International Journal of Public Administration, Management and Economic Development
Faculty of Administration and Economic Studies in Uherské Hradiště, Jagiellonian College in Toruń

On Test-Based Assessment and Evaluation

Oldřich Hájek

Newton University, 5. května 1640/65, Prague 140 21, Czechia,
oldrich.hajek@newtoncollege.cz

Jiří Novosák

Newton University, 5. května 1640/65, Prague 140 21, Czechia,
jiri.novosak@newtoncollege.cz

Jana Novosáková

Newton University, 5. května 1640/65, Prague 140 21, Czechia,
jana.novosakova@newtoncollege.cz

Blanka Vytrhlíková

Newton University, 5. května 1640/65, Prague 140 21, Czechia,
blanka.vytrhlikova@newtoncollege.cz

Abstract

The main intent of this paper is to introduce crucial components of a newly created methodology that contains practical and usable instructions how to perform tasks common to test-based assessment and evaluation (case studies). Theoretically, the methodology is based on two fundamental approaches to test-based assessment and evaluation: (a) classical test theory; and (b) item response theory, and these approaches were used (a) to identify eighteen case studies; and (b) to give step-by-step instructions how to perform them. In this regard, all case studies are specified against components of a general methodological framework, and such a specification is illustrated for a selected case study titled 'Test taker's proficiency level – test scale'. It is worth noting that a broader situational context is discussed in more detail. Finally, a link between the methodology and the Strategy for Education Policy of the Czech Republic 2030+ through their common focus on the quality of test-based assessment and evaluation is explained and emphasized.

Keywords: Assessment and evaluation; education; tests; classical test theory; item response theory

---

1. Introduction

Human capital is at the core of several theories of socio-economic development (see, e.g., Lucas, 2015; Diebolt and Hippe, 2019; Faria et al., 2016). The essence of these theories is that human capital relates positively to the long-term benefits of both human capital holders in the form of higher earnings and society in the form of economic growth. Moreover, education is regarded as a mechanism for human capital formation, and all these ideas, despite much criticism, motivate the interest in investments in education as well as in education policy (see, e.g., Gillies, 2017; Becker, 1992; Holden and Biddle, 2017).

Discussions regarding education policy cover a variety of different topics, including the development of the 21st century competencies (see, e.g., Erstad and Voogt, 2018). Voogt and Roblin (2012), Dede (2010), Binkley et al. (2013) point out in this regard that the competencies required for employment, active citizenship and self-realization in the 21st century differ from those of the 20th century. The main differences particularly reflect higher requirements for: (a) processing (understanding, evaluation and interpretation) of a wealth of easily accessible information; (b) a search for non-standardized solutions to non-recurrent problems; and (c) cooperation with complementary knowledge- and skill-holders. The importance of ICT, assessment and evaluation is also emphasised (see, e.g., Erstad and Voogt, 2018; Dede, 2010).

The highest-level strategic document that outlines a vision for education in the Czech Republic acknowledges the importance of the profound changes of the 21st century. Hence, according to the Strategy for Education Policy of the Czech Republic 2030+, the major societal changes in shaping the 21st century education have included: (a) economic and other transformations, resulting in new knowledge and skills needed for job performance; (b) a widespread and expanding use of ICT for communication and socialization; and (c) practically unlimited access to information, which must be critically examined (Fryč et al., 2020). Moreover, the strategy accentuates also the need for assessment and evaluation in education and gives an important role to large-scale testing. Regarding the last point, Voogt and Roblin (2012), Gillies (2017), Binkley et al. (2013) give the desirable qualities of assessment and evaluation in education as follows:

- the establishment of a conceptual framework for assessment and evaluation;
- the systematic identification of educational needs and feedback provision (formative assessment and evaluation);
- ICT support to assessment and evaluation;
- the fulfilment of quality criteria of assessment and evaluation.

This paper deals with the last of these four items, focusing on test-based assessment and evaluation. In this context, Thompson (2016), Voogt and Roblin (2012), Gillies (2017) claim that tests belong to the most common assessment and evaluation tools; however, their quality often does not meet desirable standards. In fact, this is what motivated us to prepare, under a project supported by the Technological Agency of the Czech Republic (hereafter referred to as 'TACR'), a methodology that contains practical and usable instructions how to perform tasks common to test-based assessment and evaluation (case studies). The link to education policy is obvious.

