Measuring Basic Mathematical Literacy in Elementary School

D. A. Federiakin, G. S. Larina, E. Yu. Kardanova

Received in September 2020

Denis Federiakin, Intern Researcher, Center for Psychometrics and Measurement in Education, Institute of Education, National Research University Higher School of Economics. Email: dafederiakin@hse.ru (corresponding author)
Galina Larina, Candidate of Sciences in Education, Research Fellow, Center for Psychometrics and Measurement in Education, Institute of Education, National Research University Higher School of Economics. Email: glarina@hse.ru
Elena Kardanova, Candidate of Sciences in Mathematical Physics, Associate Professor, Tenured Professor, Director of the Center for Psychometrics and Measurement in Education, Institute of Education, National Research University Higher School of Economics. Email: ekardanova@hse.ru

Address: 20 Myasnitskaya St, 101000 Moscow, Russian Federation.

Abstract

Measuring mathematical literacy is not easy as this construct is multicomponent and tasks often involve a lot of reading. Generally, intended users of test results want information about the overall level of respondents' mathematical literacy as well as its specific components. According to educational and psychological testing standards, reporting overall scores together with subscores simultaneously requires additional psychometric research to provide evidence for validity of all scores reported. A study performed shows that PROGRESS-ML, a test measuring basic mathematical literacy in elementary school pupils, can be used as a one-dimensional measure, allowing overall test scores to be reported. Meanwhile, reading skills do not contribute significantly to the probability of item responses, and subscores can be reported independently as complementary to the total score.

Keywords

basic mathematical literacy, complex construct, composite measure, PROGRESS-ML.

For citing

Federiakin D. A., Larina G. S., Kardanova E. Yu. (2021) Izmerenie bazovoy matematicheskoy gramotnosti v nachal'noy shkole [Measuring Basic Mathematical Literacy in Elementary School]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 2, pp. 199–226. https://doi.org/10.17323/1814–9545–2021–2–199–226

Translated from Russian by I. Zhuchkova.

Modern education and, broader, social sciences witness a growing demand for composite measures, i. e. instruments consisting of subscales that contribute in a particular manner (as a sum or a weighted sum of subscores) to the overall score on the test. Such instruments are indispensable to measure composite constructs, e. g. the so-called 21st-century skills or new literacies. These constructs comprise different components and are difficult to treat as a classical unidimensional (single-component) persons' characteristic. At the same time, prac-

titioners and policymakers who base their decisions on the results of measurement value the information about the overall level of performance as well as its individual components. This enriched information is valuable for improvement of the education system as a whole.

In terms of psychometrics, composite measures are multidimensional instruments that serve to evaluate the overall level of examinees' literacy as well as its individual components.

The Standards for Educational and Psychological Testing [American Educational Research Association, American Psychological Association, National Council on Measurement in Education 2014] state that (i) scores should not be reported for individuals unless their validity, comparability, and reliability have been established, and (ii) if a test provides more than one score, the psychometric quality of *all* the subscores should be established. These quality standards are important because just as inaccurate information at the total test score level may lead to decisions with damaging consequences, inaccurate information at the subscore level can also lead to incorrect remediation decisions [Sinharay, Puhan, Haberman 2011]. In academia, the use of low-quality subscores may result in misleading conclusions about the nature of the phenomenon analyzed.

Basic mathematical literacy is one example of composite constructs. There have been numerous attempts to define basic literacies with regard to most diverse aspects of the construct, from content area to their necessity for a life in a modern world [Froumin et al. 2018]. What all these definitions have in common is that they define literacy as the ability to solve daily life problems. Due to diversity of such problems, instruments measuring basic literacies place task models into widely heterogeneous contexts in an effort to measure effectiveness of solving specific types of problems. That is how such instruments (measuring basic mathematical literacy) become composite: they represent a certain integrated measure of an examinee's ability to solve various problems involving math operations. A number of researchers believe that such a structure of composite measures leads to nuisance skills "interfering" with problem solving (see, for example, [Grimm 2008]). As a consequence, one of the key psychometric issues with such measures is the demand for valid and psychometrically sound total score on the test as well as subscores on its individual components.

The present study seeks to test and substantiate different approaches to modelling data of the PROGRESS-ML test designed to measure basic mathematical literacy in elementary school. The article is structured as follows. First, an overview of theoretical challenges in measuring basic mathematical literacy is given. Next, the theoretical framework and structure of PROGRESS-ML are described. As long as the focus of this article is on psychometric modelling, it does not investigate deep into different ways of interpreting the theoretical framework of the test, its design peculiarities, or the relationship

between test content and elementary school curriculum — without belittling the significance of such work. Further on, the existing psychometric approaches to composite measurement models are described, and a rationale for choosing one of them is provided. After that, empirical data collected with PROGRESS-ML is analyzed. At the first stage of this analysis, the reportability of the total score on the basic mathematical literacy test is evaluated. The second stage tests the hypothesis that reading skills make an important contribution to performance on the basic mathematical literacy test. Finally, the third stage evaluates the reportability of subscores gained in specific content areas and measuring specific types of cognitive processes. The final section of this article provides a psychometric interpretation of the results obtained and describes their contribution to educational and psychological measurement methodologies.

1. Challenges in Measuring Basic Mathematical Literacy

Mathematical literacy is a competency that everyone needs to handle everyday situations, such as grocery shopping, cooking, bill payment, etc. In Russia, mathematics has been traditionally regarded as a school subject that is critically involved in other academic disciplines and as a unique tool for promoting "cognitive development in mainstream schools" [Kozlov, Kondakov 2009]. However, measuring the mathematical literacy of students is difficult due to challenges in defining the construct.

Researchers and practitioners around the world have not yet come up with any consensual concept of mathematical literacy that could be defined by a certain set of knowledge and skills [Jablonka 2003]. Mathematical literacy includes computational skills, knowledge and understanding of fundamental mathematical concepts, spatial thinking, real-world problem solving skills, ability to implement complex problem-solving algorithms and to analyze data from graphs. Even tests measuring mathematical literacy in preschool-age children are designed using divergent sets of basic knowledge and skills from an impressive list, which includes mental calculation, numerical knowledge, number sense, object counting, shape recognition, measuring, etc.

