NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Denis Federiakin

INVESTIGATING THE CROSS-NATIONAL COMPARABILITY OF TESTING USING RESPONSE TIMES

BASIC RESEARCH PROGRAM
WORKING PAPERS

SERIES: EDUCATION
WP BRP 57/EDU/2020

Denis Federiakin[1]

INVESTIGATING THE CROSS-NATIONAL COMPARABILITY OF TESTING USING RESPONSE TIMES[2]

The cross-national comparability of testing results is one of the main concerns associated with international assessment. One way to approach this concern is the theoretical framework of cross-national comparability. One such framework, proposed by Van de Vijver and Tanzer [1997], describes the bias in item meaning across countries and the bias in terms of test administration procedures and respondent behavior across countries. To study how items function in terms of respondent behavior and in terms of their psychological meaning, we introduce the concept of time-related differential item functioning (DIF). This concept is similar to traditional DIF but describes the incomparable time parameters of items across countries rather than item difficulty parameters. We discuss time-related DIF, referencing dual processing cognitive theory and illustrate how it limits the interpretability of a certain type of time-related DIF. We also demonstrate the analysis with real data, and discuss the difference introduced by response time information in parameter interpretation and modelling results.

JEL Classification: Z0.

Keywords: Cross-national comparability, method bias, item response time, Rasch model, differential item functioning.

[1] National Research University Higher School of Economics. Center for psychometrics and Educational Measurement at Institute of Education. Research intern; E-mail: dafederiakin@hse.ru

[2] Support from the Basic Research Program of the National Research University Higher School of Economics is gratefully acknowledged.

Introduction

One of the main concerns associated with international assessment is the cross-national comparability of testing materials. Cross-national comparability implies that the measure functions the same ways across groups. If the comparability of the test has not been studied, it is impossible to determine whether the differences between countries (or their absence) are due to real differences in the measured construct or to differences in other factors such as language and culture [Sahin et al., 2015]. Moreover, the incomparability of a test may cause poor psychometric results [Church et al., 2011].

One of the ways to study cross-national comparability was suggested by Van de Vijver and Tanzer [1997], who proposed three major sources of biases, which can threaten the comparability of the results obtained from international studies. These groups are construct bias, method bias and item bias. Construct bias means that the theoretical construct itself has a different meaning between groups. This refers to the conceptual incomparability of the construct structure and meaning across countries in cross-national research [Stegmueller, 2011]. Method bias refers to a range of possible reasons for the incomparability of results. Usually, researchers list problems such as differences in samples, procedural differences during the survey, differential familiarity with stimuli and response procedures [Davidov, et. al., 2014]. Item bias refers to problems at the task-level, such as unrelated to the ability of interest factors (items may require additional secondary traits or abilities), culturally specific connotations of the item content, or poor translation [Hambleton, 1993].

An unspoken feature of this approach is that the elimination of the bias within each group implies that comparability has been achieved. To prove that the results are comparable across groups, researchers and practitioners working in the field of cross-national assessment usually study construct dimensionality (configural measurement invariance to prove the absence of construct bias), differential item functioning (DIF) (scalar and metric invariance to prove the absence of item bias) [Putnick, Bornstein, 2016], and develop and standardize procedures of sampling and test administration across groups (to prove the absence of method bias) [Ercikan, Lyons-Thomas, 2013]. However, statistical procedures to ensure that the method bias is eliminated, have remained overlooked and undeveloped for a long time.

To study cross-national comparability in terms of method bias, one should study response process data. Examples of response process data are log data, eye-tracking data, and response time data [Goldhammer, et. al., 2014]. Although research in computational behavioral science seeks to incorporate these data into modelling Item Response theory, only models for the simultaneous modelling of response times and response accuracy remain well-developed [De Boeck, & Jeon,

2019]. Therefore, one of the ways to investigate cross-national comparability in terms of method bias is to study information from response times.

Although the cross-national comparability of assessment is a complex phenomenon, this study is primarily concerned with the approach to the analysis of cross-national comparability through method bias and item bias. Particularly, we seek to investigate, whether the standardization of the test administration procedures across countries is enough to prove the absence of method bias. To do so, we (i) enhance the understanding of DIF by including the cross-national comparability of item response times, (ii) we examine which effects the introduction of response times in a measurement model has on the results, and (iii) we check if the enhanced definition of DIF supports the cross-national comparability of real data.

Item Response Theory Modelling Framework

Item Response Theory (IRT) is a statistical framework describing the interaction between items (and their properties) and respondents (and their properties). The formal definition of IRT suggests that the separation of personal parameters and item parameters define the probability of a correct response [van der Linden, 2016]. The simplest model for item response accuracy is the unidimensional dichotomous Rasch model [Rasch, 1960]:

$$g(P_{pi}) = \theta_p + \delta_i, \quad (1)$$

where P_{pi} is the probability of respondent p mastering item i,

$g(.)$ is the function of choice to describe the change in probability of a correct answer with respect to ability (in this study we use inverse logit function),

θ_p is the ability measure of respondent p on a scale defined by function $g(.)$,

and δ_i is the easiness measure of item i on the same scale as θ_p (the higher the numerical value, the easier the item).