The goal of this paper is twofold: (a) to introduce crucial components of the methodology; and (b) to illustrate its application to a case study example. Concerning the latter goal, we specifically demonstrate different ways how to express the test takers' proficiency,

using tests with dichotomous items. The paper is structured as follows. The second section provides theoretical foundations for the methodology and particularly for the selected case study. The third section introduces the essence of the methodology. The fourth section illustrates the application of the methodology in the selected case study, which is further discussed in the fifth section. The last section concludes.

2. Literature review

There are several theoretical frameworks applied in test-based assessment and evaluation (see, e.g., Ziegler and Hagemann, 2015). However, as Ryan and Brockmann (2009) note, two main theoretical approaches are classical test theory (hereafter referred to as 'CTT' only) and item response theory (hereafter referred to as 'IRT' only), and this is why our methodology emanates just from these two theoretical approaches. Reflecting this, we carried out an extensive literature review on CTT and IRT in order to: (a) identify the tasks common to test-based assessment and evaluation; and (b) give step-by-step instructions how to perform these tasks. The following text present the main results of this review regarding the selected case study.

CTT is a traditional methodological approach to test-based assessment and evaluation (see, e.g., Traub, 1997; Raykov and Marcoulides, 2016). The main idea of this approach is that a latent (unobservable) variable (e.g., student competencies) is measured using appropriate tests (see, e.g., DeVellis, 2006; De Champlain, 2010). The fundamental equation of CTT is:

$$X = T + E,$$

where X is an observed test score, T is a true (unobserved) test score, and E is an error score, i.e. the difference between the observed and true test scores (see, e.g., Hambleton a Jones, 1993; Revelle, 2012; Brennan, 2011; De Champlain, 2010; Graham, 2006). Hence, the observed test scores (X) express the test takers' proficiency, and this theoretical background is, therefore, crucial for the selected case study.

IRT is a modern methodological approach to test-based assessment and evaluation (see, e.g., Thorpe and Favia, 2012; Toland, 2014; van der Linden, 2010; Rusch et al., 2017). The essence of this approach is the modelling of the relationship between the test takers' proficiency (traditionally labelled as Θ) and their patterns of responses to test items (see, e.g., DeMars, 2010; Hambleton and Jones, 1993; Toland, 2014; Orlando and Thissen, 2000; Van Zile-Tamsen, 2017). Similarly, De Champlain (2010) claims that the IRT approach estimates the probability of a correct response to a particular item as a function of item parameters and the test takers' proficiency (Θ). Hence, two major differences between CTT and IRT are: (a) different procedures for dealing with items; and (b) different ways of measuring the test takers' proficiency.

It is obvious that the way of measuring the test takers' proficiency (Θ) is crucial for the selected case study. Several IRT-based approaches have been suggested in this regard; however, van der Linden (2010) indicates the three-parameter logistic (3PL) model to be a good first option (standard). The 3PL model is defined as follows (DeMars, 2010; van der Linden, 2010; Hambleton and Jones, 1993):

$$P(x_i = 1; \theta_j) = c_i + (1 - c_i) * \frac{e^{1,7a_i(\theta_j - b_i)}}{1 + e^{1,7a_i(\theta_j - b_i)}},$$

where a_i is the discrimination parameter for item i, b_i denotes the difficulty parameter for item i, c_i is the pseudo-guessing parameter for item i, and Θ_j is the proficiency level of the test taker j. Additionally, the 3PL model may be simplified by assuming that the pseudo-guessing parameter is zero for all items. Then, the 3PL model is reduced to the two-parameter logistic (2PL) model, which excludes the pseudo-guessing parameter. Similarly, by assuming that the discrimination parameter is the same for all items, the 2PL model may be reduced to one-parameter logistic (1PL) model, and this model estimates the difficulty parameter only.

