Journal of Language Testing & Assessment (2022) Vol. 5: 13-21 **DOI: 10.23977/langta.2022.050102**
Clausius Scientific Press, Canada **ISSN 2616-2938**

Exploring Item Bank Stability through Live and Simulated Datasets

Tony Lee, David Coniam and Michael Milanovic

LanguageCert, UK

Keywords: item banks, stability, simulated dataset, Rasch, Bayesian ANOVA

Abstract: LanguageCert manages the construction of its tests, exams and assessments using a sophisticated item banking system which contains large amounts of test material that is described, inter alia, in terms of content characteristics such as macroskills, grammatical and lexical features and measurement characteristics such as Rasch difficulty estimates and fit statistics. In order to produce content and difficulty equivalent test forms, it is vital that the items in any LanguageCert bank manifest stable measurement characteristics.

The current paper is one of two linked studies exploring the stability of one of the item banks developed by LanguageCert [Note 1]. This particular bank has been used as an adaptive test bank and comprises 820 calibrated items. It has been administered to over 13,000 test takers, each of whom have taken approximately 60 items. The purpose of these two exploratory studies is to examine the stability of this adaptive test item bank from both statistical and operational perspectives.

The study compares test taker performance in the live dataset with over 13,000 test takers (where each test taker takes approximately 60 items) with a simulated 'full' dataset generated using model-based imputation. Simulation regression lines showed a good match and Rasch fit statistics were also good: thus indicating that items comprising the adaptive item bank are of high quality both in terms of content and statistical stability. Potential future stability was confirmed by results obtained from a Bayesian ANOVA. As mentioned above, such item bank stability is important when item banks are used for multiple purposes, in this context for adaptive testing and the construction of linear tests. The current study therefore lays the ground work for a follow-up study where the utility of this adaptive test item bank is verified by the construction, administration and analysis of a number of linear tests.

Introduction

This paper reports on a study investigating the stability and robustness of one of the item banks developed by LanguageCert. Given that both linear paper-based and adaptive high-stakes tests are produced from such item banks, key issues that need to be confirmed are item bank stability and item measurement quality in terms of tests generated from such item banks (see Mills & Steffen, 2000). These issues are important because test quality is a vital consideration for any organisation administering high-stakes examination.

Item Bank Size and Stability

Operationally, a key question is how to establish the stability of an item bank from a measurement perspective. In this context, we are working with an item bank containing 820 items used both as an adaptive test bank and for the generation of linear tests. 'Stability' may be defined here from two perspectives. The first is that model-fit statistics remain within acceptable ranges, even at the extreme ends of the percentile spectrum. The second, from an operational perspective, is that tests derived from the item bank produce comparable results when run with test takers.

One of the early researchers into item banking with particular reference to adaptive testing some five decades ago was Choppin (1968). Choppin's starting point was that an item bank of around 500 items was required, calibrated on 2,000 test takers. Ree (1981) conducted simulations of different adaptive test scenarios with differing test taker and item bank sample sizes. His recommendations, to an extent, echoed Choppin's findings: that an item bank comprising at least 200 items and calibrated on 2,000 test takers might be an acceptable starting point. Wainer (2000) describes an item pool consisting of some 800 items. Voss & Blumenthal (2020) describe a pool of 1,071 items calibrated on some 4,200 test takers.

Other researchers have nonetheless recommended rather larger item bank sizes. Derner et al. (2008) in discussing the construction of an item bank to measure technical skill attainment mentions 9,000 items as an optimal size, resources permitting. Similarly, Rudner (2009), in describing the development of the GMAT, states that for reasons such as security and broad construct coverage, the GMAT comprised over 9,000 items.

Among the limited number of researchers who have investigated stability (see e.g., Gao & Chen, 2005; Weiss & von Minden, 2012; Sahin & Weiss, 2015) studies have tended to focus on the theoretical construct in terms of how many (or rather how few) items might be necessary for information to be provided at appropriate θ levels (theta, for personal ability estimates) and with item parameters accurately estimated. While these studies provide an informative backdrop, the second study in this set differs somewhat in its approach to stability, in that following a simulated 'full dataset' study, an investigation into the direct construction and analysis of real world tests from a live item bank is conducted.

The LanguageCert adaptive item bank described in the current paper contains 820 items, and subsets of approximately 60 items have been administered (as adaptive tests) to approximately 13,000 test takers. This gives a live dataset of 0.78 million data points against a theoretical maximum of 10.66 million data points.

