Psychometrical Modeling of Components of Composite Constructs: Recycling Data Can Be Useful[1]

Denis Federiakin[1[0000-0003-0993-5315]] and Elena Kardanova[1[0000-0003-2280-1258]]

[1] National Research University Higher School of Economics, Potapovsky Lane 16, build. 10,
101000 Moscow, Russia
dafederiakin@hse.ru, ekardanova@hse.ru

Abstract. This paper describes a list of studies necessary to justify the simultaneous use of both the overall test score and the subscale scores when measuring complex constructs. We investigate in detail one of the strategies for modeling composite constructs, which is popular within the international comparative studies of education. This strategy is based on repetitive recalibrations of the same data using unidimensional models for reporting overall test score and multidimensional models for reporting its components. We use Monte-Carlo simulations to illustrate that repetitive recalibrations of the data using unidimensional and multidimensional models yield, basically, the same results after their transformation to the same scales. However, we also illustrate that the fit of the unidimensional models to the data may be confounded if the components of the composite vary in terms of their relations with each other and their variance. We illustrate the studied strategy for modeling composite constructs using the computer adaptive test PROGRESS-ML, which measures basic math literacy in the third grade.

Keywords: Composite Constructs, Composite Tests, Multidimensional Rasch Models, Unidimensional Rasch Models, PROGRESS-ML.

1 Introduction

Within contemporary educational sciences and broadly, in the social sciences, there is a growing need for composite measurement instruments – instruments that have a complex structure, for example, those which consist of subscales that invest in some way in the overall test score. This may be a consequence of the trend for measuring complex constructs - such as 21st century skills or new literacies. Such constructs consist of multiple components, and it is not easy to portrait them as a classic unidimensional or single-component trait of respondents. It is widely assumed that the information about the integral trait level is valuable for policymakers, while information about its components is valuable for practitioners. Such information provides important insights for improving the performance of, for example, the educational system or psychological practice at different levels of the social system.

[1] Supported by the Russian Science Foundation, grant No. 19-29-14110.

Proceedings of the 4th International Conference on Informatization of Education and E-learning Methodology: Digital Technologies in Education (IEELM-DTE 2020), Krasnoyarsk, Russia, October 6-9, 2020.

The standards for educational and psychological testing [1] clearly state that (1) test scores should not be reported to users until their validity, fairness, and reliability have been studied, and (2) if the test produces more than one test score, the psychometric quality of all reported scores must be confirmed.

This is important because inaccurate information about the overall test score can lead to decisions with undesirable social consequences, while erroneous information about subscores can lead to incorrect decisions to correct or improve the situation [2]. In an academic environment, low-quality subscores can lead to false conclusions about the nature of the phenomenon being studied.

2 Psychometrics of composite instruments

In psychometric terms, composite tests are multidimensional. Therefore, the task is to evaluate, if possible, both the overall ability and its components. Psychometric modeling of such tests consists of several stages. First of all, a researcher needs to check whether the test is essentially unidimensional. It is possible to do so by utilizing the weak definition of local item independence stating that item residual correlations are zero after extracting a single factor estimated by the unidimensional model (figure 1a) [3]. If so, a researcher can report the overall test score - of course, given that it is proven to be valid and psychometrically consistent. If the test is not unidimensional, it is necessary to use multidimensional models, and then the overall test score requires additional research using hierarchical models [4]. Two types of hierarchical models are particularly popular – models with higher-order factors (figure 1c) [5] and bifactor models [6, 7] (figure 1d). Despite the algebraic similarities [8, 9] and the fact that both groups of models assume the use of the overall test score (called the general factor in factor-analysis terminology), their interpretation is different [10, 11]. While models with higher-order factors estimate the general factor that manifests in items through subscores, bifactor models assume a complete separation of the general factor and specific factors.

Second, if a researcher intends to report subscores (for example, cognitive operations or content areas), several approaches are available. The first is to apply the unidimensional model to each subscale separately [12]. This approach is called the "Consecutive Approach". The consecutive approach is the least attractive since the number of items in each subscale is usually small. Therefore, the measurement reliability will not be high enough, and the measurement error will be too large. This leads to the impossibility of reporting subscores [1].