Besides the models, the way of measuring the test takers' proficiency (Θ) may also differ in estimation procedures. These may include the standard EM (expectation-maximization) algorithm but also other options such as Monte Carlo EM estimation or Monte Carlo with Markov chain simulations (see, e.g., Burgos, 2010). Several methods are also available to estimate the test takers' proficiency (Θ), particularly: (a) maximum likelihood; (b) expected a posteriori (EAP) estimation; and (c) maximum a-posteriori (MAP) estimation (Rupp, 2005).

Livingston (2014) notes a common desire to avoid the reporting of test results as scores. For this reason, another scale (e.g., a point scale) may be constructed. If proceeding in this manner, decisions have to be especially made about: (a) the scale unit (e.g., one point, ten points); (b) the upper and lower scale limits; and (c) the relationship between test scores and the alternative scale. Finally test results may be reported on a percentile scale (see, e.g., Livingston, 2014; Dorans, Moses and Eignor, 2010).

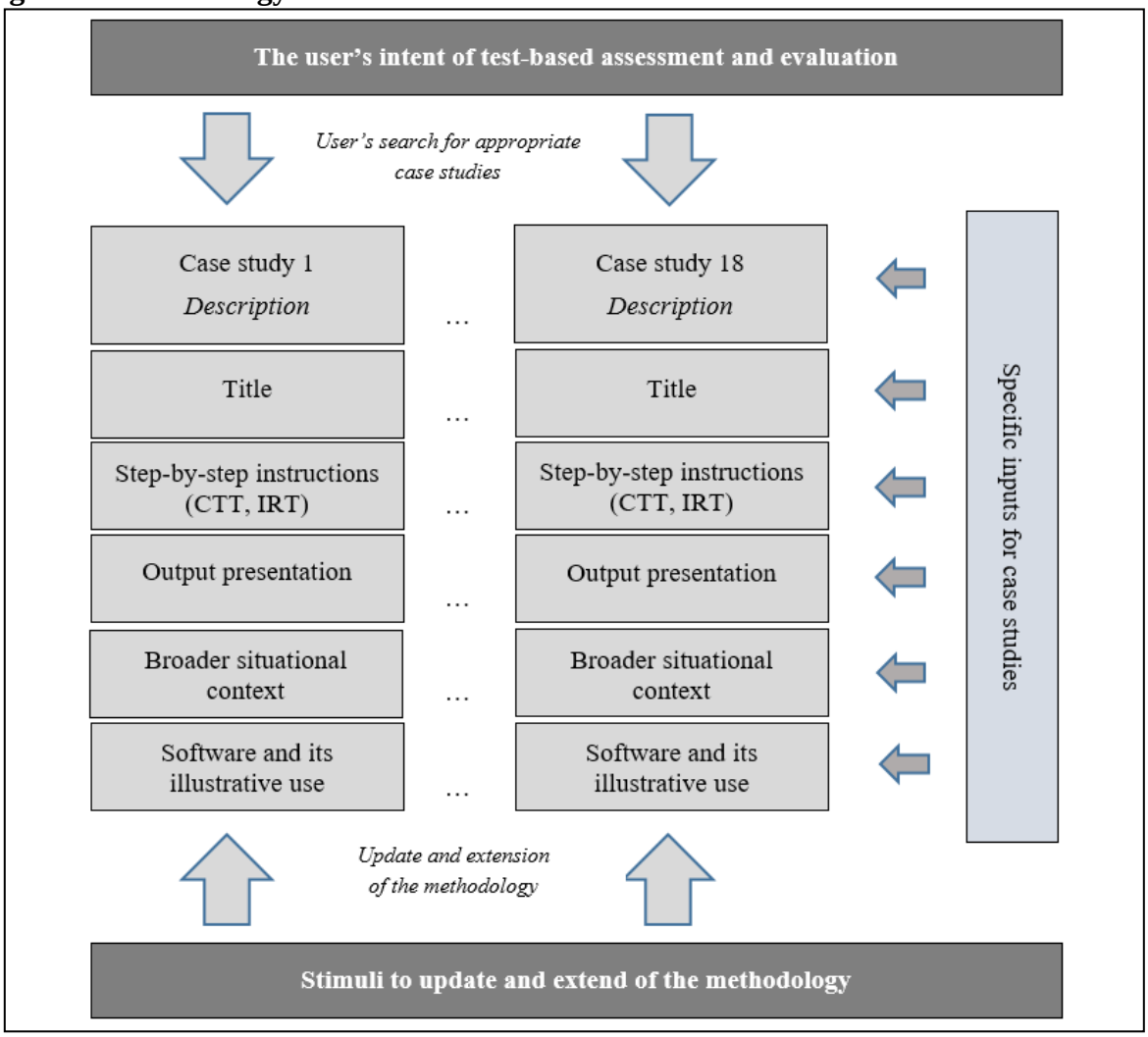
Overall, this theoretical background indicates a variety of ways how to express the test takers' proficiency. The learned information is subsequently used to elaborate the specification of the selected case study in accordance with the methodology. It is worth noting that the same procedure was adopted for other case studies covered by the methodology.