Challenges in defining and, consequently, measuring mathematical literacy of school students come, in particular, from the dependence of mathematics as a school subject on the specific school curriculum which is designed in accordance with the formal goals of mathematics education. For example, a number of curricula focus on the systemic role of quality mathematics education for scientific progress (e. g. the 2013 Concept of Mathematical Education in the Russian Federation[1]). Another approach, supported by some of the existing international

[1] Concept of Mathematical Education in the Russian Federation. Ministry of Education and Science of the Russian Federation: http://www.math.ru/conc/vers/conc-3003.pdf

surveys of education quality such as PISA[2] [OECD2019], makes it a priority that by the time school students complete the obligatory stage of education, they should possess the skills of handling various everyday situations, from grocery shopping to arranging furniture at home.

The only consensus achieved so far among researchers is on the development trajectories of particular mathematical skills in preschool and school [Purpura, Baroody, Lonigan 2013]. It is still unclear, for instance, when and how young children develop the ability to understand and process symbolic numerical representations (digits and numbers) [Benoit et al. 2013; Kolkman, Kroesbergen, Leseman 2013], to what extent the development of mathematical skills depends on other cognitive abilities [Peng et al. 2016; Toll et al. 2015], or which early competencies are the best predictors of formal math performance [Chen, Li 2014; Libertus et al. 2016; Schneider et al. 2017]. As for school math instruction, skills acquired by students become more and more divergent with every subsequent grade due to differences in curricula, which makes it virtually impossible to identify a single construct to describe mathematical literacy.

Another challenge in defining mathematical literacy is that this construct demonstrates not only what knowledge students possess but also in which situations they can apply it. Different content areas are assessed using tasks designed to induce different levels of cognitive load. For example, a problem in which students are asked to estimate the height of a column in a histogram requires less cognitive effort than the one in which students have to process information from a graph which they see for the first time. For this reason, a number of assessments also measure mathematical literacy through the prism of the cognitive processes involved in problem solving. For instance, PISA measures the level of 15-year-old students' theoretical knowledge in specific content domains (quantity, space and shape, etc.) and their ability to apply this knowledge at every stage of problem solving [OECD2013].

Finally, challenges in measuring mathematical literacy also arise from the fact that mathematical skills are closely associated with reading literacy. In a longitudinal study conducted on a low-income sample of students from the Chicago Public Schools, third grade reading comprehension was found to be a positive significant predictor of gains in mathematics skills up to eighth grade controlling for prior mathematics performance. The largest effects of reading achievement were shown for the problem solving and data interpretation [Grimm 2008]. Similar results were obtained in analyses comparing the mean scores of students across 22 countries that took part in two international assessments conducted in 2003: PISA reading scores are highly corre-

[2] Program for International Students Assessment, administered every three years to measure mathematical, reading and scientific literacy of 15-year-old school students.

lated with performance in the TIMSS "Data" content domain,[3] which measures the ability to read graphs and interpret diagrams (r = 0.91, the correlation coefficients with other content domains varied from 0.57 to 0.79) [Wu 2010]. Math disabilities (dyscalculia) often co-occurs with dyslexia[4] [Joyner, Wagner 2020], children with comorbid mathematics and reading difficulties performing worse in mathematics than children with dyscalculia only [Jordan, Hanich, Kaplan 2003].

The relationship between mathematical and reading literacy is complex and still largely dubious. On the one hand, both constructs can involve the same cognitive processes: for instance, numerical cognition was found to depend on language skills in early childhood [Gelman, Butterworth 2005]. Miscomprehension of the language of word problems can be another possible explanation of this relationship [Cummis et al. 1988]. According to Anne Newman's hierarchy of error in written mathematical tasks [Newman 1977; Casey 1978], the first two stages in solving any word problem depend directly on the reading skills of decoding and comprehension, which imply reading the task carefully, understanding the context of the problem and the question asked, and collecting all the necessary information. Errors at these stages account for 12 to 22% of all solution errors (see, for example, a review of studies in [Clements, Ellerton 1996]). Therefore, obtaining the correct solution to any written mathematical task depends on whether the pupil makes errors during the first two steps in attempting to answer it, e. g. by reading "minutes" instead of "minus" [Clements 1980].

As we can see, mathematical literacy is not a binary but complex and multifaceted construct that involves a broad array of mathematical skills. It would be difficult to evaluate the development of mathematical skills — from informal knowledge in preschool to sophisticated methods in high school — on a single scale. Besides, mathematical skills involved in solving word problems and reading graphs largely depend on reading comprehension. The construct's complexity and relationship with other intellectual skills make assessment of mathematical literacy even more challenging and require advanced psychometric models to validate the instruments.

2. The PROGRESS-ML Basic Mathematical Literacy Test

The PROGRESS-ML Basic Mathematical Literacy Test is part of the PROGRESS instrument[5] designed to measure basic literacies in elementary school pupils for the purposes of learning process support and improvement. PROGRESS measures basic literacies in mathemat-

[3] Trends in Mathematics and Science Study, administered every four years to measure mathematical and scientific literacy of fourth- and eighth-grade students.

[4] A learning disorder that involves difficulty reading and writing.

[5] The instrument was designed by the Institute of Education (Higher School of Economics).

Table 1. **Content areas covered by the PROGRESS-ML test.**

Content area	Number of items	Description
Spatial concepts	7	The items measure pupils' ability to understand spatial relationships and mentally represent 2D and 3D objects. Children are required to not only recognize different geometric shapes but also visualize new geometric objects by combining 2D or 3D figures into one.
Measuring	6	Successful performance on this module demonstrates understanding that numbers can not only be used for specifying position in a sequence but also serve as attributes (length, surface area, etc.). The tasks evaluate children's ability to manipulate numbers as measures.
Patterns and sequences	6	The items measure pupils' ability to recognize and extend arithmetic and geometric sequences as well as their level of algorithmic universal learning activities<FootnoteStart:>The part of the Russian Federal Educational Standard.<FootnoteEnd:>. To solve the tasks in this module correctly, pupils must understand how sequences work (one or more rules).
Modelling	6	This module measures the ability to translate models represented with words or geometric sequences into mathematical formulations. As in the Patterns and Sequences module, these items also measure pupils' level of algorithmic universal learning activities. However, students are required here not only to recognize the model but also to represent it using formal mathematical notation.
Data handling	5	The tasks evaluate children's ability to comprehend and interpret information from charts and graphs. In addition, pupils must understand how to perform computations using graphic information and make judgments with additional information.

ics, language, reading comprehension, and vocabulary. Measurements are performed as computerized adaptive testing.