The Rasch model has received a lot of attention in the psychometric literature over the years. It has been extended to many special cases such as polytomous items, multidimensional tests, tests with items locally dependent on personal parameters, models with a mixture of distributions of persons, models for splitting item and person parameters into parameters of attributes, etc. [see, De Boeck, & Wilson, 2004]. However, in this study we focus on one of the recent advances of this model – a model with response time parameters. As described by Goldhammer, et al. [2014], the most saturated model may be described as:

$$g(P_{pi}) = \theta_p + \delta_i + (\beta_0 + \beta_p + \beta_i) * \ln T_{pi}, \quad (2)$$

where β_0 is general time intercept,

β_p is the time parameter of respondent p,

β_i is the time parameter of item i,

T_{pi} is time taken by respondent p to give an answer to item i (note, that in the equation (2), the natural logarithm is taken of T_{pi} to normalize the response time distribution, since it has a natural bound of 0 and an inflated right tail).

The interpretation of the parameters from equation (2) deserves an extended discussion. Time-related parameters are not speediness and laboriousness parameters, as in some other models for response time and response accuracy [Molenaar, Tuerlinckx, & van der Maas, 2015]. These parameters describe how sensitive the probability of a correct response is to the change in response time. Thus, the general time intercept describes whether the test tends to be an ability test (if $\beta_0 > 0$) rather than a speediness test (if $\beta_0 < 0$). Similarly, item-specific and respondent-specific time parameters describe the measure of sensitivity (of an item or of a respondent) to the length of time. If the numerical value is positive, then with additional time the probability of a correct response increases, and vice versa. Therefore, the distributions of β_p and β_i describe how much respondents and items vary in this sensitivity. Moreover, the cross-classification parametrization of the model allows the computation of the correlations of parameters from the same facet (θ_p with β_p, and δ_i with β_i). These correlations describe the general strategies of the sample. If a correlation of θ_p with β_p is positive, then the higher the performance, the more sensitive to time limits the respondent, and vice versa. The same interpretation holds for the correlation of δ_i with β_i [Goldhammer, et al., 2014].

Differential item functioning

DIF occurs when test-takers from different groups with the same ability have different probabilities of endorsing or correctly answering a particular item [Holland & Wainer, 2012]. When an item demonstrates DIF, it no longer measures the same construct across different groups of test-takers. Thus, the test scores can no longer be considered comparable across groups [Van de Vijver & Tanzer, 1997]. The solution to a detected DIF may be “splitting” item parameters. That is, since an item does not measure the same construct across groups of respondents, its parameter cannot be the same across them. Therefore, group-specific item parameters are estimated by introducing “virtual items” – treating one real item as a series of group-specific items [Fischer, 1983].

DIF analyses in practice turn into a test of whether the difference in group-specific item parameters is statistically significant. An illustrative example of such an approach to DIF analysis

(which we use in this study) is the application of the Wald test for the purpose of DIF analysis within Rasch modelling [Fischer, Molenaar, 1995]:

$$W_i = \frac{\delta_{i|G=1} - \delta_{i|G=2}}{\sqrt{\sigma^2_{i|G=1} + \sigma^2_{i|G=2}}}, \quad (3)$$

where W_i is the z-statistic for item i,

$\delta_{i|G=1}$ is the easiness of item i for the first group,

$\delta_{i|G=2}$ is easiness of item i for the second group,

$\sigma_{i|G=1}$ is the standard error of the easiness of item i for the first group,

$\sigma_{i|G=2}$ is the standard error of the easiness of item i for the second group.

If the value of W-statistics exceeds the critical value of the z-distribution at a desired level of statistical significance, an item is said to exhibit DIF. Since the numerator of fraction (3) contains a subtraction, it is required for the group-specific item easiness parameters to lie on the same scale. This is usually achieved by an iterative purification DIF-analyses strategy, when items are analyzed for DIF one-by-one, thus the easiness parameters of all other items are estimated the same way for both groups [French, & Maller, 2007].

However, DIF analysis under the iterative purification strategy requires some way of controlling Type I errors. This is important, since multiple hypothesis testing may cause the inflation of Type I errors and the flagging of some DIF-free items as exhibiting DIF. In Rasch modelling, this is done by the introduction of DIF classification based on the effect size [see Paek, & Wilson, 2011]. If an item demonstrates a difference in group-specific item easiness parameters less than 0.426 logits (even if the DIF effect is statistically significant), DIF is marked as category A and is typically ignored. If the DIF effect size is less than 0.638 logits, DIF is marked as category B, and requires additional item content research to clarify whether the effect may be claimed as a real incomparability issue. If the DIF effect size is 0.638 or more, DIF is marked as category C and cannot be ignored. In our study we focus only on category C.