3. Methodology

The methodology of this paper is primarily based on the methodology developed as part of the TACR funded research project, which has investigated solutions to common tasks related to test-based assessment and evaluation. Conceptually, the methodology follows a case-study approach consisted of two parts. Firstly, a general framework common to all case studies has been built. Under this framework, the following components of each case study are described: (a) title; (b) brief description; (c) step-by-step instructions how to perform the case study by applying either CTT or IRT approaches; (d) output presentation; (e) broader situational context; and (f) relevant software and its illustrative use. Secondly, these components have been specified separately for each case study (see figure 1 for schema of the methodology). In sum, eighteen case studies have been identified and described (see table 1) using the literature review on CTT and IRT.

Two general principles of the methodology are noteworthy. Firstly, it is the user who decides how the methodology is used. In this regard, a brief description of each case study is provided to assist the users in their search for appropriate case studies. Secondly, the methodology is open to changes, including adaptation of the component items and addition of new components and case studies. Moreover, several R packages are used in the methodology to solve the case studies. The choice of the R environment has been motivated by its easy accessibility (General Public License) and its flexibility for advanced statistical analysis.

Figure 1: Methodology schema



Source: authors.

Table 1: Title

ID	Title
1	Test (scale) reliability
2	Test takers' mastery of an item, i.e. item difficulty
3	Item's proficiency to discriminate among test takers' of different standing on the scale, confusing (correct) responses, (potentially) incorrect scoring key
4	The quality of distractors in items
5	Change in test (scale) reliability if particular item is deleted
6	Item fairness (bias) with regard to test takers' group membership

ID	Title
7	Item quality and detection of low-quality items
8	Unusual test takers' response patterns to items
9	Test taker's proficiency level – test scale
10	Equating of test takers' scores in two tests linked together with anchoring items
11	Progress in education
12	Test unidimensionality and the number of inherent dimensions (domains)
13	Local independence of items
14	Test domains
15	The most appropriate IRT model for test assessment and evaluation
16	Assessment, evaluation and reporting of test results
17	Within and between school differences in test results
18	Determinants of test results

Source: authors.

4. Results

In this section, we illustrate the specification of the general methodological framework for the case study 9 titled 'Test taker's proficiency level – test scale'.

4.1 Brief description

The essence of this case study is to gain information on the test takers' proficiency in the tested domains, and that is the main intent of test-based assessment and evaluation. In this case study, the methodology user's task is: (a) to decide on the scale for measuring the test takers' proficiency in the tested domains; and (b) to use the scale for measuring the test takers' proficiency in the tested domains. In this regard, various methods can under certain assumptions be used. We remind that the methodology concerns the tests with dichotomous items.

4.2 Step-by-step instructions

The following procedure is used to achieve the intent of the case study, i.e. to measure the test takers' proficiency in the tested domains:

- Step 1: The methodology user decides whether he/she uses IRT models in order to construct the scale for measuring the test takers' proficiency in the tested domains.
- (No) Step 2: The methodology user decides on the scale for measuring the test takers' proficiency.
  - (No) Step 3a: The methodology user assigns a score to each test taker, i.e. the number of correctly answered items.
  - (No) Step 3b: The methodology user assigns the share of correctly answered items to each test taker.
- (No) Step 4: The methodology user decides whether he/she uses an alternative test scale, particularly: (a) a point scale with a mean *n* and standard deviation *m*; and (b) a percentile scale.
- (Yes) Step 2: The methodology user examines the assumptions of the traditional IRT models, particularly: (a) the required minimum sample size; (b) test unidimensionality (case study 12, see table 1); and (c) local independence of items (case study 12, see table 1).
  - (Yes) Step 3a: The methodology user decided that the assumptions given in the (Yes) Step 2 hold or he/she adopts measures in order to comply with the assumptions given in the (Yes) Step 2.
    - (Yes) Step 4a: The methodology user estimates the parameters of the: (a) 1PL model; (b) 2PL model; and (c) 3PL model.
    - (Yes) Step 5a: The methodology user decides on the most appropriate model (case study 15, see table 1).
    - (Yes) Step 6a: The methodology user extracts the test takers' proficiency in the tested domain (Θ).
    - (Yes) Step 7a: The methodology user decides whether he/she uses an alternative test scale, particularly: (a) a point scale with a mean *n* and standard deviation *m*; and (b) a percentile scale.
  - (Yes) Step 3b: The methodology user decided that the assumptions given in the (Yes) Step 2 do not hold and he/she estimates multidimensional IRT models.
    - (Yes) Step 4b: The methodology user estimates the parameters of a multidimensional logistic model. The number of dimensions (domains) is determined using the methodological approach given in the case study 12 (see table 1).
    - (Yes) Step 5b: The methodology user extracts the test takers' proficiency in the tested domains (Θ_i).
    - (Yes) Step 6b: The methodology user decides whether he/she uses an alternative test scale, particularly: (a) a point scale with a mean *n* and standard deviation *m*; and (b) a percentile scale.
  - (Yes) Step 3c: The methodology user decided that the assumptions given in the (Yes) Step 2 do not hold and he/she uses the CTT-based approach (see the (No) Step 2).

Note that the methodology user may consider the violation of the unidimensionality assumption also in the (No) Step 2.

4.3 Output presentation

The output of the case study may be: (a) test takers' scores – (No) Step 3a; (b) test takers' shares of correctly answered items – (No) Step 3b; (c) test takers' proficiency in tested domains – (Yes) Step 6a and (Yes) Step 5b; (d) test takers' points on an alternative scale – (No) Step 4,

(Yes) Step 7a and (Yes) Step 6b; and (e) test takers' percentiles – (No) Step 4, (Yes) Step 7a and (Yes) Step 6b.

4.4 Broader situational context (shortened)

The decision on the appropriate scale for measuring the test takers' proficiency may lead to different results. This is particularly true when more complex models that consider the discrimination and pseudo-guessing parameters are estimated (i.e., 2PL and 3PL models). Hence, the test taker who achieves a higher score than other test takers may have a lower proficiency measured on the 2PL or 3PL scales because of different weights given to particular items. Consequently, the decision on the appropriate scale for measuring the test takers' proficiency is also reflected in other case studies, including the case study 18 (see table 1) in which the scale is the dependent variable in estimated models.

4.5 Relevant software and its illustrative use (shortened)

Three R packages are recommended to solve the case study. These include the CTT package (see, e.g., Willse, 2018), ltm package (see, e.g., Rizopoulos, 2018) and mirt package (see, e.g., Chalmers, 2020). Moreover, the methodology specifies the commands needed to replicate the step-by-step instructions in the subsection 4.2.

5. Discussion

In this section, we present selected results of an application of the methodological framework for the case study 9 as given in section 4. These results are the output from a verification process of the methodology that has been carried out in cooperation with the Czech School Inspectorate. Particularly, we report the test takers' proficiency using the following test scales: (a) test scores; (b) the 1PL scale; (c) the 2PL scale; and (d) an alternative point scale with the mean set at 500 points and standard deviation at 100 points. The IRT proficiency levels were derived from the 1PL and 2PL models using the expectation-maximization maximum-likelihood (EM-ML) algorithm and expected a posteriori (EAP) estimation. Regarding the R packages, the functions rasch, ltm and factor.scores from the ltm package (see, e.g., Rizopoulos, 2018) and the function score.transform from the CTT package (see, e.g., Willse, 2018) were applied. Table 2 shows the results for several test takers.