PROGRESS-ML is designed to evaluate how well students perform in mathematics after two years of schooling. In today's postindustrial world, mathematics education must prepare students to solve problems in ever-changing environments, e. g. make quick decisions and adapt to new situations, be able to solve unfamiliar problems and navigate easily in large quantities of information. All of this transforms the concept of basic mathematical literacy. For instance, such literacy comes to involve a broader range of skills due to the increased number of data handling or problem solving contexts. The test was designed around the following definition: "Basic mathematical literacy (including data handling) is the ability to apply mathematical knowledge, reasoning, and models to real-world problems, including those in digital environments" [Froumin et al. 2018].

Table 2. **Cognitive domains evaluated in the PROGRESS mathematics test**

Cognitive domain	Number of items	Description
Knowing	12	Covers the knowledge of mathematical facts, which is fundamental to solving any problem. For example, students may be asked to add or subtract two whole numbers, estimate the height of a column in a simple histogram, or calculate how many times a geometric shape can be fitted into a picture.
Applying	14	Focuses on the ability of students to apply the acquired knowledge and skills to solve problems or situations with well-known contexts and solution algorithms. For example, a problem may ask for the rule behind a number sequence or a shape pattern.
Reasoning	4	The tasks require careful analysis of the information given in the problem in order to connect facts from different content areas and consider alternative solutions. These items use unfamiliar contexts and thus require more focus.

The test is comprised of 30 dichotomous multiple-choice items. Items were selected to meet the definition of basic mathematical literacy and at the same time make allowance for the elementary school curriculum. As a result, five content areas were identified (Table 1), and all the items were grouped into the respective modules.

Furthermore, PROGRESS-ML evaluates students' cognitive processes that are necessary to solve the tasks. The test was designed in accordance with the three cognitive domains identified within the theoretical framework of the TIMSS fourth-grade assessment [Harris, Martin 2013]: knowing, applying, and reasoning. At the same time, the test items were designed to fit the Rasch model.

The TIMSS taxonomy of cognitive domains is similar to Bloom's taxonomy of educational objectives [Bloom 1956], yet it is not identical to it and identifies only three domains, not six. Besides, a fundamentally distinctive feature of the TIMSS taxonomy is that the three cognitive domains are not ranked by difficulty or degree of abstractness, so each domain has tasks of varying difficulty.

Problem solving in PROGRESS-ML involves all the three cognitive domains of the TIMSS taxonomy. Most of the tasks are distributed approximately equally between the domains of Knowing and Applying, while the remaining small portion of items targets Reasoning[6] (Table 2).

Some of the PROGRESS-ML tasks (about 50%) can be regarded as reading-intensive, as students are supposed to read and understand the problem setting.

[6] The correctness of classifying the items under the three cognitive domains was verified by experts.

Therefore, PROGRESS-ML is a structurally complex measure that covers five content areas and evaluates three cognitive domains. Such measures are called composite and imply reporting the total score on the test (in this case, basic mathematical literacy) as well as the subscores (in this case, for content areas and/or cognitive domains).

3. Psychometric Properties of Composite Measures

Psychometric analysis of composite measures involves a few steps. First of all, it is necessary to examine whether the test is essentially unidimensional. If yes, then it is safe to report the total score on the test (given that it has been proven to be valid and psychometrically sound). If the test is not unidimensional, then multidimensional models should be used; in this case, the total score cannot be reported until a secondary analysis with hierarchical models is conducted [Schmid, Leiman 1957]. Two types of hierarchical models are especially popular: bifactor models [Reise 2012] and higher-order models [Gignac 2008]. Although both types of models have some algebraic similarities and imply reporting the total score on the test, solutions of these models are interpreted in different ways [Rijmen 2010; Mansolf, Reise 2017]. Higher-order models measure the general factor which is manifested in subscores, while bifactor models separate the effects of the general factor from those of subscores.

If the test is designed to report subscores (e. g. scores in specific content areas or cognitive domains) in addition to the total score, three approaches to modelling are possible. The first one consists in applying a unidimensional model to each subscale individually [Davey, Hirsch 1991]. However, subscales do not normally contain many items, so this results in compromised reliability and unacceptably high measurement error. Under such conditions, subscore reporting appears to be inappropriate [American Educational Research Association, American Psychological Association, National Council on Measurement in Education 2014].

The second approach is based on between-item non-compensatory multidimensional item response models [Reckase 2009]. Such models bring together a few unidimensional models in a single likelihood equation. Each latent trait of a variable is estimated based on examinees' answers to specific items only, controlling for the correlations among the latent variables. Therefore, multidimensional models use the information from each dimension to model the probability of item responses not as a function of a single latent variable but as a function of a multidimensional latent distribution of respondents (they consider correlations among latent dimensions). As a result, measurements will be more reliable than with the previous approach, adding more value to subscore reporting. Between-item non-compensatory multidimensional item response models can be used in analysis of collateral information, i. e. any information about the test, examinees, or relations between them which does not affect parameter interpretation

when added to the model but allows reducing uncertainty in parameter evaluation [Wang, Chen, Cheng 2004].

The third approach uses bifactor models. Hypothetically, they allow reporting total scores together with subscores as supplementary and independent information. However, studies show that subscores obtained using bifactor models tend to lack reliability as they describe information which was not extracted using the total score and thus often capture random noise in the data [Haberman, Sinharay 2010].

To summarize, data from composite measures cannot be used without prior psychometric analysis to ensure reliability and reportability of the total score and subscores.

3.1. Research Methodology

3.1.1. Sampling and Procedure

The study was conducted on a regionally representative sample of 6,078 third-grade pupils from two regions of Russia. The average age of pupils was 9.06 (SD = 0.46); girls accounted for 52.36% of the sample. Computer adaptive testing with a stopping rule was administered to the sample.

3.1.2. Data Analysis Methods

Psychometric analysis was performed entirely within the framework of Item Response Theory (IRT) [van der Linden 2018]. IRT postulates that items as well as examinees have some latent parameters, the interactions among which determine the probability of observing a correct response of each examinee to each item. Specifically, analysis was done through Rasch measurement modelling [Rasch 1993], which posits that items only differ in their difficulty, compared to other IRT models that use more item parameters [Wright, Stone 1979]. The specific objectivity (that guarantees separation of the parameters of items and examinees, a clear hierarchy of items along the entire continuum of ability, and numerical stability of models) justify the selection of Rasch models for psychometric analysis in the present study.