Including time-related parameters in the model begs the same logic of checking the comparability across groups. Therefore, it is possible to speak of two types of comparability on the item level – one which regards the psychological meaning of an item (that is, item position in the group-specific item hierarchy in Rasch modelling, based on response accuracy) and one which regards the item-solving behavior of respondents (that is, the similarity in how respondents process information across groups based on the response time). This additional type of time-related DIF naturally raises issues of method bias in the investigation of comparability on the item-level. Time-

related DIF may be defined as statistically significant difference across groups in time-related item parameters, controlling for the level of ability. This means that a time-related DIF occurs when respondents with the same level of ability from different groups use their time on task differently.

Nonetheless, time-related DIF analysis in the model from equation (2) requires some additional clarification. First, there is no established effect-size classification for the effect-size of time-related DIF. This means, that for now there is no way to judge the real impact of time-related DIF. Second, it is technically possible to conceive of three types of models for DIF analysis with time-effect considerations: (i) the model where the item easiness parameter is split across groups but item time-effect is not, (ii) the model where item time-effect is split across groups but the item easiness parameter is not, and (iii) the model where both the item easiness parameter and item time-effect are split. However, from an interpretational point of view it is difficult to interpret the second model – there an item is assumed to have the same psychological meaning across groups, but the behavior caused by this psychological meaning is assumed to differ across groups [Goldhammer, 2015]. Therefore, in this study we consider only models (i) and (iii).

Analysis of real data

In our exemplary analysis we use data from the Study of Undergraduate PERformance (SUPERtest project), which aims to internationally examine student learning in higher education. In particular, this study assesses the skill levels and gains among computer science and electrical engineering students across multiple countries and helps identify the contextual factors that impact student learning [Kardanova, et. al., 2016]. As part of the study, randomized nationally representative samples of fourth-year undergraduate students from China and Russia took the test. The total sample size was 1,662 Chinese students and 853 Russian students. The testing was conducted in November-December 2016. The test consisted of 25 dichotomous items in multiple choice or short constructed response format. The testing was conducted in computerized format, which allowed the project team to gather data about item response times. Time of access to the test materials was limited to 90 minutes, however none of the students used all 90 minutes.

All described IRT-models have been estimated as cross-classification models [Van den Noortgate, De Boeck, & Meulders, 2003]. Random effects of item easiness parameters and item time parameters have been constrained to be orthogonal. This was done to facilitate comparisons of the baseline model with models for time-related DIF analysis, and to save computational time. All models controlled for the average level of ability within groups to increase the trustworthiness of the analyses, since overlooking this effect may lead to confounded results [see Paek, & Wilson, 2011]. To compare the baseline models (with and without time-related effects), we used the loglikelihood ratio test, and

several information criteria, which compare the likelihood of the data given the parameter estimates, penalizing the more complicated model for extra-parameters (AIC) with respect to sample size (BIC). However, to compare the models with different numbers of random effects, we used conditional AIC (cAIC) to avoid bias due to the uncertainty of the random effect realization as suggested by Vaida and Blanchard [2005]. All models have been estimated with lme4 package v. 1.1-21 [Bates, et. al., 2019] for R v. 3.6.1.

The first step of the analysis calculated baseline models – one with response-time effects taken into consideration and one without. These models do not contain any country-specific parameters, all item parameters were treated as common across countries to facilitate model comparison. Table 1 contains information about the baseline models.

Table 1. Comparison of baseline models.

Model	-2LL	Number of parameters	AIC	Conditional AIC	BIC	Variance of respondents	Variance of items	Difference between means of distributions (SEM)	Average of Chinese sample relative to average of Russian sample (SEM)
Without response time	53135.2	4	53143.2	49465.4	53178.4	0.392	1.292	-1.033 (0.229)	0.264 (0.035)
With response time	52063.4	8	52079.4	48830.4	52149.7	0.515	2.459	-2.350 (0.319)	0.440 (0.032)

The model with response-time effects estimated 4 additional effects: variance of time effects for items (0.085), variance of time effects for respondents (0.012), general time intercept (0.345, SEM = 0.060), and the correlation between ability parameters and time effects for respondents (-0.71). Since the models are nested (the model without time effects can be seen as a model with time effects with 4 additional parameters constrained to zero), it is possible to compare their global model fit loglikelihood ratio test. The results suggest that the model with response times fits the data significantly better (χ^2=1071.8, d.f. = 4, p<0.001). The results are also supported by all information criteria.

Although the variances of time parameters are small, their introduction into the model increases the variances of respondent's ability and items easiness. It also increases the difference between averages of item distribution and person distribution and the difference between national samples. Additionally, the correlation between person effects (ability and time parameters) and the general time intercept suggest that persons who performed better on the test were less sensitive to changes in time spent to solve items. The interpretation of this from the test side implies that the test was aimed to measure not speediness or automated tasks but complex competence requiring cognitive control while solving the items.