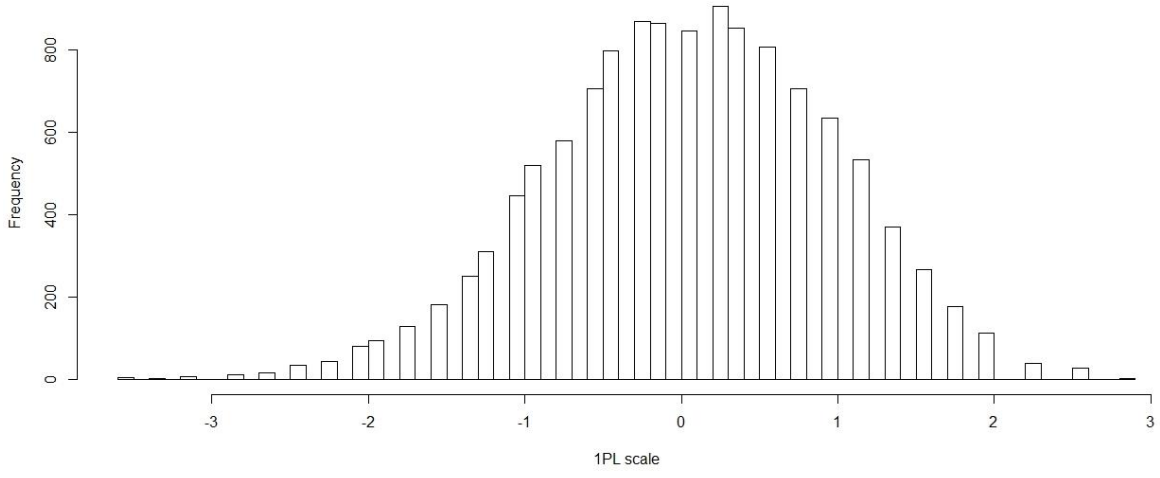
The results presented in table 2 clearly illustrates the broader situational context of the case study 9 (see subsection 4.4). Hence, there is a direct and unambiguous relationship between the test taker's score and his/her proficiency on the 1PL scale but the same relationship does not hold true for the 2PL scale (see, e.g., ID2 and ID3; and ID4 and ID10). Moreover, the test taker ID1 answered more items correctly than the test taker ID3; however, his/her proficiency level on the 2PL scale is even lower, and this is because of different weights given to particular items when estimating the 2PL model parameters. Then, it is not surprising that these findings may be less understandable to users. On the other hand, the 2PL scale allows users to better differentiate between test takers due to its more continuous nature (see figure 2 for differences in frequency distribution of proficiency levels measured on the 1PL and 2PL scales). The decision on the appropriate scale for measuring the test takers' proficiency ought to consider all these aspects.

Table 2: Test takers' proficiency – various test scales

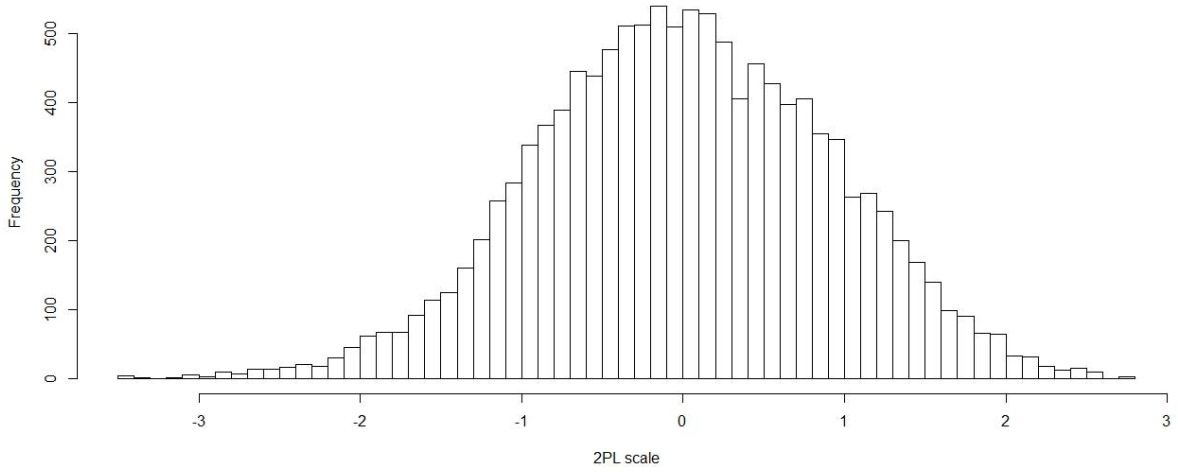
Test taker's ID	Score	1PL model		2PL model	
		θ	point scale	θ	point scale
1	25	0.9304	605	0.8051	589
2	23	0.5680	564	0.7790	586
3	23	0.5680	564	0.9503	605
4	26	1.1224	626	0.7777	586
5	16	-0.5909	434	-0.6761	426
6	18	-0.2689	470	0.0369	504
7	22	0.3935	544	0.1992	522
8	18	-0.2689	470	0.0848	509
9	17	-0.4310	452	-0.4431	451
10	26	1.1224	626	0.8309	592