The second approach involves the use of bifactor models. These models, hypothetically, allow simultaneous reports of the overall score and subscores as additional independent information. However, studies show that subscores estimated in bifactor models rarely have satisfactory reliability because they describe information not extracted by the overall score. Therefore, valuable information is often suppressed by random noise [13]. Moreover, their interpretation is difficult due to model assumptions.

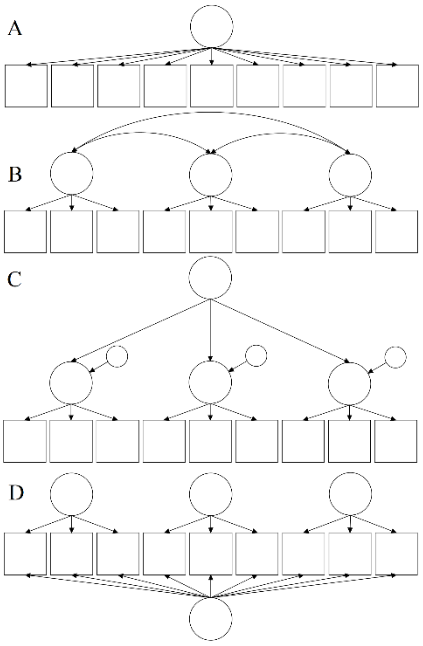

Fig. 1. Structural models for modeling composite constructs. Latent variables are drawn using circles, while observed variables are drawn using squares. One-headed arrows represent regression dependencies, while two-headed arrows represent correlations

The third approach involves the use of non-compensatory multidimensional models [14] (correlated traits models or models for between-item multidimensionality [15], figure 1b). Such models represent, essentially, several unidimensional models combined in a single likelihood equation. This approach is under investigation in this paper. From a modeling perspective, it is crucial to distinguish this analysis strategy from the bifactor modeling and consecutive approach. The described approach breaks the general factor into its parts, proportional to the number of items dedicated to a particular dimension. Each latent trait is calculated based on respondents' responses to the corresponding items and considering the latent variables' estimated correlations. Thus, multidimensional models use information about each dimension and compute the probability of completing or endorsing an item as a function of several latent variables, taking into account the relationships between them. As a result, such measurements' reliability will be greater compared to the consecutive approach. Therefore, it is more likely that it will be possible to report subscores. At the same time, bifactor modeling suggests modeling additional subscale-specific components, which add up to the general factor to produce the observed item scores. Consequently, the interpretation of the subscale-specific scores from bifactor models is too convoluted for the most practical tasks. As a result, the application of the bifactor models is mostly limited to modeling testlet-based assessments and local item dependence conditional on person parameters.

The third analysis strategy illustrates the use of collateral information. Collateral information is any information about items, respondents, or their interaction, which, being introduced in the measurement model, does not change the parameters' interpretation. However, collateral information reduces the uncertainty in the estimates [16]. In this case, for each subscale, the responses to all other subscales (together with the correlation matrix of latent dimensions) are collateral information [17].

Thus, to use the results of composite tests, regardless of the chosen strategy of data analysis, it is necessary to conduct extensive psychometric research. It is necessary to decide whether the overall test score and subscores are reliable and psychometrically consistent enough to be reported to users.

2.1 Modeling components of the composite

Breaking the overall test score into its components is popular within cross-national comparative studies of education. For example, PISA [18] and TIMSS [19] use repetitive recalibrations of their testing data to decompose the overall test score into the components, which produce it. TIMSS uses its theoretical framework to report subscores on cognitive operations required to solve an item. From a statistical point of view, de facto, it leads to ignoring model-fit indices and recirculation of the data. Nevertheless, its interpretation allows researchers to describe the composition of the overall test scores in terms of how respondents achieve those test scores. This enables policymakers to make decisions based on the information described in terms of social sciences.

