MEASURING MATHEMATICAL COMPETENCIES WITH A LARGE-SCALE MULTI-FORMAT ASSESSMENT

Sam Ehrlich

Department of Teaching and Learning, Stockholm University, Sweden

Large-scale multi-format assessments include short-response, constructed-response, oral, and investigative mathematics tasks, offering different ways for students to demonstrate their competencies. However, given their complex design, it is non-trivial to assume that items designed to operationalise the same competency work together to produce sum scores that can be inferred as measures of that competency. Using Rasch modelling, I found that although there was consistent evidence that items from a Swedish large-scale multi-format assessment measured some curriculum-defined competencies, evidence for the measurement of others was mixed, underscoring the opportunities and challenges of using such assessments to measure complex educational constructs at scale.

INTRODUCTION

Large-scale mathematics assessments serve important national and international purposes: monitoring educational systems, informing policy, and supporting teachers' judgments of student learning. Because they are administered to thousands of students under special conditions, they are necessarily constrained. Consequently, many large-scale assessments rely on selected-response and/or short constructed-response items, although students learn and communicate mathematics in various ways. The lack of concordance between how mathematics is taught and learned, and how it is assessed at scale, has long been recognised as a threat to the validity of inferences drawn from such assessments (Bennett et al., 2012; Braswell & Kupin, 1993; de Lange, 2007).

Multi-format assessments attempt to address this. Rather than assessing students only through short- and constructed-response items (mixed format), multi-format assessments include oral and investigative mathematics tasks, with the rationale that this provides opportunities to activate students' competencies in various ways. This approach strengthens the concordance between complex educational constructs and their operationalisations, thereby contributing to the validity of sum-score interpretations.

In Sweden, this design principle is explicit: annually administered national tests of mathematics at the end of compulsory schooling are composed of items of various formats designed to predominantly operationalise curriculum-defined mathematical competencies as five variables: communication, conceptual, problem-solving, reasoning, and strategic, summarised in Table 1:

2026. In H. Krzywacki, M. Hähkiöniemi, E. Haataja, M.S. Hannula & P. Hästö (Eds.). *Proceedings of the 49th Conference of the International Group for the Psychology of Mathematics Education: Research Reports*, pp. 403-410. PME.

Table 1: The mathematical competencies from the Swedish compulsory school curriculum (The Swedish National Agency for Education, 2011, p. 56).

Mathematical competency	Variable
Use mathematical forms of expression to discuss, reason and give an account of questions, calculations and conclusions.	Communication
Use and analyse mathematical concepts and their interrelationships.	Conceptual
Formulate and solve problems using mathematics and assess selected strategies and methods.	Problem-solving
Apply and follow mathematical reasoning.	Reasoning
Choose and use appropriate mathematical methods to perform calculations and solve routine tasks.	Strategic

However, multi-format design introduces challenges, not least because it implies different scoring practices, raising questions about measurement and reliability. Items also appear in different parts of the assessment, and elicit different types of cognitive and social engagement. These complexities make it non-trivial to assume that items intended to measure the same competency function collectively to produce sum scores that can be inferred as measures of that competency. Consequently, a critical question arises: To what extent do items from a large-scale, multi-format mathematics assessment work together to measure mathematical competencies?

Measurement theory

Typically, designers of mathematics assessment items aim to produce observations that serve as empirical evidence of what students have learnt, summarised in a total, or sum score. This approach views measurement as a process that produces information on properties of students, such that the information produced is in the form of values of the properties that assessors are interested in and intend to measure (Mari et al., 2023). For sum scores to serve as informative values, or measures, of “more” or “less” mathematical competency, a single variable is assumed to account for the consistent pattern in responses to items.

To test this assumption, a unidimensional model is necessary so that what is observed can be compared with what would be expected under the assumption that responses arise from a single variable. If empirical evidence suggests that the response data fit a unidimensional model, we gain support for interpreting sum scores as measures of a single property that the items were designed to measure.

The Rasch model (Rasch, 1980) is an example of a unidimensional model, where the only amount governing the probability π_{si} of student s correctly responding to item i is the difference between the item’s difficulty, β_i and the student’s competency, θ_s:

$$\pi_{si} = \exp(\theta_s - \beta_i) / [1 + \exp(\theta_s - \beta_i)] \quad (1)$$

Using this model, we can estimate a student's competency based on their total correct responses, or sum score. Similarly, we can estimate an item's difficulty based on the total number of students who get the item correct. The higher the competency of a student compared to an item's difficulty, the greater the probability of a correct response. Statistical analyses enable the examination of the extent to which the difference between these two primarily explains the variation in the empirical data. We can then examine any variation not explained by this to determine if there is a systematic structure that suggests an impact on scores, as this would indicate where item design may be functioning differently than intended.