In assessment situations where items need to be calibrated to a common scale, analysis needs to take account of extensive amounts of missing data (Roth, 1998). This is particularly the case with the current item bank and adaptive test. As mentioned above, the adaptive bank contains many hundreds of items, with responses available for each test taker to only a small number of items. For the reliability of such an item bank to be demonstrated, item statistics therefore need to be computed such that missing values in the dataset are taken account of. This may be achieved by using imputed values (Peugh & Enders, 2004).

A number of methods for evaluating the effect of missing data have been explored: model-based imputation (Huisman & Molenaar, 2001); pairwise deletion (Zhang & Walker, 2008); maximum likelihood (Schminkey et al, 2016); multiple imputation (Li et al., 2015). The consensus would appear to converge on model-multiple imputation, and it is this method which has been adopted in the current study.

Multiple Imputation in the Current Study

The current study describes the analysis of this adaptive test item bank where missing data is simulated using the Rasch measurement program Winsteps (Linacre, 2018). Imputed missing data values have been generated via model-based multiple imputation, with the starting point for the simulation being, as mentioned, the 13,000 test takers, with their individual responses to 60 items.

Methodology

Figure 1 below presents a snapshot of data of actual test taker responses in the adaptive test dataset.

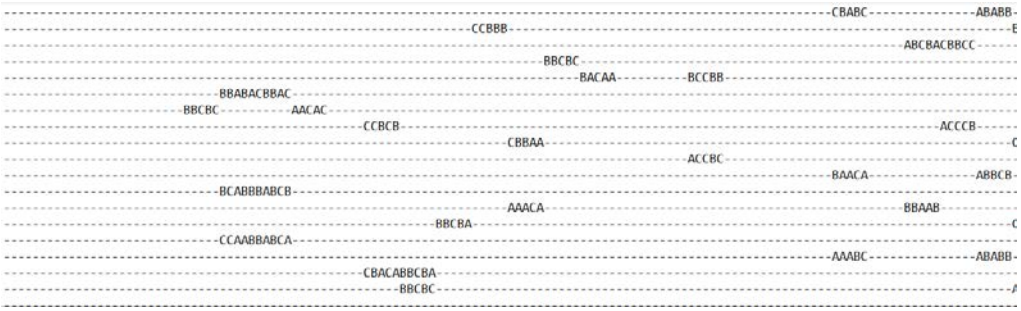

Figure 1. Data points in the live LTE adaptive test dataset

As may be seen from Figure 1, the data occurs as chunks, spread out across a vast data space.
The analysis in the current study has been conducted using the software Winsteps (Linacre, 2018), which simulates data using model-based multiple imputation. In the analysis, simulated data was imputed from the dataset presented in Figure 1 above. The 'full' dataset has been constructed – not on the basis of random responses – but via the software imputing a dataset for each test taker for all 820 items based on their limited set of actual responses. From the existing 0.78 million data points (as in Figure 1), the full dataset contains 10.66 million data points. Figure 2 presents a sample of the simulated dataset.

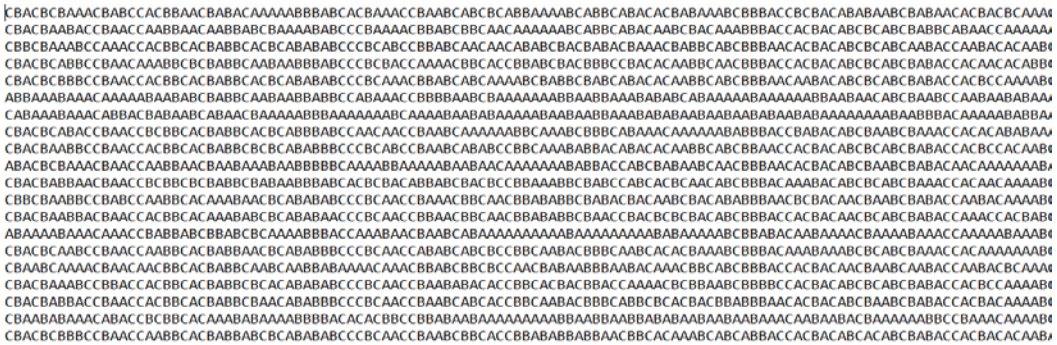

Figure 2. Simulated 'full' dataset

Hypotheses

The overarching hypothesis is that the simulation will return statistics within acceptable values, indicative of item bank stability. The study pursues the following hypotheses.
1. Regression line (R^2) values will be a minimum of 0.75 [the rule of thumb for 'substantial' R^2 values (Ringle & Sinkovics, 2009)].
2. Rasch infit and outfit statistics will be within acceptable ranges at the 25th and 75th percentiles: between 0.5 - 1.5.

