Application of the Contemporary Psychometrics for Assessing Economic Literacy

Elena Kardanova, Andrei Dementiev, Ilya Denisov, Irina Zueva, Denis Federiakin

Received in August 2023

Elena Kardanova — Candidate of Sciences (PhD) in Differential Equations, Dynamic Systems and Optimal Control; Scientific Supervisor, Centre for Psychometrics and Measurement in Education, Institute of Education, HSE University. Address: 16/10 Potapovskiy Lane, 101000 Moscow, Russian Federation. E-mail: ekardanova@hse.ru. ORCID: https://orcid.org/0000-0003-2280-1258 (corresponding author)

Andrei Dementiev — Leading Expert, Laboratory for Competency Modelling and Assessment in Higher Education, Institute of Education, HSE University. E-mail: adementiev@hse.ru. ORCID: https://orcid.org/0000-0001-6449-7659

Ilya Denisov — Analyst, Centre for Psychometrics and Measurement in Education, Institute of Education, HSE University. E-mail: idenisov@hse.ru. ORCID: https://orcid.org/0000-0002-8741-2141

Irina Zueva — Analyst, Laboratory for Competency Modelling and Assessment in Higher Education, Institute of Education, HSE University. E-mail: izueva@hse.ru

Denis Federiakin — Research Fellow, Department of Economic Education, Johannes Gutenberg University of Mainz, Mainz, Germany. E-mail: denis.federiakin@uni-mainz.de. ORCID: https://orcid.org/0000-0003-0993-5315

Abstract

Currently, new skills and various types of “new literacies” relevant to the modern world are becoming issues of growing importance. One of them is economic literacy; however, there are only few assessment instruments that fulfil the academic requirements for its assessment among university students. One of such internationally established instruments is the Test of Understanding in College Economics (TUCE), which is a popular tool in empirical studies of economic literacy in many countries around the world. Despite its advantages, the currently available version of the TUCE designed for American colleges back in 2006, is prone to cheating and provides limited opportunities for formative feedback.

The purpose of this paper is to present the Updated Test of Understanding in College Economics (U-TUCE). In developing the U-TUCE, we utilized the capabilities of contemporary psychometrics, which offer sufficient advances in overcoming all limitations of the original TUCE mentioned before. First, we present a revised theoretical framework of the U-TUCE, highlighting that the test measures different types of mastery of economic literacy. Second, we describe the approaches used for modifying the TUCE items and developing new items. A half of the original test items have been replaced or redesigned to reflect the economic context that has changed since 2006. Third, we utilize the logic of automatic item generation algorithms to increaseg the level of test protection against cheating. We made all changes in such a way as to maintain comparability with the previous versions of the TUCE test if necessary. Finally, the use of the Item Response Theory (IRT) is paired up with that of Cognitive Diagnostic Modeling (CDM) to ensure the quality of the U-TUCE and enhance its formative value. We show that IRT can be used to estimate the construct as a whole (which is of interest to researchers, administrators, and policy makers), while CDM provides information relating to each of the construct components, which are of interest to educational practitioners and students themselves. The results of the data analyses show that the test can be used for both purposes simultaneously.

Keywords economic literacy, psychometric modeling, Item Response Theory, cognitive diagnostic modeling, automated item generation, formative feedback

For citing Kardanova E.Yu., Dementiev A.V., Denisov I.S., Zueva I.O., Federiakin D.A. (2024) Application of the Contemporary Psychometrics for Assessing Economic Literacy. *Voprosy obrazovanoya / Educational Studies Moscow*, no 3 (1), pp. 45–66. https://doi.org/10.17323/vo-2024-17846

The modern labor market requires university graduates to have not only professional knowledge, but also generic competences, which are related to personal success in various professional and personal areas. Today, the issue of the new skills and various types of “new literacies”, such as digital, financial, or information, relevant to the modern world, has become crucial. One of them is economic literacy, which can be defined as the ability to recognize basic economic problems in everyday life and apply the methods and principles of economic analysis to solve them. Despite the importance of economic literacy, there are only few assessment instruments that fulfil the requirements for evaluating university students on it. One of the internationally established instruments, due to its well-developed theoretical framework and the fact that it complies with the modern standards of test development [AERA, APA, and NCME, 2014], is the Test of Understanding of College Economics (TUCE) [Walstad, Rebeck, 2008].