An important advantage of Rasch models is that they allow assessing the variance of latent abilities (random effects) [Kamata 2001]. In case the variance of an ability approaches zero, this ability does not load on the items enough to affect the probability of item responses. Besides, just as all multidimensional IRT models, Rasch models allow estimating directly the correlations and variances of latent abilities "cleansed" of random error variance in the distribution of scores across the items [De Boeck, Wilson 2004].

Unidimensional and multidimensional models were used in the present study to test a series of hypotheses about possible latent abilities required to solve the tasks.

The Rasch unidimensional measurement model [Wright, Stone 1979] was used to test the hypothesis that the total score on the test can be safely reported to individuals. Test dimensionality was analyzed by applying principal component analysis (PCA) to standardized model residuals, which represent standardized deviations of responses from values expected under the employed model [Linacre 1998; Smith 2002].

Thereby, the variance unexplained by the model is decomposed into components. In theory, if the unidimensionality assumption is confirmed, correlations among the residuals should be close to zero. In this case, PCA will not extract components that systematically explain more variance than others. Hence, the distribution of the variance explained by the components will be close to uniform. It is also generally accepted that if the eigenvalue of the component explaining most of the variance is less than 2, the distribution of residuals captures random noise and the test is thus unidimensional [Linacre 2021]. Otherwise, the test has more than one dimension.

The item fit was assessed using unweighted (OutFit) and weighted (InFit) mean square (MnSq) statistics [Linacre 2002]. These statistics are also based on standardized residuals [Wright, Masters 1990] and have an expected value of 1. The present study considers the acceptable range of fit statistics to be (0.75; 1.33), which is treated in research literature as a range productive for measurements [Adams, Khoo 1996].

Finally, psychometric soundness of the total test score was assessed by measuring its reliability and measurement error.

The hypothesis on the importance of reading skills in solving mathematical problems was tested using within-item compensatory multidimensional models [Adams, Wilson, Wang 1997]. Such models imply that more than one latent ability is required to solve any task (e. g. mathematics and reading skills), and linear combinations of such latent traits modulate the actual probability of item responses.

Fifteen most reading-intensive items were selected to assess reading skills. Next, an incomplete bifactor Rasch model was calibrated, which allows that mathematical literacy is measured by all the items while a selected group of items additionally measures reading comprehension. This model can be regarded as an extended Rasch testlet model [Paek et al. 2009], an oblique bifactor IRT model that allows direct estimation of correlations between two latent abilities. Therefore, the primary dimension that loads on all the items can be interpreted as basic mathematical literacy "cleansed" from the contribution of reading. At the same time, the additional ability that loads on the selected 15 tasks can be interpreted as reading skills that are manifested in the basic mathematical literacy test.

Finally, the subscore reporting hypothesis was tested using Rasch between-item non-compensatory multidimensional models [Adams, Wilson, Wang 1997], which imply that each item belongs to only one particular subscale and there is no general factor. The preference of a between-item model in the present study has to do with criticism against general factor models, in particular because the subscores obtained in such models are difficult to interpret. For a bifactor model, this involves low subscore reliability and constraints imposed on the variance–covariance matrix of latent variables that make interpretation of any obtained scores challenging [Bonifay, Lane, Reise 2017]. At the same time, higher-order models do not use subscores at all be-

cause the key information about the construct components is already described by the general factor.

Goodness of the global fit of all models described above was compared to that of the Rasch unidimensional model, which served as the baseline. The global fit was assessed using the Akaike Information Criterion (AIC) [Akaike 1974]) and the Bayesian Information Criterion (BIC) [Schwarz 1978]. These information criteria include a penalty for extra-parameters (AIC) with respect to the sample size (BIC). They estimate the relative quality of models, the ones with the lowest values being preferred. Local fit was assessed using the OutFit and InFit statistics described above.

Reliability of all the models was assessed using expected a posteriori (EAP) estimation of ability [Bock, Mislevy 1982]. The EAP method works particularly well with multidimensional measures as it utilizes information about the multidimensional ability distribution as well as the entire patterns of item responses to measure ability along each dimension. The use of EAP in this case is justified because the instrument is not designed for the analysis on particular dimensions separately without using the other ones. Therefore, the use of multidimensional IRT models in this context involves a different understanding of collateral information. In this case, for each subscale, data from all the other subscales (included in covariance matrix of the latent variables) represents collateral information [Wu, Tam, Jen 2016]. As a result, measurement reliability improves. The posterior standard deviation can be treated as the standard error measurement. The ultimate reliability is determined by the ratio of this error variance to the estimated latent ability variance [Adams 2005].

All the models applied can be regarded as special cases of the Multidimensional Random Coefficients Multinomial Logit Model [Adams, Wilson, Wang 1997]. All the models were estimated using a Quasi-Monte Carlo algorithm in the TAM package (version 3.5–19) [Robitzsch, Kiefer, Wu 2020] for *R* (version 3.6.2) under the Marginal Maximum Likelihood estimator, which makes a parametric assumption about (multidimensional) normality of the ability distribution [Bock, Aitkin 1981]. All the models were identified by keeping the sample mean at zero for each dimension. This is especially important when identifying within-item multidimensional models. Dimensionality was assessed using the 'psych' package (version 1.9.12.31) [Revelle 2020]. Residuals were evaluated using the parameter estimates obtained by Warm's weighted maximum likelihood estimation [Warm 1989], which has proved to be less biased than other methods for ability point-estimation that are popular in IRT.

4. Results

4.1. Testing the Total Score Reporting Hypothesis

Table 3 displays the results of testing the conformance of items to the unidimensional Rasch model. All the items show a reasonable fit with goodness-of-fit statistics staying within the acceptable range.