After such a comparison, the correlations between the corresponding parameter estimates across models were computed. The correlation of item easiness across two models was 0.728, and the correlation of respondent ability was 0.954. Thus, IRT-models with response-time effects somehow change the latent space they describe, relative to the IRT-models without response time. This change in parameter space is shown in Figure 1. Figure 1A shows how person ability has changed after introducing time-related effects to the model (the x-axis is the estimated ability from the Rasch model without time effects; y-axis is the estimated ability from the model with time effects; the dots are colored based on the magnitude of standardized residuals from the univariate linear regression of the x-axis on the y-axis represented by a blue line). Figure 1B shows how item easiness has changed after introducing time-related effects into the model (Figure 1B has the same composition as Figure 1A). Both figures have the metric of the originally estimated logit scale without adjustment for the general model intercept.

As Figure 1A suggests, most of the difference in the ability space occurs near the middle of the ability distribution and within its right tail. However, the change in order of respondents is non-uniform – time effects provoke an overestimation of ability and an underestimation of it relative to the Rasch model without time effects. As Figure 1A illustrates, if response-time effects are considered, there are no sufficient statistics for the ability estimation (which is the raw score for the Rasch model without considering time effects [Andersen, 1977]). Figure 1B also suggests a non-uniform change in item space. However, the main difference is located in the left tail of the item easiness distribution. Overall, response-time effects cause a larger change in the parameter estimates for more difficult items and more capable respondents. The lower correlation of item parameters relative to person parameters may be, at least partially, due to the larger sample of respondents (P=2,515) relative to the number of items (I=25).

Figure 1A. Change in space of person ability

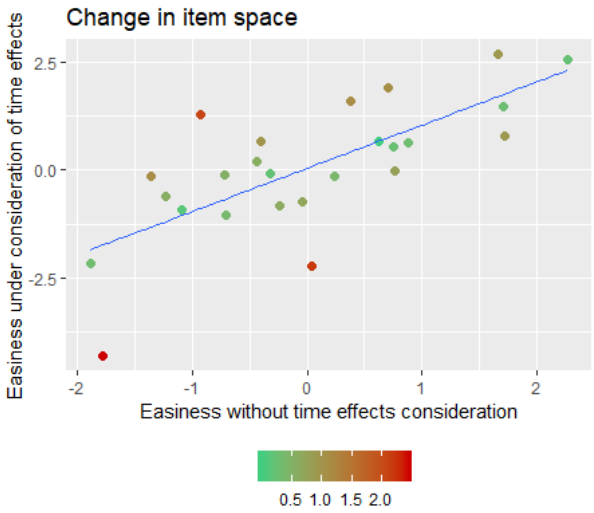

Figure 1B. Change in space of item easiness

After this comparison, a traditional DIF analysis was conducted without consideration of response-time effects. These results served as a baseline for the comparisons with the other steps. These results are presented in Table 2. The results suggest that the test contains 5 items with DIF of category C (4 in favor of China). After splitting the item easiness parameters of DIF-exhibiting items across countries and repeating the analysis among the remaining items, no more DIF were discovered.

The results of the DIF analysis considering time effects are presented in Tables 3 and 4. Table 3 contains results from the models where item easiness parameters are split across countries, but item time-effects are not. The results from Table 3 suggest that 6 items exhibit DIF of category C across countries (3 in favor of China). Table 4 contains the results from the models where both item easiness parameters and item time parameters were split across countries. The results from Table 4 suggest that 8 items exhibit DIF in terms of item easiness (1 in favor of China), and 7 items exhibit time-related DIF. After splitting the incomparable parameters of DIF-exhibiting items across countries and repeating the analysis among the remaining parameters, no more DIF were discovered.

Particularly interesting results come from the final models (where both item parameters were split across countries). They suggest 9 items in total with some form of DIF. Only 5 items demonstrate total DIF (in both easiness and time parameters), 2 items demonstrate DIF only in easiness, and 2 items demonstrate DIF only in time effects. The latter is difficult to interpret. Even if items have different psychological meaning (no DIF in easiness, items have the same psychological interpretation and are stable within the construct across countries), they still can exhibit DIF in time parameters (difference in response behavior). An explanation for such findings may be the effect size of time-related DIF: it is smaller than the 5 items which demonstrate total DIF. Therefore, it could be an example of Type I error.

Table 5 contains the results of the comparisons of the items marked as DIF-items across all tested models. The results suggest that as the complexity of the model grows, DIF may either gain effect size or lose it, but it does not change the group which it favors. It appears that response-time information may significantly alter the results of DIF analysis compared to the traditional approach by detecting new DIF items and making previously discovered DIF items disappear.

Conclusion and discussion

Any cross-national research requires establishing the level of cross-national data comparability appropriate for the comparison. For a meaningful comparison it must always be clear if the obtained difference (or absence) exists because of a real difference in the measured ability or because of differential measure functioning [Sahin, et al., 2015]. Since cross-national comparability is a complex phenomenon, plenty of preparatory and post hoc studies should be conducted to confirm this.

The rapid development of computerized testing allows researchers to gather more and more information about the way respondents perceive and process information to arrive at a conclusion. However, to process such data, it is necessary to have a fitting statistical model which utilizes this information [De Boeck, & Jeon, 2019]. This means that items become described by more and more parameters, which more precisely describe the probability of a correct response in terms of information processing. The increasing number of model parameters allow for a more sophisticated analysis of comparability across groups.