However, the difference and equivalence between multidimensional and unidimensional models is a challenging area of psychometric research. Many studies have already touched upon the idea of the unidimensional interpretation of multidimensional measurements. For example, Reckase et al. [20] showed that if the test items are selected according to specific conditions, the unidimensional model can fit such data. However, it requires strict guiding the process of test development by the psychometric parameters of the items. Several researchers have also tried to conceptualize the fit of the unidimensional models to multidimensional data in terms of the general factor's strength. For example, Drasgow and Parsons [21] demonstrated that if the general factor is "strong" (if the factors in the multidimensional model are firmly positively correlated), then the unidimensional model can fit the data well. Our paper describes the same phenomenon directly in terms of the correlation matrix of latent dimensions. Many other researchers studied how model modification can allow the unidimensional model to fit multidimensional data. The main implication of those findings is that it is possible to use the overall test score even if the general factor is weak as long as the multidimensional structure of the data is explicitly modeled [22].

Nonetheless, much research found that the differences between parameter estimates from multidimensional and unidimensional models are expressed, mainly, in item parameters. Numerous researches have highlighted unpredictable distortion in the item parameters estimated when the model's dimensionality is misspecified regarding the data-generating model [23]. However, another conclusion from this stream of research concerns the stability of the person parameters. As DeMar noted (although, in another context), "if the focus is on estimated θ's and not on the item parameters, any of the models will perform satisfactorily" [24]. Reise et al. [25] summarized that the correlation of person parameters from different models tends to be close to 1 regardless of the model's misspecification.

3 Simulation study

To illustrate the possibility of fitting the unidimensional models to the multidimensional data, we perform a small-scale Monte-Carlo simulation study. We generate the data under the multidimensional Rasch model and calibrate both unidimensional (misspecified) and multidimensional (correctly specified) Rasch models on the data. We then compare the average of the multiple person abilities from the multidimensional model and the estimated person ability from the unidimensional model. To compare the results, we used the Pearson linear correlation.

We also analyze the essential unidimensionality of the simulated data by utilizing residual analysis. To do so, we apply principal components analysis to the standardized response residuals under the unidimensional model. This is standard practice for the analysis of unidimensionality under the Rasch modeling paradigm. This method rests upon the assumption that if the data is unidimensional (and does not exhibit local item dependence conditional on person parameters), the residuals are noise, and any significant principal component cannot be extracted from the data [26; 27]. To analyze local fit, we used Rasch InFit and OutFit item-wise statistics [28], particularly their range from the maximum to minimum values. The larger range in InFit and OutFit means that some items deviate from the model prediction and do not fit the Rasch model, while smaller variance means that all items fit the Rasch model.

We conduct the simulations for 2000 respondents responding to 30 dichotomous items, separated into five subscales equally (6 items per subscale). We carry out 100 replications for randomly varying positive definite variance-covariance matrices with positive manifold (where all latent dimensions are non-negatively correlated). Note, however, that during random varying of the variance-covariance matrix, we also alter the variance of latent dimensions. To control this source of the difference of the results, we also carry out 50 replications for three fixed variance-covariance matrices of person parameters (where all correlations were equal to 0.80, 0.50, or 0.20, and all variances are equal). For the randomly varying variance-covariance matrices, we calculate the difference between correlations by taking the standard deviation of the values in the lower triangle of the correlation matrix. We do so to analyze the fit of the unidimensional models conditional on the difference between the variance-covariance matrix values. Both the multidimensional model and unidimensional model can be considered as special cases of the Multidimensional Random Coefficients Multinomial Logit Model [15]. The quasi-Monte-Carlo algorithm implemented in the Tam v. 3.5-19 package [29] for the R V. 3.6.2 software was used to estimate all models.

3.1 Results of the simulation study

The average correlation between person parameters from the unidimensional model and the average of person parameters from the multidimensional model is 0.99 ($p < 0.01$) with a standard deviation of less than 0.01 across all simulated conditions. These results hold for any case – whether the variance-covariance matrix was fixed or not. This result is in agreement with other similar research, suggesting that person parameters are more stable in the situation of model dimensionality misspecification.