METHOD

Instrument and sample

The Swedish national mathematics assessment for 15- and 16-year-old students completing their compulsory schooling consists of five sets of items, designed by test developers and content experts. Each set is intended to measure one mathematical competency (Table 1) and includes short-response, constructed-response, oral, and investigative tasks. Administered and scored by teachers using detailed scoring guides and analytic rubrics, approximately 100,000 students take the assessment annually. I analysed a subsample of 2,248 students assessed in the 2015/6 academic year following the rationales and protocols described next.

Data analysis

Data were analysed using RUMM2030 (Andrich et al., 2010), which is designed to facilitate examinations of the fit of data to the Rasch model. Data fitting the Rasch model means that the responses align with what would be expected from using only the sum scores to explain variation in responses. For an individual item to fit the Rasch model means that students with higher sum scores are more likely to answer an item correctly than students with lower sum scores, such that the observed pattern of responses across the performance range is similar to the pattern predicted by the Rasch model. A poorly fitting individual item indicates that the item may function differently from the other items in a set; that is, something other than the sum score describes the observed pattern. Reasons for this may vary and could include poor discrimination between students, guessing, ambiguity in item phrasing, or other factors. I counted an item as misfitting if it had a large item-fit residual and an item characteristic curve that differed from what would be expected under the model.

Examination of essential unidimensionality

After seeing how well the data fit the unidimensional Rasch model, I examined other possible dimensions in the data. To do this, I analysed residuals using principal components analysis to identify any remaining systematic structure in the unexplained

variation in responses, following Smith (2002). The first principal component represents the largest source of residual variance after accounting for a primary dimension, and I interpreted it as a potential secondary dimension. Items loading most strongly, either positively or negatively, on this first residual contrast are those most likely to challenge the assumption of unidimensionality. I estimated student competencies based on these item subsets and compared the estimates using paired t-tests. If fewer than 5% of these comparisons were statistically significant, I considered the secondary dimension to contribute negligible variance and evidence that the assumption of essential unidimensionality was satisfied.

Identification of local item dependence

To see the extent to which the items have more in common beyond the competency variable they were designed to measure, I wanted to identify locally dependent items. To do this, I analysed correlations among item residuals following Christensen et al. (2017). Except for the single dimension they have in common, responses to individual items should be independent of one another, as responses to one item should not be influenced by responses to another item beyond what can be explained by the sum score. Such dependence can arise from shared wording, common stimuli, or similar content; therefore, identifying local item dependence is necessary to ensure that items contribute unique information about the competency variable. I counted pairs of items as over-correlated if their residual correlation was 0.2 more than the average residual correlation.

Following the analyses described above, I considered data that (1) was similar to what would be expected assuming a unidimensional Rasch model, with non-significant chi-squared statistics, both overall and individually for at least 90% of items; (2) exhibited essential unidimensionality with less than 5% significant t-tests; and (3) demonstrated local item independence, with no over-correlated item residuals, as evidence that sets of items worked together to measure their respective competency variables. For these variables, I computed the average student competencies and the average item difficulties, and estimated reliability using coefficient alpha as an indicator of the consistency of student responses across items and the Person Separation Index (PSI) as an indicator of how well the items distinguished between students.

RESULTS

Table 2 overleaf summarises the results of the analyses described above. Whilst overall and individual item fit varied across competency variables, analyses of essential unidimensionality showed low proportions of significant person-level t-tests across all competency variables (2–4%), suggesting that secondary dimensions contributed negligible variance and providing evidence that each set of items satisfied the assumption of essential unidimensionality. Residual correlation analyses revealed little evidence of local item dependence within the sets of items, with over-correlations being rare or absent across competency variables.

Overall, items designed to measure the conceptual competency did not fit the Rasch model, indicating inconsistent functioning of items within this scale, reflected by almost one in five items not functioning as expected. Similarly, for items designed to measure the problem-solving competency, 60% of the items did not function as expected.

Overall and individually, items designed to measure the communication, strategic, and reasoning competencies fit the Rasch model, with consistent evidence that these sets of items produced sum scores that can be interpreted as reliable measures of their respective competency variables. Table 3 shows the average student competency and item difficulty, as well as reliability indices for these three competency variables:

Table 3: Reliability, Item Difficulty, and Student Competency Estimates

Variable	Reliability indices		Item difficulties (logits)		Student competencies (logits)	
	a	PSI	M	SD	M	SD
Communication	.71	.73	0.0	2.4	–0.1	1.7
Reasoning	.63	.70	0.0	1.0	–0.9	1.4
Strategic	.79	.76	0.0	1.5	0.9	1.5

Sets of items can thus be compared in terms of their overall difficulty for this sample of students. By substituting the values from Table 3 into equation (1), I calculated that for the communication variable, an average student had a 48% chance of correctly answering an item of average difficulty, whilst this probability was lower for the reasoning variable (29%). The easiest competency for the students was the strategic competency, with a higher probability of correct response of approximately 71%.