One of the ways to approach the scope of the procedures for establishing cross-national comparability is proposed by Van de Vijver and Tanzer [1997]. They described three sources of bias in cross-national research – construct bias, item bias, and method bias. This paper is focused on studying method bias and item bias. Particularly, we study how respondents react to the items across groups. If they react the same way across groups, it supports the claim for the high level of cross-national comparability of the results and enhances the results of comparisons. To illustrate this, we use real data about the competences of Russian and Chinese engineering students.

To describe the issues of cross-national comparability, we describe an item-level aspect of method bias – time-related DIF. Just like traditional DIF, time-related DIF implies that an item provokes incomparable behavior across countries. A traditional interpretation of the usual DIF is that item measures not (only) the target construct but some additional nuance of the dimension which differs between respondent groups [Shealy, Stout, 1993]. This implies that the psychological meaning of the item differs across the countries. Therefore, the interpretation of time-related DIF regards the

cognitive behavior caused by differential psychological meaning. A time-related DIF implies that items incite different cognitive processes or strategies of item information processing across countries.

According to the Dual Processing theory (which is a source of inspiration for the IRT-models described here [Goldhammer, 2015]), there should not be differing results of behavior if this behavior itself does not differ from sample to sample. This limits the interpretable possibilities for types of DIF. We consider two types of DIF: (i) DIF only in the easiness parameters (regardless of whether or not time parameters are estimated) – this type regards only item meaning and item content, (ii) total DIF – expressing in both the item easiness parameters and item time parameters, this type regards the meaning of the items as well as behavior they provoke. However, the results suggest that it is possible to discover another type of DIF existing in item-provoked behavior and not in item meaning. We believe such results may be caused by statistical fluctuations. We also suggest further research is needed to establish the time-related DIF effect-size classification, analogous to the existing classification based on usual DIF effect size.

Among other important results of this study, we discover that as the complexity of the model grows, some items stop demonstrating DIF and other item start to do so. This may be expected, since models with time effects describe a slightly different latent space of model parameters relative to the model without time effects, although the latent spaces appear to be highly correlated (especially in terms of ability estimates). Moreover, the introduction of a time effect into the model inflates all the effects in our case – both item sample and respondent sample become more heterogeneous; the difference between them, and between the Russian and Chinese samples in terms of ability, increases. This suggests that although the variation of time effects is small relative to the ability variation, these models change the interpretation of ability itself, as other research has shown [De Boeck, & Jeon, 2019; Bolsinova, et. al, 2017].

References

Andersen E. B. Sufficient statistics and latent trait models / Psychometrika. 1977. № 42(1). С. 69.

Bates D., Mächler M., Bolker B., Walker S. Fitting linear mixed-effects models using lme4 // arXiv preprint:1406.5823. 2014.

Bolsinova M., Tijmstra J., Molenaar D., De Boeck P. Conditional dependence between response time and accuracy: an overview of its possible sources and directions for distinguishing between them // Frontiers in psychology. 2017. № 8. С. 202.

Church A.T., Alvarez J.M., Mai N.T.Q., French B.F., Katigbak M.S., Ortiz F.A. Are cross-cultural comparisons of personality profiles meaningful? Differential item and facet functioning in the Revised NEO Personality Inventory // Journal of Personality and Social Psychology. 2011. № 101. С. 1068.

Davidov E., Meuleman B., Cieciuch J., Schmidt P., Billiet J. Measurement Equivalence in Cross-National Research // Annual Review of Sociology. 2014. № 40(1). С. 55.

De Boeck P., Jeon M. An overview of models for response times and processes in cognitive tests // Frontiers in psychology. 2019. № 10. С. 102.

Ercikan K., Lyons-Thomas J. Adapting tests for use in other languages and cultures. С. 545 / In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, M. C. Rodriguez (Eds.). 1st Ed. APA handbook of testing and assessment in psychology. Vol. 3. Testing and assessment in school psychology and education. 2013. Washington, DC: American Psychological Association.

Explanatory item response models: A generalized linear and nonlinear approach / P. De Boeck, M. Wilson (Eds.). 1st Ed. NY: Springer Science & Business Media. 2004.

Fischer G.H. Logistic latent trait models with linear constraints // Psychometrika. 1983. №48. С. 3.

Rasch Models: Foundations, Recent Developments and Applications / G. Fischer, I. Molenaar (Eds.). 1st Ed. NY: Springer-Verlag. 1995.

French B.F., Maller S.J. Iterative purification and effect size use with logistic regression for differential item functioning detection // Educational and Psychological Measurement. 2007. № 67(3). С. 373.

Goldhammer F. Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control // Measurement: interdisciplinary research and perspectives. 2015. № 13(3-4). С. 133.

Goldhammer F., Naumann J., Stelter A., Tóth K., Rölke H., Klieme E. The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment // Journal of Educational Psychology. 2014. № 106(3). С. 608.