Further, the results of dimensionality analysis using PCA on unidimensional model residuals do vary depending on the size of correlations of latent dimensions in the data-generating multidimensional model. They suggest that the eigenvalue of the first component depends on the mean correlation of those dimensions ($r = -0.48$, $p < 0.01$, figure 2) and less depends on differences in the values of correlation matrix ($r = -0.25$, $p < 0.05$, figure 3). Note, however, that the critical value for the first eigenvalue is 2 [30, 31]. Since the first component's eigenvalue is larger than the critical value, all unidimensional models are critically misspecified for the simulated data, and, therefore, their results are inconsistent.

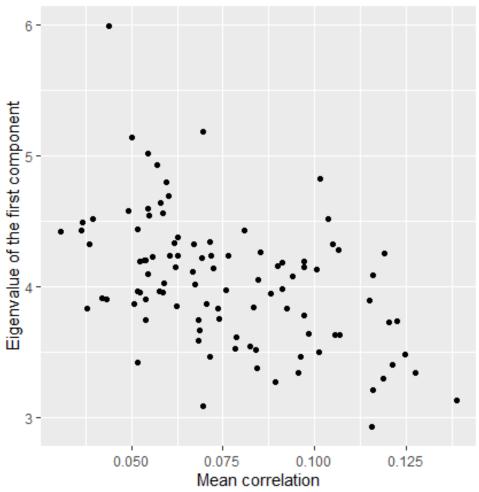

Fig. 2. Scatterplot of eigenvalues of the first component from PCA applied to the standardized model residuals versus mean correlation of latent dimensions from the simulations with randomly varying variance-covariance matrices. Each point represents a single simulation

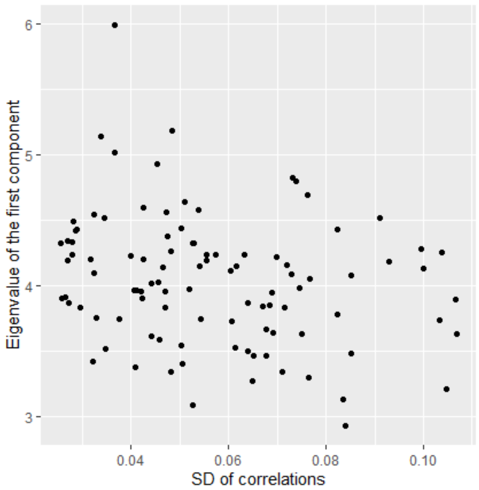

Fig. 3. Scatterplot of eigenvalues of the first component from PCA applied to the standardized model residuals versus standard deviation of correlations of latent dimensions from the simulations with randomly varying variance-covariance matrices. Each point represents a single simulation

To support these findings, we additionally analyzed the eigenvalue of the first component from PCA applied to the unidimensional model residuals when the variance-covariance matrix was fixed. We compared the eigenvalue of the first component across different values of the fixed correlation and the varied matrix. The results are presented in figure 4. They also suggest that the analyzed eigenvalue depends on the size of the fixed correlation. However, they never exceed the critical value of 2. Therefore, the data with small (or absent) variance in the values of the correlation matrix of underlying latent factors can be considered unidimensional.

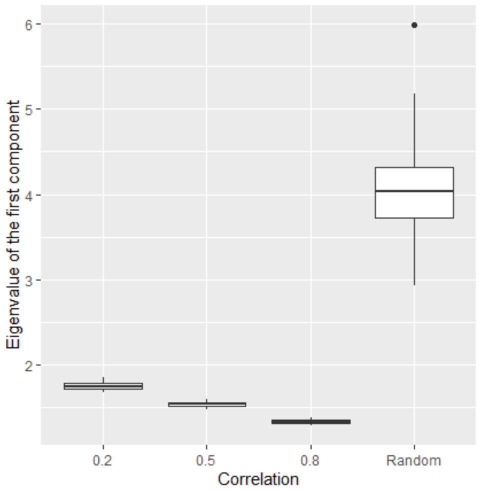

Fig. 4. Boxplot of the variance of the first component from PCA applied to the standardized model residuals depending on conditions for simulations

Next, we analyzed item fit statistics. The results are presented in figure 4. We compared values of item fit statistics across different conditions of simulations similarly with previous results. The results are presented in figures 5 and 6. We discovered similar findings: the range of item fit statistics from unidimensional models in case of randomly varied variance-covariance matrix exceeds that of the fixed variance-covariance matrix. However, since the model used for data-generating is the Rasch model as well as the model used for data analysis, item fit statistics do not react to differences in item discrimination parameters. Instead, they react to the violation of unidimensionality, which is expected [27].