Table 3. **Testing the conformance of items to the unidimensional Rasch model.**

Module	Item	Number of responses	Difficulty	SE	InFit MnSq	OutFit MnSq
Spatial concepts	I01	6,041	–1.18	0.03	0.97	0.96
	I02	5,975	–0.56	0.03	1.03	1.04
	I03	5,997	–0.23	0.03	1.05	1.06
	I04	5,430	0.37	0.03	1.04	1.06
	I05	4,987	0.69	0.03	1.00	0.99
	I06	4,125	1.19	0.04	1.08	1.14
	I07	3,098	1.94	0.05	0.98	0.97
Measuring	I08	5,843	–1.51	0.03	1.07	1.13
	I09	5,860	–1.47	0.03	1.03	1.05
	I10	5,811	–1.69	0.04	0.95	0.88
	I11	5,626	–0.76	0.03	1.02	1.01
	I12	5,375	–0.65	0.03	0.93	0.89
	I13	5,080	–0.59	0.03	0.95	0.92
Patterns and sequences	I14	5,560	–2.41	0.05	0.95	0.87
	I15	5,473	–2.06	0.04	1.02	1.04
	I16	5,340	–1.42	0.04	0.96	0.92
	I17	5,009	–1.16	0.03	1.00	1.03
	I18	4,755	–0.56	0.03	1.06	1.07
	I19	4,398	–0.12	0.03	1.03	1.04
Modelling	I20	4,603	–0.62	0.03	0.98	0.98
	I21	4,263	–0.41	0.03	0.89	0.86
	I22	3,826	–0.37	0.04	1.10	1.14
	I23	3,013	1.78	0.05	1.04	1.11
	I24	2,423	0.72	0.05	1.06	1.12
	I25	1,702	0.73	0.05	1.00	1.01
Data handling	I26	2,808	–2.31	0.06	0.98	0.98
	I27	2,469	–2.20	0.06	0.93	0.83
	I28	2,320	–1.35	0.05	0.89	0.83
	I29	1,969	–1.60	0.06	0.88	0.76
	I30	1,708	1.05	0.06	0.95	0.93

Table 4. **Comparing the baseline model with the model evaluating the contribution of reading comprehension.**

Model	Log-likelihood	Sample size	Number of parameters	AIC	BIC
Unidimensional	144,255.6	6,078	31	144,318	144,526
With reading comprehension as a latent dimension	142,638.5		33	142,705	142,926

PCA of model residuals reveals that the eigenvalue of the first component is 1.45, which accounts for 4.2% of the residual variance. The eigenvalues of the following four components fall within the range of (1.15; 1.20), and the variance is distributed approximately uniformly among the principal components (about 4%). Consequently, the unidimensional model describes the distribution of response probabilities adequately and the test can be treated as unidimensional.

The EAP reliability of the unidimensional model in measuring mathematical literacy equals 0.76 (ability variance = 0.93); Cronbach's alpha is 0.81, which is fairly high.

Therefore, the test can be considered unidimensional based on the above analysis even despite different methods of item grouping, which means that the total score on the mathematical literacy test can be reported as psychometrically stable.

The unidimensional Rasch model served as the baseline for comparing all the other models.

4.2. Testing the Hypothesis about the Contribution of Reading Skills

Table 4 shows the results of comparing the unidimensional Rasch model with the Rasch model calibrated for measuring the contribution of reading skills in the probability of item responses.

The model measuring the contribution of reading comprehension looks more preferable in terms of global fit. However, the variance of reading skills is only 0.02, which is 52.35 times lower than that of mathematical literacy (0.89) from this model. Similarly, the reliability of mathematical literacy was found to be 0.75, which is 41.83 times higher than that of reading comprehension (0.01). Consequently, examinees do not differ in the latent ability measured by the selected 15 items. Furthermore, Pearson's correlation coefficient between the dimensions of reading skills and mathematical literacy is insignificantly different from zero ($r = 0.01$, $p > 0.05$, according to the *t*-test for Pearson's correlation), contradicting previous findings (e. g. [Grimm 2008]). The reason may consist in low variance and, as a result, low reliability of reading skills measurement: with such reliability and variance values, differences in examinees' reading skills are almost entirely attributable to random fluctuations. Based on the results of analysis, one may say that reading comprehension may contribute to scores in this

Table 5. **Comparing the unidimensional Rasch model with the Rasch models for content areas and cognitive domains.**

Model	Log-likelihood	Sample size	Number of parameters	AIC	BIC
Unidimensional	144,255.6	6,078	31	144,318	144,526
Content areas	143,875.7		45	143,966	144,268
Cognitive domains	143,965.4		36	144,037	144,279

Table 6. **Reliability, variance, and correlation coefficients for the content areas model.**

Dimension (content area)	Spatial concepts	Measuring	Patterns and sequences	Modelling	Data handling
Spatial concepts		0.85	0.80	0.83	0.80
Measuring			0.85	0.90	0.83
Patterns and sequences				0.86	0.84
Modelling					0.83
Variance	0.89	1.23	1.12	1.06	2.95
Reliability	0.68	0.71	0.67	0.68	0.63
Number of items	7	6	6	6	5

test just as in other instruments, but the contribution is so small that it is essentially unidentifiable.

4.3. Testing the Subscore Reporting Hypothesis

Table 5 shows the results of comparing the unidimensional Rasch model with the Rasch models calibrated for validating the theory-based content areas and cognitive domains.

Data from Table 5 indicates that either of the two concurrent models fits data better than the unidimensional model. Consequently, the taxonomy behind test design actually guides the examinees towards expected behavior.

The results of analyzing the variance–covariance matrices and reliability coefficients for each dimension of the models applied are given in Tables 6 and 7.

The coefficients in Table 6 allow concluding that, firstly, reliabilities of all the dimensions are sufficiently high for using the test as a longitudinal survey instrument. Despite the small number of items, the basic mathematical literacy test can be used for longitudinal assessments thanks to the reliability analysis method (EAP estimation of ability) applied under multidimensional model. Secondly, all the content areas correlate with one another at approximately the same level (0.8–0.9),

Table 7. **Reliability, variance, and correlation coefficients for the cognitive domains model.**

Dimension (cognitive domain)	Knowing	Applying	Reasoning
Knowing		0.95	0.85
Applying			0.85
Variance	1.37	0.82	0.60
Reliability	0.75	0.74	0.61
Number of items	12	14	4

adding to the argument for unidimensionality even though the multidimensional model fits the data statistically better. These results indicate that items from every content area load equally on the general factor of mathematical literacy.

A similar inference can be made about the model assessing the validity of cognitive domains (Table 7): reliabilities of the dimensions are sufficiently high for using the test as a longitudinal survey instrument. It is worth focusing on the reasoning dimension which consists of four dichotomous items only. Such a small number of items basically makes raw subscores on this scale unreportable, unlike the scores on this latent dimension. Analysis of the correlation matrix of cognitive domains also supports the hypothesis about the test being essentially unidimensional.