3. The Bayes Factor will be in the range of 30-100 or higher, indicative of very strong evidence for the hypothesis of interest.

Rasch Measurement

The current study, as mentioned, is predicated on the use of the Rasch model, a brief overview of which is provided below.

The use of the Rasch model enables different facets (e.g., person ability and item difficulty) to be modelled together, converting raw data into measures which have a constant interval meaning (Wright, 1997) and which provide objective and linear measurement from ordered category responses (Linacre, 2012). This is not unlike measuring length using a ruler, with the units of measurement in Rasch analysis (referred to as 'logits') evenly spaced along the ruler. Once a common metric has been established for measuring different phenomena (test takers and test items, for example), person ability estimates may then be calculated independently from the items used, with item difficulty estimates also being calculated independently from the sample recruited.

In this manner, the Rasch model enables persons and items to be calibrated onto a single unidimensional latent trait scale – also known as the one-parameter IRT (Item Response Theory) model (Bond & Fox, 2007). Person measures and item difficulties are placed on an ordered trait continuum by which direct comparisons between person measures and item difficulties can be easily conducted, and results subsequently interpreted with a more general meaning.

Rasch Model Fit

Broad criteria in assessing model fit are the *Infit* and *Outfit* mean square statistics (i.e., estimates of population variance, or standard error) and the *Standardised Infit* and *Outfit* (i.e., Z-score) statistic. These statistics are outlined briefly below.

Infit may be seen as the 'big picture' in that it scrutinises the internal structure of an item or person. High infit mean square values indicate rather scattered information within the item or person, providing a confused picture about the placement of the item or person. Very small infit values indicate only very small variation and, provide therefore, little information to articulate clear and meaningful judgments about an item or person.

Outfit gives a picture of 'outliers', that is responses from persons or items that appear to be considerably out of line with where a person or item would expect to be placed.

For both Infit and Outfit, a perfect fit of 1.0 indicates that obtained values match expected values 100%. While acceptable ranges of tolerance for fit vary, acceptable ranges are generally taken as from 0.5 for the lower limit to 1.5 for the upper limit (Lunz & Stahl, 1990). 1.5 to 2.0 is considered just about acceptable, with figures beyond 2.0 unacceptable.

Bayesian Statistics

Bayesian statistical methods describe the conditional probability of an event based on data as well as prior information or beliefs about the event, with probabilities computed and updated after obtaining new data – see Andraszewicz et al. (2015).

Since Bayesian statistics treat probability as a degree of belief, permitting inferences about future events to be estimated in a positive way – other than simply of failure to reject the alternative hypothesis, as in standard statistical testing.

In Bayesian statistics, the critical statistic is the *Bayes Factor (BF)* – the ratio of likelihood between the null and the alternative hypothesis. Jeffreys (1961) proposes cutoff levels for interpreting the strength of Bayes Factors, recommending cutoff levels ranging from 1 (no evidence for the alternative hypothesis) to 10-30 (strong evidence), to 30-100 (very strong evidence), to > 100 (extreme evidence for the alternative hypothesis).

The *credible interval* is the Bayesian statistics version of the standard ("frequentist") statistics *confidence interval*. The credible interval represents the spectrum in which a specified percentage, e.g., 95%, of cases would fall. It has a direct interpretation as "the probability that ρ is in the specified interval" (Hoekstra et al., 2014).

Results

To test the hypotheses, the simulation was run to explore how a potentially-complete dataset compared with the live dataset which contained missing data values.
Table 1 presents the regression line calculated for the simulation.

Table 1. Simulation regression line

Sample size	13,000
R^2	0.99

As can be seen, the simulation shows an extremely good match with the regression line of 0.99 – markedly above the target of 0.75 (Ringle & Sinkovics, 2009). This result gives an indication of the stability of the calibration of the adaptive item bank under investigation, which is underpinned, it has to be assumed, by the quality of the items which constitute the bank.