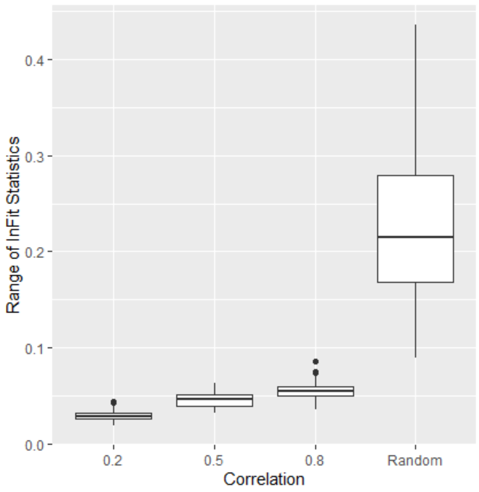

Fig. 5. Range of Rasch InFit item-wise statistic depending on conditions for simulations

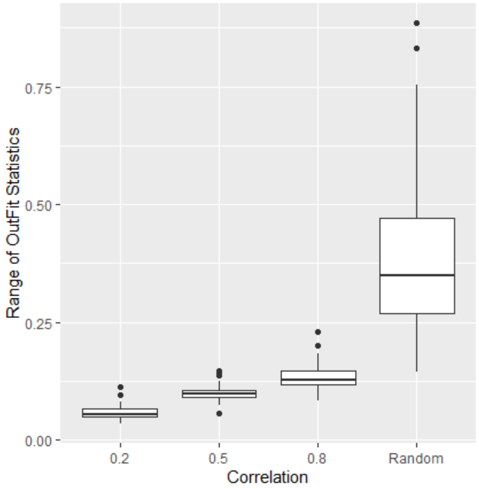

Fig. 6. Range of Rasch OutFit item-wise statistic depending on conditions for simulations

Thus, we showed that the unidimensional IRT model could fit the data well even if the data was actually generated under the multidimensional model. This is fair for the cases where the values in the correlation matrix of latent dimensions are positive, correlations are strong, and they do not vary much. However, regardless of that, the average of the ability estimates from the multidimensional model is equal to the ability estimate from the unidimensional model. For this, of course, their transformation to the scales with the same numerical values is necessary (e.g., linear transformation to the scale N(500,100)). This finding is in agreement with previous studies, which found that the person parameters are not as sensitive to the model dimensionality misspecification as the item parameters.

Nonetheless, psychometric consistency of the unidimensional score can be confounded if there is variation in the correlation matrix of "true" latent dimensions. If this is a case, the extraction of the overall test score from multidimensional data cannot be conducted by averaging the multidimensional model's estimates. Additional research on "sufficient unidimensionality" of the data is crucial for overall test score reporting.

4 Real data example

This section demonstrates the scope of psychometric studies necessary for reporting both overall test scores and specific scores, interpreting them as components of the composite construct. We do so by applying them for the PROGRESS-ML basic mathematical literacy test.

The PROGRESS-ML test evaluates how well a student is oriented in mathematics after completing two years of primary school. When developing the test, we relied on the following definition of basic mathematical literacy [32]: "basic mathematical literacy (including working with data) – the ability to apply mathematical tools, reasoning, and modeling in everyday life, including in the digital environment".

The PROGRESS-ML basic math literacy test consists of 30 dichotomous items. The assessment is built as a computerized adaptive test with an automated stopping rule.