The TUCE test is designed to assess the knowledge of students studying the principles of economics from college entry level to that of a university degree. The test consists of two content areas: micro- and macroeconomics, each of which contains 30 dichotomous multiple-choice items. All items are said to measure three levels of cognitive comprehension of economics: recognition and understanding (where students are expected to simply demonstrate the basic familiarity with economic concepts), explicit application (where students are faced with a real-life situation that forces them to apply explicitly stated methods of economic analysis) and implicit application (where students need to decide which method of economic analysis to apply).

The access to the TUCE test is open, which makes it a popular tool in empirical studies of the quality of economic education in many countries around the world. In particular, the TUCE is the basis for the assessment instruments used for the international study of the quality of economic education of university students WiWiKom[1]. In 2020–2021, the Russian version of the WiWiKom test was developed in accordance with international standards [International Test Commission, 2017], including translating and adapting the instrument for the national educational context, psychometric analysis of the instrument, and piloting [Federiakin et al., 2022].

Along with advantages, the TUCE test has several drawbacks that limit its use. First, it was originally developed for American colleges, which restricts its application in an international context. Second, the current version of the test (TUCE-4) was developed in 2006 and no longer fully reflects the ever-changing landscape of economic skills. Third, although the test is presented in two versions, both of them are publicly available, which makes cheating possible. Finally, the test provides limited opportunities for formative feedback. Information on test scores may be sufficient for making some administrative decisions, but not for giving useful feedback to students.

Thus, despite the appeal of the existing Russian version of the TUCE test, it seems reasonable to improve the TUCE test while considering its limitations. The study was conducted at the National Research University Higher School of Economics in 2021–2022. As a result, the Updated Test of Understanding in College Economics (U-TUCE) was developed. The U-TUCE is a modification of the TUCE test and is intended for mass testing of students. In developing the U-TUCE, the capabilities of the contemporary psychometrics were utilized, which offer sufficient advances in overcoming all of the revealed limitations of the original TUCE.

The purpose of this paper is to demonstrate the possibilities of modern psychometrics for the development of the U-TUCE. Specifically, we (i) present a revised theoretical framework of the U-TUCE, (ii) describe the approaches used for modifying the TUCE items and developing new items based on the updated theoretical framework, (iii) describe the logic of automatic item generation algorithms enhancing the anti-cheating protection, (iv) utilize the Item Response Theory (IRT) and Cognitive Diagnostic Modeling (CDM) to ensure and enhance the quality of pilot studies of the U-TUCE.

1. Theoretical framework for assessing economic literacy

1.1. Definition of economic literacy

Economic literacy is the ability to identify the basic concepts and principles of economic functioning at the micro- and macrolevels and to apply methods of economic analysis to justify solutions to practical problems faced by households, firms, and the government. Economic literacy goes beyond declarative knowledge of basic economic terms, definitions, and concepts. Being a functional litera-

[1] https://www.wiwi–kompetenz.de/

cy, it implies, by definition, the ability to choose and justify optimal solutions in life situations, formulated from the perspective of the economic theory and described in everyday language. A high level of economic literacy implies the ability to find and apply the most appropriate methods of economic analysis to solving complex problems that require integral knowledge of micro- and macroeconomics. One of the critical parts of this construct, however, is the limited use of mathematical apparatus in the solution-search behavior.

In order to obtain a reliable, valid, and differentiated assessment of economic literacy, test specification should define the range of topics and depth of mastery of the subject area. The detailed specification of the U-TUCE, both cognition- and content-wise, is an attempt to specify the very notion of economic literacy as a universal competence measured at different levels.

1.2. Cognitive specification of U-TUCE and the facets of economic literacy

The TUCE is based on a three-level taxonomy, which is a projection of the modified Bloom's taxonomy [Walstad, Rebeck, 2008] onto the domain of economic knowledge. The cognitive specification of the TUCE test is based on a taxonomy consisting of three competence levels: recognition and understanding; explicit application; implicit application. The recognition and understanding level corresponds to a combination of the first two categories of Bloom's modified taxonomy — "remember and understand", the level of explicit application generally corresponds to the level "apply" and, finally, the level of implicit application corresponds to the higher levels "analyze" and "evaluate".

The TUCE was first developed in 1967 [Fels, 1967] and has undergone many changes since then, most of which occurred in the version of the test published in 2006 [Walstad et al., 2007]. Nevertheless, the three-level cognitive taxonomy has retained its ascending structure although the authors of the test highlight the very important point that the exact organization of the respondents' cognitive process while solving the items is not known [Walstad, Rebeck, 2008]. In particular, an implicit application question may turn out to be a recognition question, if, for example, the learning of the principles of economics was based on case studies and has not been forgotten by a respondent. In this regard, the task of developing a more detailed cognitive taxonomy of the test in relation to the educational content of a particular educational program is extremely relevant.