Item-model Fit Statistics

Table 2 below presents a comparison of item-model fit statistics between the live ('sparse') and simulated ('full') dataset. Unacceptably high values are flagged in red font.

Table 2. Live and Simulated datasets (N=820): **Item fit statistics**

Percentile statistics	**Infit**			**Outfit**	
	Live dataset	Simulated dataset		Live dataset	Simulated dataset
Mean	1.03	1.00		1.07	1.00
Std. Deviation	0.34	0.01		0.45	0.03
Minimum (1st percentile)	0.54	0.98		0.98	0.93
25th percentile	0.93	1.00		0.89	0.99
50th percentile	0.97	1.00		0.95	1.00
75th percentile	1.05	1.00		1.10	1.01
Maximum (99th percentile)	4.76	1.02		5.18	1.37

As can be seen, at the 25th and 75th percentiles, fit statistics for values in the existing live dataset are well within the acceptable range of 0.5 – 1.5. It is only at maximum values that both infit and outfit mean squares emerge as being unacceptably high.
The simulated 'full' dataset presents a picture of stability – even at minimum and maximum percentile values (See Table 2). The larger standard deviations which emerge with the live dataset may be accounted for by the fact that each test taker in the live dataset has only 60 data points, as opposed to over 820 in the case for every item in the simulations.
The results using the simulated data suggest that the quality of the test items in the adaptive test item bank is high and that the adaptive test as it is currently calibrated would appear to be robust.

Person-model Fit Statistics

The explorations above have been at item level. To further explore the stability of the item bank, person-model fit statistics are now reported. Person values and misfit are a possibly greater encumbrance than item values – certainly when these are all calibrated – due to the fact that test takers may guess, leave blanks, cheat etc. (see Meijer, 1996).

The current study builds on research by Coniam et al. (2021), which documented different phases of measurement scale development for the LanguageCert Test of English (LTE), validating the LanguageCert Item Difficulty (LID) scale. Test taker results are reported against CEFR (the Common European Framework of Reference for Languages) levels, which have been defined on the basis of LanguageCert Item Difficulty (LID) scale scores; these are laid out in Table 3 below.

Table 3. LID scale

CEFR level	Mid point
C2	160
C1	140
B2	120
B1	100
A2	80
A1	60

The LID scale in Table 3 above is key for the interpretation of Table 4 below, which presents person-model fit statistics for the live and simulated datasets, with unacceptably high values again in red font. LID values are also included in the table in order to provide a more in depth picture of comparability.

Table 4. Live and simulated datasets (N=13,000): **Person fit statistics**

	Live dataset				**Simulated dataset**		
	LID values	Infit	Outfit		LID values	Infit	Outfit
Mean	120.80	1.00	1.03		121.12	1.00	1.00
SD	18.86	0.17	0.39		18.93	0.04	0.15
Minimum (1st percentile)	37.96	0.43	0.19		42.03	0.84	0.46
25th percentile	108.46	0.89	0.80		108.63	0.97	0.92
50th percentile	120.45	0.98	0.94		120.82	1.00	0.98
75th percentile	134.43	1.10	1.15		134.76	1.03	1.05
Maximum (99th percentile)	180.64	2.00	8.47		182.84	1.19	3.73

As can be seen, at the 25th, 50th and 75th percentiles, LID measures are constant with both datasets – indicative that the simulated dataset is a good extrapolation of the live dataset.

Fit statistics are within acceptable values. It is again only at maximum values that outfit mean squares in particular emerge as unacceptably high. This may well be due to outliers, i.e., test takers who have scored higher than they might have been expected to as a result of correct guesses. There is less misfit in infit and outfit mean square values in the simulated dataset than in the live dataset. This again suggests that – even though indications are that values computed from the current live dataset are stable and reliable – as the dataset increases in larger size, its stability will improve even further.

Bayesian Statistic Results

Bayesian statistics permit, as mentioned, the exploration of the probability-based future robustness of the adaptive test. To this end, a Bayesian ANOVA was run on the simulated dataset. The Bayesian H_0 for ANOVA (as with the null hypothesis in standard [frequentist] statistics), is that there will be no significant difference among test means.

The descriptives for the simulation are presented in Table 5.

Table 5. Simulation Descriptives (N=820)

		95% Credible Intervals	
Mean	SD	Lower	Upper
100.29	36.15	97.82	102.76

The 95% credible intervals indicate that the fluctuations of the item bank mean in future events would be less than three LID scale points above and below the mean with an extreme difference of about five LID scale points – approximately one quarter of a CEFR level.