The content of the test was selected in a way that, on the one hand, it meets the definition of basic mathematical literacy, and on the other hand, it takes into account the content of the Russian Federal Educational Standard. As a result, we identified five content areas: spatial representations, measurement of quantities, regularities, modeling, and information processing. Test items are grouped into blocks according to the content area.

Additionally, the PROGRESS-ML test evaluates students' cognitive processes required to solve the items. When developing the test items, we used the TIMSS' theoretical framework for the 4th grade [33]. Therefore, in addition to assessing the content area, three cognitive operations groups are measured — knowing, application, and reasoning.

Thus, the PROGRESS-ML test is a composite tool: it includes five content areas and reflects three cognitive operations groups. It is assumed that the test results will report the students' overall test score (in this case, the level of their basic mathematical literacy), as well as subscores (in this case, content areas and cognitive operations).

The sample consisted of 6078 the 3rd grade students from two regions of the Russian Federation. The samples were representative for the regions. Average age = 9.06 years (SD = 0.46), number of girls = 52.36%.

4.1 Results of the analysis of the real data

In the analysis of standardized residuals by the PCA, we found that the first component's eigenvalue is 1.45, which corresponds to 4.2% of the residual variance. The next four components' eigenvalues are in the interval from 1.15 to 1.2. The distribution of the explained variance of residuals among the components is almost uniform – about 4% per component. Therefore, we conclude that the unidimensional model sufficiently describes the response probability distribution across persons, and the test can be considered unidimensional.

The model Expected-a-Posteriori reliability [34] of the entire test score from the unidimensional model was 0.76. For comparison, we calculated the reliability using the methods of Classical Test Theory (CTT): Greatest Lower Bound (GLB) [35] reliability was 0.86, the Cronbach's α [36] was 0.81. However, it is essential to note that the design of testing (computerized adaptive) implies that not all items are administrated to all respondents, and the CTT parameters become unstable in the presence of missing responses. Therefore, even though, in our example, the reliability estimated in the CTT (both GLB and Cronbach's α) is slightly higher than the reliability of the scores evaluated in the IRT, these indices should not be trusted.

Overall, the analysis results suggest that the test can be considered unidimensional, even though there are different ways to group items. This implies that it is possible to report one overall test score of mathematical literacy based on the test results, which will have good reliability and psychometric consistency.

Then, we calibrated the multidimensional IRT model to estimate if they will have good psychometric characteristics. The reliability analysis results by content areas are shown in table 1, by cognitive operations are shown in table 2.

Table 1. Analysis of relations between content areas

Content area	Spatial representation	Measurement of quantities	Regularities	Modeling	Information processing
Spatial representation		0.85	0.80	0.83	0.80
Measurements			0.85	0.90	0.83
Regularities				0.86	0.84
Modeling					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

Table 2. Analysis of relations between cognitive operations

Cognitive operation	Knowing	Application	Reasoning
Knowing		0.95	0.85
Application			0.85
Variance	1.37	0.82	0.60
Reliability	0.75	0.74	0.61
Number of items	12	14	4

From the tables, we can conclude that all dimensions have sufficient reliability for the monitoring test use. Despite the small number of items per subscale, relatively high reliability is possible due to the approach used for IRT modeling. In fact, such a small number of items per dimension makes raw subtest scores unusable. Additionally, we looked at correlations between the latent dimensions: both content areas and cognitive operations correlate approximately equally – at the level of 0.8-0.9. Based on the simulation study, we conclude that this can be seen as an additional argument in favor of the unidimensional model, even though multidimensional models fit the data better than the unidimensional model according to the AIC [37] and BIC [38] indices. These indices can estimate the relative model fit to the data introducing a penalty for extra model parameters (AIC) with respect to sample size (BIC). The lower values of these indices indicate a better model fit. These indices are presented in table 3.

Table 3. Analysis of model fit

Model	Deviance	Sample	Number of parameters	AIC	BIC
Unidimensional	144255.6		31	144318	144526
Content areas	143875.7	6078	45	143966	144268
Cognitive operations	143965.4		36	144037	144279

Thus, multidimensional IRT models allowed us to get reasonably reliable subscores (for both content areas and cognitive operations) and therefore made it possible to report them to users. Moreover, the described reliability estimates are derived from IRT models in which no context variables were entered. Note that the introduction of these variables into the model (using latent regression modeling) leads to the estimation of more reliable scores for subscales due to explaining ability variance.