5. Conclusion

Social sciences have been using increasingly more often composite measures, which imply reporting the total score as well as subscores. One of the possible strategies of applying the measurement results obtained with such instruments could consist in reporting raw subscores [Wilson, Gochyyev 2020]. However, psychometric analysis is required to find out how much value raw subscores add to the total score [Haberman 2005]. In most cases, raw subscores are not psychometrically sound, in particular due to their low reliability [Haberman, Sinharay 2010].

Another, more popular strategy suggests using complex psychometric models that are often difficult to interpret [Bonifay, Lane, Reise 2017]. This primarily applies to bifactor models: the estimation of their parameters requires essential, sometimes unrealistic assumptions that make it extremely hard to interpret the test results [Wilson, Gochyyev 2020]. Even with the recent advances in oblique bifactor models, it is still a long way to developing a single framework for their interpretation and completing the analysis of their psychometric

properties [Kanonire, Federiakin, Uglanova 2020]. Yet another strategy consists in using higher-order models [Gignac 2008], in which subscores work as indicators of the general factor. However, such models do not imply subscore reporting at all, which limits their applicability without belittling their academic value.

Subscores are in high demand among practitioners as they not only measure the level of performance on a construct but also describe how exactly it was achieved. To meet this demand, researchers involved in international school student assessments use the fact that a person's mean score across all the subscales in a multidimensional model is equal to their score in a unidimensional model, provided that subscales undergo linear transformation into scales with identical numerical parameters (for example, with the mean of 500 and SD of 100) [Foy, Yin 2015]. This allows researchers to avoid restricting the interpretation to a single model and avoid the use of overparametrized psychometric models [Brandt, Duckor, Wilson 2014].

A similar strategy was used to report scores on the PROGRESS-ML mathematical literacy test. To provide justification for using the total score, the items were tested for unidimensionality using PCA of model residuals and goodness-of-fit statistics. Results indicate that the test can be used as a unidimensional measure, which means that the overall mathematical literacy score can be safely reported to end users.

Next, the reportability of subscores in addition to the total score was tested. Since mathematical literacy is a complex multicomponent construct, its subscores have added value for end users. Subscore reportability was assessed to enhance the applied value of test results in compliance with the Standards for Educational and Psychological Testing [American Educational Research Association, American Psychological Association, National Council on Measurement in Education 2014]. Item recalibration in other models — the most suitable approach for this measure — showed that subscores obtained on the construct components are psychometrically sound and can be reported to end users.

Therefore, the total score is the key result of the test. However, additional item recalibrations across the content areas and cognitive domains allow describing how exactly the overall score on the test was achieved. In fact, the total score is decomposed into its components. Information about correlations among the subscores makes it possible to use even relatively small scales (e. g. the 'Reasoning' scale from the model for cognitive domains consists of four items only) with fairly high reliability.

In addition, the contribution of reading skills in the probability of item responses was assessed. Expert evaluations were used to measure the reading intensity of items, allowing to identify the second potential dimension and evaluate its variance and correlation with the primary dimension. Reading comprehension was found to make no significant contribution to the probability of item responses. The approach tested in the present study has a great potential for generali-

zation and can be used to analyze the contribution of nuisance dimensions in other measures.

This article describes the psychometric properties of the PROGRESS-ML basic mathematical literacy test. A three-stage analysis showed that (i) this test can be used as a unidimensional instrument, i. e. its total score can be reported to end users; (ii) reading comprehension does not contribute significantly to the probability of item responses; and (iii) subscores obtained on test components can be reported to end users in addition to the total score.

The reported study was funded by the Russian Foundation for Basic Research (RFBR) as part of research project no. 19–29–14110.

References

Adams R. J. (2005) Reliability as a Measurement Design Effect. *Studies in Educational Evaluation*, vol. 31, no 2–3, pp. 162–172. doi: https://doi.org/10.1016/j.stueduc.2005.05.008.

Adams R. J., Khoo S. T. (1996) *Quest*. Melbourne, Australia: Australian Council for Educational Research.

Adams R. J., Wilson M., Wang W. C. (1997) The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement*, vol. 21, no 1, pp. 1–23. doi: https://doi.org/10.1177/0146621697211001.

Akaike H. (1974) A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, vol. 19, no 6, pp. 716–723. doi: https://doi.org/10.1109/TAC.1974.1100705.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014) *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Benoit L., Lehalle H., Molina M., Tijus C., Jouen F. (2013) Young Children's Mapping between Arrays, Number Words, and Digits. *Cognition*, vol. 129, no 1, pp. 95–101. doi: https://doi.org/10.1016/j.cognition.2013.06.005.

Bloom B. S. (ed.) (1956) *Taxonomy of Educational Objectives: The Classification of Educational Goals*. New York: Longmans, Green and Company.

Bock R. D., Aitkin M. (1981) Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm. *Psychometrika*, vol. 46, no 4, pp. 443–459. doi: https://doi.org/10.1007/BF02293801.

Bock R. D., Mislevy R. J. (1982) Adaptive EAP Estimation of Ability in a Microcomputer Environment. *Applied Psychological Measurement*, vol. 6, no 4, pp. 431–444. doi: https://doi.org/10.1177/014662168200600405.

Bonifay W., Lane S. P., Reise S. P. (2017) Three Concerns with Applying a Bifactor Model as a Structure of Psychopathology. *Clinical Psychological Science*, vol. 5, no 1, pp. 184–186. doi: https://doi.org/10.1177/2167702616657069.

Brandt S., Duckor B., Wilson M. (2014) *A Utility Based Validation Study for the Dimensionality of the Performance Assessment for California Teachers (PACT)*. Paper presented at Annual Conference of the American Educational Research Association. Available at: https://www.researchgate.net/publication/281645866_A_Utility_Based_Validation_Study_for_the_Dimensionality_of_the_Performance_Assessment_for_California_Teachers_PACT (accessed 2 April 2021).

Casey D. P. (1978) Failing Students: A Strategy of Error Analysis. Aspects of Motivation (ed. P. Costello), Melbourne: Mathematical Association of Victoria, pp. 295–306.

Chen Q., Li J. (2014) Association between Individual Differences in Non-Symbol-

ic Number Acuity and Math Performance: A Meta-Analysis. *Acta Psychologica*, vol. 148, May, pp. 163–172. doi: https://doi.org/10.1016/j.actpsy.2014.01.016.

