking options represented in the sample and only one participant selected {LMH}. In addition, only one participant explicitly compared options {H} and {M}, indicating that further data collection is needed to better understand response processes associated with the Truck Intent Item.

TENSIONS IN MEASURE DEVELOPMENT

Measure development for complex constructs such as attentiveness remains a persistent challenge for the mathematics education community, one that often garners superficial nods to statistics associated with reliability and validity (e.g., Cronbach's alpha) or else is sidestepped altogether in lieu of qualitative assessment. This could be due, in part, to the tensions which arise when attempting to address issues revealed by Rasch analysis while also operating within the context of mathematics education. For example, from the measurement perspective, the DASI-QRI's low person "test" reliability and separation index indicate the need for additional items. Yet from the mathematics education perspective, additional items, especially when considering the

cognitive complexity required to elicit evidence of attentiveness, place an undue burden on test-takers in terms of both time and fatigue. Quality instrument development from the measurement development perspective also often depends upon ready availability of large participant pools. In contrast, recruitment of participants within the mathematics education community - especially within the context of a larger instructional intervention project - can be a significant challenge. Furthermore, quality instruments and their items are expected to perform roughly the same across all administrations with an implicit assumption that test-takers complete the assessment with fidelity. However, instrument use in the mathematics education community is often embedded in a mathematics course or professional development as part of an intervention; thus, test-taker motivation and investment in completing the assessment with fidelity can vary depending on the timing of administration and test-taker perceptions of the assessment. Could this be why the Truck Intent Item performed well in some administrations and raised issues of concern in another? Did performance decrease simply because participants missed or ignored the course context when they completed the post? Are variances in Rasch analysis results due to instrument or item failings that can be addressed via revision, or are they due to something else?

FINAL THOUGHTS

We have aimed to highlight the complexity of measure development when meaningfully attending to both Rasch measurement principles and mathematics education's focus on high-quality operationalization of complex theoretical constructs, particularly within the context of developing a measure associated with an instructional intervention. The issues which surface when considering each perspective precipitate different kinds of development work, the outcomes of which can impact the other. This often warrants additional cycles of evidence collection, analysis, and revision, and elicits tensions from the mathematics education side, as we must also consider persistently small sample sizes, the length of the assessment, the time demands on instructors and teacher candidates, and alignment between item design and the cognitive complexity of attentiveness we wish to measure.

ACKNOWLEDGEMENTS

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

Ball, D. L., Thames, M. H., & Phelps, G. C. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*(5), 389-407.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5-31.

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

Carney, M. B., Totorica, T., Cavey, L. O., & Lowenthal, P. R. (2019). Developing a construct map for teacher 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.

Cavey, L. O., Libberton, J., Totorica, T., Carney, M., Lowenthal, P. (2020). VCAST learning modules: A functions & modeling course innovation. In J. Goodell and S. Kock (Eds.), *Preparing STEM teachers: A replication model* (pp. 259-275). Information Age Publishing.

Council for the Accreditation of Educator Preparation (CAEP). (2022). 2022 CAEP Initial Level Standards. Washington, DC: Author. Retrieved from http://caepnet.org/standards/2022-itp/introduction

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.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371-406.

Jacobs, V. R., Lamb, L. L., & Philipp, R. A. (2010). Professional noticing of children's mathematical thinking. *Journal for Research in Mathematics Education*, *41*(2), 169-202.

Learning Mathematics For Teaching. (2005). *Mathematical knowledge for teaching measures.* Ann Arbor, MI.

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.

Shepard, L. A. (2018). Learning progressions as tools for assessment and learning. *Applied Measurement in Education, 31*(2), 165-174.

Sherin, M. G., & van Es, E. A. (2005). Using video to support teachers' ability to notice classroom interactions. *Journal of Technology and Teacher Education, 13*(3), 475-491.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1-22.

Treffers, A. (1987). *Three dimensions: A model of goal and theory description in mathematics instruction – The Wiskobas Project*. Dordrecht, The Netherlands: Reidel.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, N.J.: Lawrence Erlbaum Associates.