5 Conclusion

Contemporary psychometric literature notes the growing popularity of composite tests designed to produce both the overall test score and the subscores. There are several strategies for processing such test data. They include the use of raw test scores or the application of hierarchical models. However, in most cases, raw test scores cannot be used due to their low reliability [13], and hierarchical models require extraordinary caution in use due to their complex mathematical nature and interpretation.

In this paper, we describe the strategy for modeling subscores as components of the overall composite test score. This strategy is based upon repetitive recalibration of the same data using unidimensional and multidimensional models. We demonstrate that the average of the ability estimates from the multidimensional models is equal to the ability estimate from the unidimensional model estimated on the same data. Interestingly, this statement holds regardless of whether or not the unidimensional model fits the data. However, the application of any statistical models in social sciences needs to be backed by checking its assumptions and thinking through its theoretical consequences. Therefore, the unidimensional model's meaningfulness must be argued in terms of both model fit and construct definition. As we demonstrate in our simulation study, the unidimensional model does not always fit the data despite the equivalence of its estimates to the average of the estimates from the multidimensional models. This means that the unidimensional model's adequacy needs to be verified either way if a researcher intends to follow the described approach in modeling the composite constructs.

We also provide an example of the described strategy for modeling the composite constructs using the PROGRESS-ML basic mathematical literacy test. We demonstrate that the use of IRT models allows us to report the respondent's overall test score and subscores with respect to test specification. For this test, the main result of testing is the respondent's overall test score. However, repeated recalibration of data based on content areas and cognitive operations groups required for solving items allows us to report subscores on those dimensions. These estimates possess greater reliability and simpler interpretation than estimates from other approaches to modeling composite constructs. The essence of these results is the decomposition of the overall test score into the components that make it up.

References

1. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education: Standards for educational and psychological testing. American Educational Research Association, Washington, DC (2014).
2. Sinharay, S., Puhan, G., & Haberman, S.J. An NCME instructional module on subscores. Educational Measurement: Issues and Practice, 30(3), 29–40 (2011).
3. Hattie, J.: Methodology review: Assessing unidimensionality of tests and items. Applied Psychological Measurement, 9, 139–164 (1985).
4. Yung, Y.F., Thissen, D., & McLeod, L.D.: On the relationship between the higher-order factor model and the hierarchical factor model. Psychometrika, 64(2), 113-128 (1999).