Clements M.A.K. (1980) Analyzing Children' s Errors on Written Mathematical Tasks. *Educational Studies in Mathematics*, vol. 11, no 1, pp. 1–21. doi: https://doi.org/10.2307/3482042.

Clements M.A., Ellerton N. (1996) *The Newman Procedure for Analysing Errors on Written Mathematical Tasks*. Available at: https://compasstech.com.au/ARNOLD/PAGES/newman.htm (accessed 2 April 2021).

Cummins D.D., Kintsch W., Reusser K., Weimer R. (1988) The Role of Understanding in Solving Word Problems. *Cognitive Psychology*, vol. 20, no 4, pp. 405–438. doi: https://doi.org/10.1016/0010–0285(88)90011–4.

Davey T., Hirsch T.M. (1991) *Concurrent and Consecutive Estimates of Examinee Ability Profiles*. Paper presented at the Annual Meeting of the Psychometric Society (New Brunswick, NJ).

De Boeck P., Wilson M. (2004) *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York, NY: Springer Science & Business Media.

Foy P., Yin L. (2015) Scaling the TIMSS2015 Achievement Data. *Methods and Procedures in TIMSS2015* (eds M.O. Martin, I.V.S. Mullis, M. Hooper), Chestnut Hill, MA: Boston College. P. 13.1–13.62. Available at: https://timssandpirls.bc.edu/publications/timss/2015-methods/chapter-13.html (accessed 2 April 2021).

Froumin I.D., Dobryakova M.S., Barannikov K.A., Remorenko I.M. (2018) *Universalnye kompetentnosti i novaya gramotnost: chemu uchit segodnya dlya uspekha zavtra. Predvaritelnye vyvody mezhdunarodnogo doklada o tendentsiyakh transformatsii shkolnogo obrazovaniya* [Key Competences and New Literacy: From Slogans to School Reality. Preliminary Results of the International Report of Major Trends in the On-Going Transformation of School Education. A Summary for Discussion]. Moscow: HSE. Available at: https://ioe.hse.ru/data/2018/07/12/1151646087/2_19.pdf (accessed 2 April 2021).

Gelman R., Butterworth B. (2005) Number and Language: How Are They Related? *Trends in Cognitive Sciences*, vol. 9, no 1, pp. 6–10. doi: https://doi.org/10.1016/j.tics.2004.11.004.

Gignac G.E. (2008) Higher-Order Models versus Direct Hierarchical Models: g as Superordinate or Breadth Factor? *Psychology Science Quarterly*, vol. 50, no 1, pp. 21–43. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.531.4178&rep=rep1&type=pdf (accessed 2 April 2021).

Grimm K.J. (2008) Longitudinal Associations Between Reading and Mathematics Achievement. *Developmental Neuropsychology*, vol. 33, no 3, pp. 410–426. doi: https://doi.org/10.1080/87565640801982486.

Haberman S.J. (2005) *When Can Subscores Have Value? ETS RR-05–08*. Princeton, NJ: ETS. doi: https://doi.org/10.1002/j.2333–8504.2005.tb01985.x.

Haberman S.J., Sinharay S. (2010) Reporting of Subscores Using Multidimensional Item Response Theory. *Psychometrika*, vol. 75, no 2, pp. 209–227. doi: https://doi.org/10.1007/s11336–010–9158–4.

Harris K.M., Martin M.O. (eds) (2013) *TIMSS2015 Assessment Framework*. Available at: https://timssandpirls.bc.edu/timss2015/downloads/T15_Frameworks_Full_Book.pdf (accessed 2 April 2021).

Jablonka E. (2003) Mathematical Literacy. *Second International Handbook of Mathematics Education* (eds A.J. Bishop, M.A. Clements, C. Keitel, J. Kilpatrick, F.K.S. Leung), Dordrecht, Netherlands: Kluwer Academic, pp. 75–102. doi: https://doi.org/10.5951/mtlt.2019.0397.

Jordan N.C., Hanich L.B., Kaplan D. (2003) A Longitudinal Study of Mathematical Competencies in Children with Specific Mathematics Difficulties versus Children with Comorbid Mathematics and Reading Difficulties. *Child Development*, vol. 74, no 3, pp. 834–850. doi: https://doi.org/10.1111/1467–8624.00571.

Joyner R.E., Wagner R.K. (2020) Co-Occurrence of Reading Disabilities and Math Disabilities: A Meta-Analysis. *Scientific Studies of Reading: The Official Journal of*

the Society for the Scientific Study of Reading, vol. 24, no 1, pp. 14–22. doi: https://doi.org/10.1080/10888438.2019.1593420.

Kamata A. (2001) Item Analysis by the Hierarchical Generalized Linear Model. *Journal of Educational Measurement*, vol. 38, no 1, pp. 79–93. doi: https://doi.org/10.1111/j.1745–3984.2001.tb01117.x.

Kanonire T., Federiakin D. A., Uglanova I. L. (2020) Multicomponent Framework for Students' Subjective Well-Being in Elementary School. *School Psychology Quarterly*, vol. 35, no 5, pp. 321–331. doi: https://doi.org/10.1037/spq0000397.

Kolkman M. E., Kroesbergen E. H., Leseman P. P.M. (2013) Early Numerical Development and the Role of Non-Symbolic and Symbolic Skills. *Learning and Instruction*, vol. 25, June, pp. 95–103. doi: https://doi.org/10.1016/j.learninstruc.2012.12.001.

Kozlov V. V., Kondakov A. M. (2009) *Fundamentalnoe yadro soderzhaniya obshchego obrazovaniya* [The Fundamental Core of the Content of General Education]. Moscow: Prosveshchenie. Available at: https://kpfu.ru/docs/F1999935214/fundamentalnoe.yadro.pdf (accessed 2 April 2021).

Libertus M. E., Odic D., Feigenson L., Halberda J. (2016) The Precision of Mapping between Number Words and the Approximate Number System Predicts Children's Formal Math Abilities. *Journal of Experimental Child Psychology*, vol. 150, October, pp. 207–226. doi: https://doi.org/10.1016/j.jecp.2016.06.003.

Linacre J. M. (1998) Structure in Rasch Residuals: Why Principal Components Analysis. *Rasch Measurement Transactions*, vol. 12, no 2, pp. 636. Available at: https://www.rasch.org/rmt/rmt122m.htm (accessed 2 April 2021).