Against the above backdrop, the overall estimation from the Bayesian ANOVA is provided in Table 6.

Table 6. Bayesian ANOVA estimations

Models	P(M)	P(M|data)	BF_M	BF_{01}	error %
Null model	0.5	1	14,830	1	
Simulations	0.5	0.00006	0.00006	14,830	0.0007

As mentioned, the critical statistic is the BF_{01} Bayes Factor. This represents the ratio of BF_0 (the null hypothesis of nil mean differences) to BF_1 (the alternative hypothesis of existence of mean difference). The target Bayes Factor was 30-100; the figure of 14,830 obtained is far beyond this figure, into the range of above 100: "extreme evidence" (after Jeffreys, 1961) in favour of the no difference in mean in the ANOVA results.

Conclusion

The current study has explored how an item bank used for adaptive testing purposes may be assessed in terms of robustness. In the study, item bank stability was investigated using a simulated 'full' dataset generated through model-based imputation. Three hypotheses were pursued in this study.

Hypothesis 1 was that the regression line (R^2) value of the simulation would be a minimum of 0.75. The R^2 values for the simulation was 0.99, and this hypothesis was accepted.

Hypothesis 2 was that Rasch infit and outfit statistics would be within acceptable ranges at the 25^{th} and 75^{th} percentiles. For both live dataset values and simulated dataset values (the latter using the 'full' dataset) at the both percentiles, fit statistics were well within acceptable ranges. This hypothesis was therefore also accepted. There was evidence of misfit with outfit mean squares although this was only at the maximum value end of the scale.

Hypothesis 3 was that the Bayes Factor would be in the range of at least 30-100. The Bayes Factor which emerged was 14,830 – well above the target of 100, and indicative of "extreme evidence".

The conclusion which may be drawn from the comparison of the 'full' and (comparatively sparser) live dataset was that as the live dataset expands in terms of data points (i.e., items and test takers), stability is likely to improve further. Such apparent stability lends credence to the claim that the items that comprise the adaptive item bank are of good quality and have been well set – and lends support to the robustness of the bank as an assessment instrument.

The current study has been laying the background for a follow-up study. The ground work – item bank stability – has now been established. The follow-up study involves a real-world use of the item bank. This study will involve the construction, administration to a representative sample of test takers, and analysis of a number of linear tests derived from the adaptive item bank. This study is reported in Coniam et al. (2022).

While the explorations reported in the current paper relate to the analysis of a specific item bank, the methodology may be useful to any researcher developing an item bank. Creating a simulated 'full' dataset allows for a view of the stability of the item bank to be evaluated, with the two statistics used in the current study offering a picture onto stability. A regression line above 0.75 gives a first line indication of the stability of the calibration of the item bank. The crucial figures, however, are calibration values and the Rasch infit and outfit statistics at the 25th and 75th percentiles. If the infit and outfit figures are within acceptable values, this is further evidence of stability in the item bank. Finally, a Bayesian ANOVA permits a prediction to be made as to the likely future stability of the item bank. If the Bayes Factor obtained from the ANOVA is 30-100 or higher, this is further "very strong evidence" as to the likely longterm stability of the item bank.

Notes

1. Reference is made in this paper to "one" item bank. It should be noted that LanguageCert tests access multiple parallel item banks.

References

[1] Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R., Verhagen, J., & Wagenmakers, E. J. (2015). An introduction to Bayesian hypothesis testing for management research. Journal of Management, 41(2), 521-543. https://doi.org/10.1177/0149206314560412.

[2] Bond, T. G., & Fox, C. M. (2007). Applying the Rasch model: fundamental measurement in the human sciences (2nd ed.). Mahwah, N.J.: Erlbaum.

[3] Choppin, B. (1968). Item Bank using sample-free calibration. Nature, 219, 870-872. https://doi.org/10.1038/219870a0.

[4] Coniam, D., Lee, T., Milanovic, M. & Pike, N. (2021). Validating the LanguageCert Test of English scale: The adaptive test. LanguageCert: London, UK.

[5] Coniam, D., Lee, T., Milanovic, M. (2022). Exploring Item Bank Stability in the Creation of Multiple Test Forms. LanguageCert: London, UK.