Hambleton R. K. Translating Achievement Tests for Use in Cross-National Studies. NY: International Association for the Evaluation of Educational Achievement; Washington, DC: National Center for Education Statistics (ED). 1993.

Differential Item Functioning / P.W. Holland, H. Wainer (Eds.). 1st Ed. NY: Routledge. 1993.

Kardanova E., Loyalka P., Chirikov I., Liu L., Li G., Wang H., Enchikova E., Shi H., Johnson N. Developing instruments to assess and compare the quality of engineering education: the case of China and Russia // Assessment & Evaluation in Higher Education. 2016. № 41(5). С. 770.

Molenaar D., Tuerlinckx F., van der Maas H.L. A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times // Multivariate Behavioral Research. 2015. № 50(1). С. 56.

Paek I., Wilson M. Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel–Haenszel procedure in short test and small sample conditions // Educational and Psychological Measurement. 2011. № 71(6). С. 1023.

Putnick D. L., Bornstein M. H. Measurement invariance conventions and reporting: The state of the art and future directions for psychological research // Developmental review. 2016. № 41. С. 71.

Rasch G. Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. Oxford: Nielsen & Lydiche. 1960.

Sahin H., French B.F., Hand B., Gunel M. Detection of differential item functioning in the Cornell Critical Thinking Test between Turkish and United States students // European Journal of Psychological Assessment. 2014. № 31. С. 238.

Shealy R., Stout W. A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF / Psychometrika. 1993. № 58(2). С. 159.

Stegmueller D. Apples and oranges? The problem of equivalence in comparative research // Political Analysis. 2011. № 19(4). С. 471.

Vaida F., Blanchard S. Conditional Akaike information for mixed-effects models / Biometrika. 2005. № 92. С. 351.

van de Vijver F.J., Tanzer N.K. Bias and equivalence in cross-cultural assessment: An overview // European Review of Applied Psychology. 1997. № 47. С. 263.

van den Noortgate W., De Boeck P., Meulders M. Cross-classification multilevel logistic models in psychometrics // Journal of Educational and Behavioral Statistics. 2003. № 28(4). С. 369.

Handbook of Item Response Theory, Volume 1: Models / van der Linden W.J. (Ed.). 1st Ed. FL: CRC Press. 2016.

Appendixes

Table 2. Analysis of DIF without consideration of response time effects.

	Country-specific item easiness									
	Russia		China		DIF effect size (Russian easiness minus Chinese easiness)					
Item	Estimate	SEM	Estimate	SEM	Estimate	SEM	z-value	p-value	DIF flag	DIF favors
1	0.577	0.074	0.608	0.052	-0.031	0.090	-0.347	p>0.05		
2	0.925	0.073	0.802	0.053	0.123	0.090	1.373	p>0.05		
3	-0.785	0.142	-1.412	0.120	0.627	0.186	3.376	p<0.005	B	RUS
4	-0.299	0.122	-0.915	0.100	0.616	0.158	3.906	p<0.001	B	RUS
5	-1.954	0.161	-1.071	0.074	-0.884	0.178	-4.970	p<0.001	C	CHN
6	0.099	0.078	0.973	0.053	-0.874	0.095	-9.248	p<0.001	C	CHN
7	0.089	0.080	-0.650	0.064	0.740	0.102	7.247	p<0.001	C	RUS
8	-0.309	0.122	-0.297	0.085	-0.012	0.149	-0.080	p>0.05		
9	-0.226	0.084	0.644	0.053	-0.870	0.099	-8.749	p<0.001	C	CHN
10	-0.606	0.096	-1.037	0.072	0.431	0.120	3.592	p<0.001	B	RUS
11	-0.705	0.138	-1.227	0.112	0.522	0.177	2.948	p<0.005	B	RUS
12	-0.333	0.123	-0.461	0.089	0.128	0.151	0.849	p>0.05		
13	-0.392	0.125	-0.827	0.098	0.435	0.159	2.738	p<0.01	B	RUS
14	0.546	0.104	0.044	0.080	0.502	0.131	3.838	p<0.001	B	RUS

	Country-specific item easiness				DIF effect size (Russian easiness minus Chinese easiness)					
	Russia		China							
Item	Estimate	SEM	Estimate	SEM	Estimate	SEM	z-value	p-value	DIF flag	DIF favors
15	1.519	0.075	1.641	0.057	-0.121	0.094	-1.287	p>0.05		
16	0.327	0.078	-0.236	0.058	0.563	0.098	5.768	p<0.001	B	RUS
17	-1.491	0.136	-1.958	0.104	0.467	0.171	2.726	p<0.01	B	RUS
18	0.551	0.074	0.813	0.053	-0.262	0.091	-2.878	p<0.005		
19	-1.349	0.180	-1.883	0.147	0.534	0.232	2.300	p<0.05	B	RUS
20	0.059	0.082	0.037	0.056	0.022	0.099	0.217	p>0.05		
21	0.742	0.073	0.729	0.053	0.013	0.090	0.147	p>0.05		
22	1.090	0.101	2.015	0.088	-0.925	0.134	-6.922	p<0.001	C	CHN
23	1.634	0.076	1.660	0.058	-0.025	0.096	-0.266	p>0.05		
24	2.246	0.086	2.138	0.065	0.108	0.108	1.001	p>0.05		
25	0.153	0.110	-0.443	0.088	0.596	0.141	4.236	p<0.001	B	RUS

Table 3. Analysis of DIF considering of response time effects (item time effects are not split across groups).