Linacre J. M. (2002) What Do Infit and Outfit, Mean-Square and Standardized Mean. *Rasch Measurement Transactions*, vol. 16, no 2, pp. 878. Available at: https://www.rasch.org/rmt/rmt162f.htm (accessed 2 April 2021).

Linacre J. M. (2021) A User's Guide to Winsteps and Ministep: Rasch-Model Computer Programs. Program Manual 4.8.0. Available at: https://www.winsteps.com/manuals.htm (accessed 2 April 2021).

Linden W. J. van der (ed.) (2018) *Handbook of Item Response Theory*. Boca Raton, FL: CRC.

Mansolf M., Reise S. P. (2017) When and Why the Second-Order and Bifactor Models Are Distinguishable. *Intelligence*, vol. 61, February, pp. 120–129. doi: https://doi.org/10.1016/j.intell.2017.01.012.

Newman M. A. (1977) An Analysis of Sixth-Grade Pupils' Errors on Written Mathematical Tasks. *Research in Mathematics Education in Australia* (eds M. A. Clements, J. Foyster), Melbourne: Swinburne College, vol. 1, pp. 239–258.

OECD (2013) *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. Paris: OECD. doi: https://doi.org/10.1787/9789264190511-en.

OECD (2019) *PISA 2018 Assessment and Analytical Framework*. Paris: OECD. doi: https://doi.org/10.1017/CBO9781107415324.004.

Paek I., Yon H., Wilson M., Kang T. (2009) Random Parameter Structure and the Testlet Model: Extension of the Rasch Testlet Model. *Journal of Applied Measurement*, vol. 10, no 4, pp. 394–407. Available at:: https://bearcenter.berkeley.edu/sites/default/files/Wilson%20%2327.pdf (accessed 2 April 2021).

Peng P., Namkung J., Barnes M., Sun C. (2016) A Meta-Analysis of Mathematics and Working Memory: Moderating Effects of Working Memory Domain, Type of Mathematics Skill, and Sample Characteristics. *Journal of Educational Psychology*, vol. 108, no 4, pp. 455–473. doi: https://doi.org/10.1037/edu0000079.

Purpura D. J., Baroody A. J., Lonigan C. J. (2013) The Transition from Informal to Formal Mathematical Knowledge: Mediation by Numeral Knowledge. *Journal of Educational Psychology*, vol. 105, no 2, pp. 453–464. doi: https://doi.org/10.1037/a0031753.

Rasch G. (1993) *Probabilistic Models for Some Intelligence and Attainment Tes*ts. Chicago, IL: MESA.

Reckase M. D. (2009) *Multidimensional Item Response Theory*. NY: Springer-Verlag.

Reise S. P. (2012) The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research*, vol. 47, no 5, pp. 667–696. doi: https://doi.org/10.1080/00273171.2012.715555.

Revelle W. (2020) *Package 'Psych'— Version: 1.9.12.31.* Available at: https://cran.r-project.org/web/packages/psych/index.html (accessed 2 April 2021).

Rijmen F. (2010) Formal Relations and an Empirical Comparison among the Bi-Factor, the Testlet, and a Second-Order Multidimensional IRT Model. *Journal of Educational Measurement*, vol. 47, no 3, pp. 361–372. doi: https://doi.org/10.1111/j.1745–3984.2010.00118.x.

Robitzsch A., Kiefer T., Wu M. (2020) *Package 'TAM'. Test Analysis Modules — Version: 3.5–19.* Available at: https://cran.r-project.org/web/packages/TAM/index.html (accessed 2 April 2021).

Schmid J., Leiman J. M. (1957) The Development of Hierarchical Factor Solutions. *Psychometrika*, vol. 22, no 1, pp. 53–61. doi: https://doi.org/10.1007/BF02289209.

Schneider M., Beeres K., Coban L., Merz S., Schmidt S. S., Stricker J., de Smedt B. (2017) Associations of Non-Symbolic and Symbolic Numerical Magnitude Processing with Mathematical Competence: A Meta-Analysis. *Developmental Science*, vol. 20, May, no e12372. doi: https://doi.org/10.1111/desc.12372.

Schwarz G. (1978) Estimating the Dimension of a Model. *The Annals of Statistics*, vol. 6, no 2, pp. 461–464. doi: http://dx.doi.org/10.1214/aos/1176344136.

Sinharay S., Puhan G., Haberman S. J. (2011) An NCME Instructional Module on Subscores. *Educational Measurement: Issues and Practice*, vol. 30, no 3, pp. 29–40. doi: https://doi.org/10.1111/j.1745–3992.2011.00208.x.

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*, vol. 3, no 2, pp. 205–231.

Toll S. W.M., van Viersen S., Kroesbergen E. H., van Luit J. E.H. (2015) The Development of (Non-)Symbolic Comparison Skills throughout Kindergarten and Their Relations with Basic Mathematical Skills. *Learning and Individual Differences*, vol. 38, February, pp. 10–17. doi: https://doi.org/10.1016/j.lindif.2014.12.006.

Wang W., Chen P., Cheng Y. (2004) Improving Measurement Precision of Test Batteries Using Multidimensional Item Response Models. *Psychological Methods*, vol. 9, no 1, pp. 116–136. doi: https://doi.org/10.1037/1082–989X.9.1.116.

Warm T. A. (1989) Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika*, vol. 54, no 3, pp. 427–450. doi: https://doi.org/10.1007/BF02294627.

Wilson M. R. (2005) *Constructing Measures: An Item Response Modeling Approach*. New Jersey: Routledge.

Wilson M., Gochyyev P. (2020) Having Your Cake and Eating It Too: Multiple Dimensions and a Composite. *Measurement*, vol. 151, November, no 107247. doi: https://doi.org/10.1016/j.measurement.2019.107247.

Wright B. D., Masters G. N. (1990) Computation of OUTFIT and INFIT Statistics. *Rasch Measurement Transaction*, vol. 3, no 4, pp. 84–85. Available at: https://www.rasch.org/rmt/rmt34e.htm (accessed 2 April 2021).

Wright B. D., Stone M. H. (1979) *Best Test Design*. Chicago, IL: MESA.

Wu M. (2010) *Comparing the Similarities and Differences of PISA 2003 and TIMSS. OECD Education Working Papers no 32*. doi: https://doi.org/10.1787/5km4psnm13nx-en.

Wu M., Tam H. P., Jen T. H. (2016) *Multidimensional IRT Models in Book: Educational Measurement for Applied Researchers. Theory into Practice*. Singapore: Springer.