[6] Derner, S., Klein, S., & Hilber, D. (2008). Assessing the Feasibility of a Test Item Bank and Assessment Clearinghouse: Strategies to Measure Technical Skill Attainment of Career and Technical Education Participants. MPR Associates, Inc.

[7] Gao, F., & Chen, L. (2005). Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three parameter logistic model. Applied Measurement in Education, 18(4), 351-380. doi:10.1207/s15324818ame1804_2.

[8] Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. Psychonomic Bulletin & Review, 21, 1157-1164. https://doi.org/10.3758/s13423-013-0572-3

[9] Huisman, M., & Molenaar W. I. (2001). Imputation of missing scale data with item response models. In Boomsma, A., van Duijn, M., & Snijders, T. (Eds.). Essays on item response theory (pp. 221-244). New York: Springer-Verlag. https://doi.org/10.1007/978-1-4613-0169-1_13.

[10] Jeffreys, H. 1961. Theory of probability (3rd ed.). New York: Oxford University Press.

[11] Li, P., Stuart, E. A., & Allison, D. B. (2015). Multiple imputation: a flexible tool for handling missing data. Jama, 314(18), 1966-1967. https://doi.org/10.1001/jama.2015.15281.

[12] Linacre, J. M. (2012). A user's guide to WINSTEPS. Chicago, IL: Winsteps.com.

[13] Linacre, J. M. (2018). Winsteps Rasch measurement computer program user's guide. Beaverton, OR.

[14] Lunz, M. & Stahl, J. (1990). Judge consistency and severity across grading periods. Evaluation and the Health Profession, 13, 425-444. https://doi.org/10.1177/016327879001300405.

[15] Meijer, R. R. (1996). Person-fit research: An introduction. Applied Measurement in Education, 9(1), 3-8. https://doi.org/10.1207/s15324818ame0901_2.

[16] Mills, C. N., & Steffen, M. (2000). The GRE computer adaptive test: Operational issues. In Computerized adaptive testing: Theory and practice (pp. 75-99). Dordrecht: Springer. https://doi.org/10.1007/0-306-47531-6_4.

[17] Mislevy, R., & Wu, P. (1988). Inferring examinee ability when some item responses are missing (RR-88-48-ONR). Princeton NJ: Educational Testing Service. https://doi.org/10.21236/ADA201421.

[18] Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. Review of Educational Research, 74, 525-556. https://doi.org/10.3102/00346543074004525.

[19] Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. Advances in International Marketing, 20, 277-319. https://doi.org/10.1108/S1474-7979(2009)0000020014.
[20] Roth, P. (1994). Missing data: A conceptual review for applied psychologists. Personnel Psychology, 47, 537-560. https://doi.org/10.1111/j.1744-6570.1994.tb01736.x.
[21] Rudner, L. M. (2009). Implementing the graduate management admission test computerized adaptive test. In Elements of adaptive testing (pp. 151-165). New York, NY: Springer. https://doi.org/10.1007/978-0-387-85461-8_8.
[22] Sahin, A., & Weiss, D. J. (2015). Effects of calibration sample size and item bank size on ability estimation in computerized adaptive testing. Educational Sciences: Theory & Practice, 15(6), 1585-1595.
[23] Schminkey, D. L., von Oertzen, T., & Bullock, L. (2016). Handling missing data with multilevel structural equation modeling and full information maximum likelihood techniques. Research in Nursing & Health, 39(4), 286-297. https://doi.org/10.1002/nur.21724.
[24] Voss, S., & Blumenthal, Y. (2020). Assessing the Word Recognition Skills of German Elementary Students in Silent Reading-Psychometric Properties of an Item Pool to Generate Curriculum-Based Measurements. Education Sciences, 10(2), 35. https://doi.org/10.3390/educsci10020035.
[25] Vriens, M., & Melton, E. (2002). Managing missing data. Marketing Research, 14(3), 12.
[26] Weiss, D. J., & von Minden, S. V. (2012). A comparison of item parameter estimates from Xcalibre 4.1 and Bilog-MG. St. Paul, MN: Assessment Systems Corporation.
[27] Wright, B. D. (1997). A history of social science measurement. Educational Measurement: Issues and Practice, 16(4), 33-45. https://doi.org/10.1111/j.1745-3992.1997.tb00606.x.
[28] Zhang B., & Walker C. M. (2008). Impact of missing data on person-model fit and person trait estimation. Applied Psychological Measurement 32(6) 466-479. https://doi.org/10.1177/0146621607307692.