Source: own elaboration in cooperation with the Czech School Inspectorate

Figure 2: Frequency distribution of proficiency levels measured on the 1PL (above) and 2PL (below) scales



Source: authors.



Source: own elaboration in cooperation with the Czech School Inspectorate

It is worth noting that the case study 9 is closely related to other case studies. Three of them are worth mentioning here. Firstly, it is desirable to examine whether the assumptions of particular models and techniques hold (e.g., case studies 1, 11 and 12). If not, the test takers' proficiency estimates may be biased or unreliable. Secondly, goodness of fit measures between observed and model simulated values can be used to decide on the appropriate IRT-based scale (case study 15). Thirdly, the number of test-inherent factors (domains, dimensions) may be determined by the step-by-step instructions given in the case study 12. It is crucial for the estimation of multidimensional models.

6. Conclusion

The Strategy for Education Policy of the Czech Republic 2030+ acknowledging the importance of the profound changes of the 21st century outlines a vision and objectives of Czech education for the time period until the year 2030+. To examine whether the vision and objectives are being achieved or political interventions are desirable and effective, the strategy emphasizes the role of assessment and evaluation in education, including large-scale testing. Consequently, the question on the quality of test-based assessment and evaluation is crucial.

This paper reflects the essential role that test-based assessment and evaluation has in achieving the vision and objectives of the Strategy for Education Policy of the Czech Republic 2030+ and introduces a methodology how to perform tasks common to test-based assessment and evaluation, i.e. case studies. In this regard, the essence of the methodology is presented using the case study of measuring test takers' proficiency on different scales for illustration. The other case studies of the methodology are specified analogous. In this way, a link between the methodology and the Strategy for Education Policy of the Czech Republic 2030+ is created through their common focus on the quality of test-based assessment and evaluation.

Acknowledgment

This paper was prepared as a result of the project TL01000385 titled 'Evaluation Methodology of Verification Testing Results in Primary and Secondary Education and Its

Application in Model Case Studies' and supported by the Technological Agency of the Czech Republic under the programme 'TL – ETA Programme for Applied Research, Experimental Development and Innovation in Social Sciences and Humanities'. The authors gratefully acknowledge the financial support from the Technological Agency of the Czech Republic.