	Country-specific item easiness									
	Russia		China		DIF effect size (Russian easiness minus Chinese easiness)					
Item	Estimate	SEM	Estimate	SEM	Estimate	SEM	z-value	p-value	DIF flag	DIF favors
1	0.711	0.254	0.625	0.236	0.086	0.347	0.249	p>0.05		
2	0.703	0.184	0.602	0.158	0.100	0.243	0.413	p>0.05		
3	-0.071	0.260	-0.805	0.239	0.734	0.354	2.077	p<0.05	C	RUS
4	-0.004	0.075	-0.215	0.076	0.211	0.106	1.981	p<0.05	A	RUS
5	-0.854	0.329	-0.098	0.277	-0.756	0.430	-1.759	p>0.05		
6	1.036	0.185	1.773	0.153	-0.737	0.240	-3.069	p<0.005	C	CHN
7	1.482	0.208	0.559	0.179	0.922	0.275	3.357	p<0.001	C	RUS
8	-0.141	0.291	-0.077	0.232	-0.064	0.372	-0.173	p>0.05		
9	0.880	0.189	1.586	0.163	-0.706	0.249	-2.832	p<0.005	C	CHN
10	1.860	0.216	1.151	0.184	0.709	0.284	2.498	p<0.05	C	RUS
11	-0.411	0.307	-0.988	0.277	0.577	0.414	1.394	p>0.05		
12	0.378	0.234	0.163	0.199	0.215	0.307	0.701	p>0.05		
13	-0.694	0.285	-1.008	0.241	0.313	0.373	0.840	p>0.05		
14	0.152	0.256	-0.107	0.199	0.259	0.325	0.798	p>0.05		
15	2.590	0.239	2.540	0.212	0.050	0.320	0.157	p>0.05		
16	-0.085	0.261	-0.675	0.240	0.590	0.355	1.663	p>0.05		

	Country-specific item easiness									
	Russia		China		DIF effect size (Russian easiness minus Chinese easiness)					
Item	Estimate	SEM	Estimate	SEM	Estimate	SEM	z-value	p-value	DIF flag	DIF favors
17	-1.511	0.404	-1.953	0.352	0.443	0.536	0.827	p>0.05		
18	0.287	0.187	0.526	0.165	-0.240	0.250	-0.958	p>0.05		
19	-3.497	0.495	-3.970	0.461	0.473	0.676	0.699	p>0.05		
20	-2.546	0.329	-2.232	0.288	-0.315	0.437	-0.721	p>0.05		
21	-0.072	0.202	-0.359	0.163	0.287	0.260	1.103	p>0.05		
22	-0.048	0.278	1.270	0.211	-1.318	0.349	-3.775	p<0.001	C	CHN
23	0.603	0.187	0.757	0.159	-0.154	0.246	-0.628	p>0.05		
24	2.562	0.213	2.422	0.177	0.140	0.277	0.505	p>0.05		
25	-0.343	0.273	-0.803	0.231	0.461	0.358	1.286	p>0.05		

Table 4. Analysis of DIF considering of response time effects (item time effects and item easiness parameters are split across groups).