Table 2: Results of analyses

Analysis		Communication	Conceptual	Strategic	Problem-solving	Reasoning
Data fit to the Rasch model	No. of items	12	26	21	12	19
	Sample size[a]	250	250	250	250	250
	χ^2 (DF)	33.90 (24)	100.99 (51)*	62.75 (42)	67.49 (42)	20.25 (15)
Individual item fit	Sample size	2,248	2,248	2,248	2,248	2,248
	No. of misfitting items	1 (8%)	5 (19%)	1 (5%)	7 (58%)	0 (0%)
Essential unidimensionality	Sample size[b]	2,212	2,238	2,248	2,220	2,158
	No. of significant t-tests	75 (3%)	97 (4%)	51 (2%)	64 (3%)	56 (3%)
Local item dependence	No. of correlations	66	325	210	66	171
	No. of over-correlations	0 (0%)	2 (1.5%)	0 (0%)	1 (1%)	0 (0%)

Note. * $p < .001$. [a] To mitigate the sensitivity of the chi-square statistic to large samples, I used responses from a random subsample of 250 students. [b] I excluded students with zero or maximum sum scores because residuals cannot be estimated for them.

DISCUSSION

This study examined the extent to which items from a large-scale, multi-format mathematics assessment work together to measure curriculum-defined mathematical competencies. Using Rasch modelling, I found evidence that items designed to operationalise the communication, reasoning, and strategic competencies functioned coherently as unidimensional scales, supporting interpretations of their sum scores as measures of these competencies. In contrast, evidence for the measurement of conceptual and problem-solving competencies was mixed, with substantial proportions of misfitting items, raising questions about the extent to which items consistently captured these competencies.

Measuring competencies in multi-format assessments

These findings provide empirical support for the claim that multi-format assessments can measure some complex mathematical competencies at scale. In particular, the communication, reasoning, and strategic competencies showed good overall and individual item fit to the Rasch model, evidence of essential unidimensionality, and little indication of local item dependence, suggesting that, despite variation in item formats and scoring practices, these items collectively activated a common underlying variable for each competency. From a measurement perspective, this supports the validity of interpreting sum scores for these competencies as representing “more” or “less” of the intended construct.

At the same time, the results highlight important limits to what multi-format assessments can achieve. For the conceptual and problem-solving competencies, a large proportion of items did not fit a unidimensional model, suggesting that factors beyond a single underlying competency variable may have influenced responses to these items. For problem-solving and conceptual mathematical competencies in particular, the high rate of misfitting items suggests that the items may not have elicited sufficiently coherent response patterns to support a single sum-score interpretation, indicating that the item design or scoring processes did not adequately capture the complexity of these competencies.

For the competencies that did function coherently, differences in average student competency and item difficulty provide insight into student performance. The relatively low probability of success on reasoning items, compared to communication and strategic items, suggests that reasoning tasks may have posed higher cognitive demands for this cohort of students. This finding invites further investigation into whether these differences reflect curricular emphases, instructional practices, or features of task design and scoring.

Limitations and future research

This study was limited to a single assessment and relied on Rasch-based analyses available within one software environment, following only select protocols. Future research could extend this work by examining alternative modelling approaches,

exploring possible differential item functioning across student groups, or analysing misfitting items to understand how task features influence responses. Further work could also examine challenges teachers may face in assessing conceptual and problem-solving competencies in this context.

References

Andrich, D., Lyne, A., Sheridan, B., & Luo, G. (2010). *RUMM2030* [Computer software]. RUMM Laboratory.

Bennett, J., Tognolini, J., & Pickering, S. (2012). Establishing and applying performance standards for curriculum-based examinations. *Assessment in Education: Principles, Policy & Practice*, *19*(3), 321–339. https://doi.org/10.1080/0969594X.2011.614219

Braswell, J., & Kupin, J. (1993). Item formats for assessment in mathematics. In W. Ward C. & R. E. Bennett (Eds.), *Construction Versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment*. Routledge.

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical Values for Yen's Q3: Identification of Local Dependence in the Rasch Model Using Residual Correlations. *Applied Psychological Measurement*, *41*(3), 178–194. https://doi.org/10.1177/0146621616677520

de Lange, J. (2007). Large-scale assessment and mathematics education. In F. K. Lester (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning: A Project of the National Council of Teachers of Mathematics*. Information Age Publishers.

Mari, L., Wilson, M., & Maul, A. (2023). *Measurement Across the Sciences: Developing a Shared Concept System for Measurement*. Springer International Publishing. https://doi.org/10.1007/978-3-031-22448-5

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests (Expanded ed.)*. University of Chicago Press. (Original work published 1960).

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, *3*(2), 205–231.

The Swedish National Agency for Education. (2011). *Curriculum for the compulsory school, preschool class and school-age educare 2011 revised 2018*. https://www.skolverket.se/publikationer?id=3984

Acknowledgments

This work was supported by the ASSESS research school funded by the Swedish Research Council under Grant [2021–00540]. Special thanks to the test developers and content experts at PRIM, Stockholm University, Sweden.