References

- Becker, G. (1992). Human capital and the economy. *Proceedings of the American Philosophical Society*, 136(1), 85–92.
- Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2013). Defining twenty-first century skills. In P. Griffin, B. McGraw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17-66). Springer.
- Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24(1), 1-21.
- Burgos, J. G. (2010). Bayesian methods in psychological research: the case of IRT. *International Journal of Psychological Research*, 3(1), 163-175.
- Chalmers, P. (2020). *Package 'mirt'*. [online]. Available from: <https://cran.r-project.org/web/packages/mirt/mirt.pdf>.
- Dede, C. (2010). Comparing frameworks for 21st century skills. In J. Bellanca & R. Brandt (Eds.). *21st century skills: rethinking how students learn* (pp. 51-75). Solution Tree Press.
- De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, 44(1), 109-117.
- DeVellis, R. F. (2006). Classical test theory. *Medical Care*, 44(11/S3), 50-59.
- Diebolt, C., & Hippe, R. (2019). The long-run impact of human capital on innovation and economic development in the regions of Europe. *Applied Economics*, 51(5), 542-563.
- Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). Principles and practices of test score equating. *ETS Research Report Series*, 2, 1-41.
- Erstad, O., & Voogt, J. (2018). The twenty-first century curriculum: issues and challenges. In J. Voogt, G. Knezek, R. Christensen, & K. W. Lai (Eds.), *Second Handbook of Information Technology in Primary and Secondary Education* (pp. 19-36). Springer.
- Faria, H. J., Montesinos-Yufa, H. M., Morales, D. R., & Navarro, C. E. (2016). Unbundling the roles of human capital and institutions in economic development. *European Journal of Political Economy*, 45, 108-128.
- Fryč, J., Matušková, Z., Katzová, P., Kovář, K., Beran, J., Valachová, I., Seifert, L., Běťáková, M., & Hrdlička, F. (2020). *Strategie vzdělávací politiky České republiky do roku 2030+ [Strategy for Education Policy of the Czech Republic 2030+]*. Ministry of Education.
- Gillies, D. (2017). Human capital theory in education. In M. A. Peters (Ed.), *Encyclopedia of educational philosophy and theory* (pp. 1053-1057). Springer.
- Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. *Educational and Psychological Measurement*, 66(6), 930–944.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38-47.
- Holden, L., & Biddle, J. (2017). The introduction of human capital theory into education policy in the United States. *History of Political Economy*, 49(4), 537-574.

- Livingston, S. A. (2014). *Equating test scores (without IRT)*. Educational Testing Service.
- Lucas, R. E. (2015). Human capital and growth. *American Economic Review*, 105(5), 85-88.
- Orlando. M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24(1), 50–64.
- Raykov, T., & Marcoulides, G. A. (2016). On the relationship between classical test theory and item response theory: from one to the other and back. *Educational and Psychological Measurement*, 76(2), 325-338.
- Revelle, W. (2012). *An introduction to psychometric theory with applications in R*. Northwestern University.
- Rizopoulos, D. (2018). *Package 'ltm'*. [online]. Available from: <https://cran.r-project.org/web/packages/ltm/ltm.pdf>.
- Rupp, A. P. (2005). Maximum likelihood item response theory estimation. In D. C. Howell, & B. S. Everitt (Eds.), *Encyclopedia of Statistics in Behavioral Science*. John Wiley.
- Rusch, T., Lowry, P. B., Mair, P., & Treiblmaier, H. (2017). Breaking free from the limitations of classical test theory: Developing and measuring information systems scales using item response theory. *Information & Management*, 54(2), 189-203.
- Ryan, J., & Brockmann, F. (2009). *A practitioner's introduction to equating with primers on classical test theory and item response theory*. Council of Chief State School Officers.
- Thompson (2016). *Introduction to classical test theory with CITAS*. Assessment Systems Corporation.
- Thorpe, G. L., & Favia, A. (2012). *Data analysis using item response theory methodology: an introduction to selected programs and applications*. The University of Maine.
- Toland, M. D. (2014). Practical guide to conducting an item response theory analysis. *The Journal of Early Adolescence*, 34(1), 120-151.
- Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, 16(4), 8-14.
- Van der Linden, W. J. (2010). Item response theory. In P. Peterson, R. Tierney, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (pp. 81-88). Elsevier.
- Van Zile-Tamsen, C. (2017). Using Rasch analysis to inform rating scale development. *Research in Higher Education*, 58(8), 922-933.
- Voogt, J., & Roblin, N. P. (2012). A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *Journal of Curriculum Studies*, 44(3), 299-321.
- Willse, J. T. (2018). *Package 'CTT'*. [online]. Available from: <https://cran.r-project.org/web/packages/CTT/index.html>.
- Ziegler, M., & Hagemann, D. (2015). Testing the unidimensionality of items. *European Journal of Psychological Assessment*, 31(4), 231-237.