5. Gignac, G.E.: Higher-order models versus direct hierarchical models: g as superordinate or breadth factor? Psychology Science, 50(1), 21 (2008).
6. Holzinger, K.J., & Swineford, F. A study in factor analysis: The stability of a bi-factor solution. Supplementary educational monographs (1939).
7. Reise, S.P. The rediscovery of bifactor measurement models. Multivariate behavioral research, 47(5), 667-696, (2012).
8. Schmid, J., & Leiman, J.M.: The development of hierarchical factor solutions. Psychometrika, 22(1), 53-61 (1957).
9. Rijmen, F. Formal relations and an empirical comparison among the bi‐ factor, the testlet, and a second‐ order multidimensional IRT model. Journal of Educational Measurement, 47(3), 361-372, (2010).
10. Brunner, M., Nagy, G., & Wilhelm, O.: A tutorial on hierarchically structured constructs. Journal of personality, 80(4), 796-846 (2012).
11. Mansolf, M., & Reise, S.P.: When and why the second-order and bifactor models are distinguishable. Intelligence, 61, 120-129 (2017).
12. Davey, T. Hirsch, T.M.: Concurrent and Consecutive estimates of examinee ability profiles. Paper presented at the Annual Meeting of the Psychometric Society, New Brunswick, NJ. (1991).
13. Haberman, S.J., & Sinharay, S.: Reporting of subscores using multidimensional item response theory. Psychometrika, 75(2), 209-227 (2010).
14. Reckase, M.D. Multidimensional item response theory models. In Multidimensional item response theory (pp. 79-112). Springer, New York, NY (2009).
15. Adams, R.J., Wilson, M., & Wang, W.C.: The multidimensional random coefficients multinomial logit model. Applied psychological measurement, 21(1), 1-23 (1997).
16. Wang, W., Chen, P., & Cheng, Y.: Improving measurement precision of test batteries using multidimensional item response models. Psychological Methods, 9(1), 116-136, (2004).
17. Wu, M., Tam, H.P., & Jen, T.H. (2016). Multidimensional IRT Models in Book: Educational measurement for applied researchers. Theory into practice.
18. Scaling PISA Data (Chapter 9). In: PISA 2018 Technical Report. OECD, Paris (2019).
19. Foy, P., & Yin, L. Scaling the TIMSS 2015 Achievement Data. In: Martin, M.O., Mullis, I.V., & Hooper, M. (eds.) Methods and procedures in TIMSS 2015, pp. 13.1-13.62. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA (2016).
20. Reckase, M.D., Ackerman, T.A., & Carlson, J.E.: Building a unidimensional test using multidimensional items. Journal of Educational Measurement 25(3), 193-203 (1988).
21. Drasgow, F., &Parsons, C.: Application of unidimensional item response theory models to multidimensional data. Applied Psychological Measurement, 7, 189–199 (1983).
22. Ip, E.H.: Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. British Journal of Mathematical and Statistical Psychology, 63(2), 395-416 (2010).
23. Steinberg, L., & Thissen, D.: Uses of item response theory and the testlet concept in the measurement of psychopathology. Psychological Methods, 1, 81–97 (1996).
24. DeMars, C.E. Application of the bi-factor multidimensional item response theory model to testlet-based tests. Journal of Educational Measurement, 43, 145–168 (2006).
25. Reise, S.P., Cook, K.F., & Moore, T.M. Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In Reise, S.P., & Revicki, D.A. (eds.) Handbook of item response theory modeling. Routledge, New York (2014).
26. Linacre, J.M.: Structure in Rasch residuals: why principal components analysis. Rasch measurement transactions, 12(2), 636 (1998).

27. Smith, E.V.: Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. Journal of Applied Measurement, 3, 205–231 (2002).
28. Linacre, J.M.: What do infit and outfit, mean-square and standardized mean. Rasch Measurement Transactions, 16(2), 878, (2002).
29. Robitzsch, A., Kiefer, T., Wu, M. Package 'TAM'. Test Analysis Modules – Version: 3.5-19, (2020).
30. Raîche, G.: Critical eigenvalue sizes in standardized residual principal components analysis. Rasch measurement transactions, 19(1), 1012 (2005).
31. Linacre, J.M.: Winsteps® Rasch measurement computer program User's Guide. Beaverton, OR: Winsteps.com (2018).
32. Фрумин, И.Д., Добрякова, М.С., Баранников, К.А., & Реморенко, И.М. Универсальные компетентности и новая грамотность: чему учить сегодня для успеха завтра. Предварительные выводы международного доклада о тенденциях трансформации школьного образования (2(19); Современная Аналитика Образования) (2018).
33. Mullis, I.V., & Martin, M.O.: TIMSS 2019 Assessment Frameworks. International Association for the Evaluation of Educational Achievement, Amsterdam, The Netherlands (2017).
34. Bock, R.D., & Mislevy, R.J: Adaptive EAP estimation of ability in a microcomputer environment. Applied psychological measurement, 6(4), 431-444 (1982).
35. Jackson, P.H., & Agunwamba, C.C.: Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: Algebraic lower bounds. Psychometrika, 42(4), 567-578 (1977).
36. Cronbach, L. J.: Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297-334 (1951).
37. Akaike, H.: A new look at the statistical model identification. IEEE transactions on automatic control, 19(6), 716-723 (1974).
38. Schwarz, G.: Estimating the dimension of a model. The annals of statistics, 6(2), 461-464 (1978).