Item	Country-specific parameters								DIF effect size									
	Item easiness				Time parameters				Russian easiness minus Chinese easiness						Russian time parameters minus Chinese time parameters			
	Russia		China		Russia		China											
	Estimate	SEM	Estimate	SEM	Estimate	SEM	Estimate	SEM	Estimate	SEM	z-value	p-value	DIF flag	DIF favors	Estimate	SEM	z-value	p-value
1	1.252	0.464	0.444	0.281	-0.216	0.089	-0.075	0.056	0.808	0.542	1.490	p>0.05			-0.142	0.105	-1.346	p>0.05
2	-0.172	0.360	0.943	0.179	0.216	0.081	-0.074	0.045	-1.115	0.402	-2.778	p<0.05			0.290	0.093	3.128	p<0.005
3	0.761	0.391	-1.227	0.297	-0.412	0.117	-0.023	0.081	1.988	0.491	4.047	p<0.001	C	RUS	-0.390	0.143	-2.729	p<0.01
4	0.549	0.402	-0.339	0.275	-0.187	0.126	-0.131	0.089	0.888	0.488	1.821	p<0.05	C	RUS	-0.056	0.154	-0.361	p>0.05
5	-1.301	0.843	-0.031	0.298	-0.203	0.165	-0.310	0.063	-1.270	0.894	-1.420	p>0.05			0.108	0.177	0.610	p>0.05
6	1.706	0.382	1.670	0.169	-0.408	0.089	-0.221	0.044	0.036	0.418	0.085	p>0.05			-0.187	0.099	-1.881	p>0.05
7	1.145	0.388	0.802	0.205	-0.273	0.097	-0.427	0.058	0.343	0.438	0.783	p>0.05			0.154	0.113	1.362	p>0.05
8	0.308	0.529	-0.196	0.269	-0.159	0.133	-0.005	0.079	0.504	0.594	0.849	p>0.05			-0.154	0.155	-0.991	p>0.05
9	0.461	0.407	1.738	0.179	-0.174	0.104	-0.327	0.049	-1.277	0.445	-2.869	p<0.005	C	CHN	0.153	0.115	1.327	p>0.05
10	1.638	0.431	1.252	0.205	-0.572	0.108	-0.659	0.057	0.386	0.477	0.808	p>0.05			0.086	0.122	0.708	p>0.05
11	-1.175	0.548	-0.702	0.317	0.196	0.154	-0.121	0.098	-0.473	0.633	-0.748	p>0.05			0.316	0.183	1.730	p>0.05
12	1.661	0.391	-0.298	0.235	-0.571	0.126	0.014	0.076	1.959	0.457	4.290	p<0.001	C	RUS	-0.585	0.147	-3.970	p<0.001
13	-0.172	0.452	-1.264	0.294	-0.013	0.129	0.228	0.092	1.092	0.539	2.026	p<0.05	C	RUS	-0.241	0.158	-1.520	p>0.05
14	0.540	0.385	-0.212	0.246	0.077	0.118	0.244	0.094	0.752	0.457	1.645	p>0.05			-0.168	0.151	-1.114	p>0.05
15	4.129	0.450	2.057	0.250	-0.621	0.095	-0.174	0.059	2.071	0.515	4.021	p<0.001	C	RUS	-0.447	0.112	-3.996	p<0.001
16	-0.676	0.505	-0.436	0.277	0.142	0.103	-0.033	0.060	-0.240	0.576	-0.417	p>0.05			0.174	0.119	1.463	p>0.05
17	-0.371	0.660	-2.652	0.455	-0.295	0.142	0.117	0.103	2.281	0.802	2.846	p<0.005	C	RUS	-0.411	0.175	-2.347	p<0.05
18	1.017	0.349	0.337	0.191	-0.119	0.090	0.129	0.053	0.681	0.398	1.710	p>0.05			-0.248	0.105	-2.368	p<0.05
19	-3.351	0.754	-4.785	0.654	0.538	0.182	0.787	0.162	1.434	0.998	1.436	p>0.05			-0.249	0.243	-1.025	p>0.05
20	-3.028	0.626	-2.127	0.344	0.533	0.121	0.418	0.075	-0.901	0.714	-1.261	p>0.05			0.116	0.142	0.813	p>0.05
21	-0.385	0.378	-0.022	0.187	0.247	0.089	0.227	0.053	-0.364	0.422	-0.862	p>0.05			0.020	0.104	0.190	p>0.05
22	0.689	0.432	1.051	0.256	0.090	0.108	0.355	0.084	-0.361	0.502	-0.720	p>0.05			-0.265	0.136	-1.944	p>0.05

Item	Country-specific parameters								DIF effect size									
	Item easiness				Time parameters				Russian easiness minus Chinese easiness						Russian time parameters minus Chinese time parameters			
	Russia		China		Russia		China											
	Estimate	SEM	Estimate	SEM	Estimate	SEM	Estimate	SEM	Estimate	SEM	z-value	p-value	DIF flag	DIF favors	Estimate	SEM	z-value	p-value
23	0.736	0.315	0.786	0.188	0.218	0.079	0.246	0.053	-0.051	0.367	-0.138	p>0.05			-0.027	0.096	-0.288	p>0.05
24	3.992	0.353	1.933	0.214	-0.501	0.113	0.178	0.084	2.059	0.412	4.994	p<0.001	C	RUS	-0.679	0.141	-4.828	p<0.001
25	0.039	0.405	-0.945	0.292	0.109	0.121	0.281	0.100	0.984	0.499	1.970	p<0.05			-0.172	0.157	-1.097	p>0.05

Table 5. Comparisons of the results across all models.

	Model without time effects	Models with time effects		
		Model with split item easiness parameters	Model where all items parameters are split	
Item	Item easiness (DIF in favor)	Item easiness (DIF in favor)	Item easiness (DIF in favor)	Item time parameter
2				+
3		RUS	RUS	+
4			RUS	
5	CHN			
6	CHN	CHN		
7	RUS	RUS		
9	CHN	CHN		
10		RUS		
12			RUS	+
13			RUS	
15			RUS	+
17			RUS	+
18				+
22	CHN	CHN		
24			RUS	+

Contact details:

Denis Federiakin

National Research University Higher School of Economics (Moscow, Russia). Center for psychometrics and Educational Measurement at Institute of Education

Research intern

101000 Moscow, Poratovsky lane 16, build. 10, room 4.10.

E-mail: dafederiakin@hse.ru

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.