In developing the taxonomy for the U-TUCE, we utilize the fact that economic literacy is a composite construct, which we define as consisting of various combinations of three types of knowledge: declarative, procedural, and functional:

- Declarative Knowledge (DK) is the ability to recognize definitions and concepts identified within a basic glossary of economic terms. DK also includes the ability to make connections between theories and concepts and to delve into specific details, and is manifested primarily in an educational context.
- Procedural Knowledge (PK) is the ability to apply standard methods of economic analysis to solving problems formulated in explicit form. PK is aimed at using known algorithms, model inference, technical calculation procedures, logical reasoning, and deductive inferences specific to the economic field.
- Functional Knowledge (FK) is the ability to generalize and use practical experience of solving economic problems formulated in implicit form. FK is always contextual, closely related to practical activity and consists in the ability to find the most probable answer to a non-standard question as a result of inductive reasoning.

The three facets of economic literacy can substitute or complement each other. It is worth noting that this taxonomy does not require these types of knowledge to be nested parts of a hierarchy of cognitive skills. In practice, economic literacy as a composite competence can exist in various combinations of the three facets, resulting in different cognitive profiles. Depending on the educational and professional context, approaches to solving practical problems may differ, for example, between students and experts. Therefore, it is more appropriate to speak of the development of economic literacy in terms of a specific cognitive profile rather than the achieved "level" of knowledge.

For example, a combination of DK and PK allows solving standard problems formulated explicitly using basic glossary terms. A typical "student-analyst" is able to solve the task by applying textbook economic models and interpreting the results. At the same time, a typical "practitioner-analyst", answering a similar question, may use a different set of cognitive skills, formed as a generalization of practical experience and/or probabilistic approach in selecting alternative solutions. A deep mastery of terminology may not be required if the task is clearly stated in explicit terms. Another example of a specific competence profile is a combination of DK and FK. A typical "student-practitioner" is able to recognize complex concepts formulated implicitly as a description of a life situation. To solve the task, such a "student-practitioner" recalls a similar example and reproduces its solution in a similar way. A typical "expert", however, to solve the task, compares a concrete situation with a library of solved cases and selects the appropriate term from the problem situation. PK as application of standard analytical algorithmms is not required for them since the solution is selected from a set of similar cases.

1.3. Content specification of the U-TUCE test

In terms of content, the *U-TUCE* content covers six microeconomics topics and six basic-level macroeconomics ones (see Table 1).

Table 1. **Content areas of U-TUCE**

Content area	Topic	Number of items
Microeconomics	The Basic Economic Problem	4
	Markets and Price Determination	7
	Theories of the Firm	6
	Theory of Consumer Behavior	5
	Factor Markets	4
	The Role of Government in a Market Economy	4
Macroeconomics	Measuring Aggregate Economic Performance	4
	Aggregate Supply and Aggregate Demand	7
	Money and Financial Markets	5
	Monetary and Fiscal Policies	6
	Limitations of Macroeconomic Policies	3
	International Economics	5

The content structured in this way largely echoes the content specification of the TUCE, but we updated the content of the test for it to better reflect the modern economic theories. For example, the thematic structure of TUCE did not include the “Theory of Consumer Behavior”, which is necessary to test the level of economic literacy with regard to the principles of household management. Our analysis of modern educational programs revealed that this topic is included in many of them. This is what determined the need to identify this topic as a distinct content aspect of the U-TUCE.

The coverage of thematic sections proposed in the U-TUCE is neither exhaustive nor the only one possible. A certain selectivity and limitation to the basic topics is necessary due to the fact that economic literacy is supposed to be a generic competence essential to all students, including those not majoring in economics.

At the same time, in light of the content approach to economic literacy, the thematic division of the test should not be interpreted as an attempt to assess competences in the relevant topics separately from each other. These topics in the modern science are based on a common methodology and have many content overlaps. Thus, economic literacy as a functional literacy reflects the ability to identify a relevant area of knowledge for the analysis of a specific economic situation, which is a cross-thematic skill. This emphasizes the metacognitive nature of economic literacy as a generic competence. This is also reflected in the U-TUCE specification, where none of the items are flagged as measuring only one topic. The items are cross-classified in terms of two or three content topics and one type of knowledge, reflecting the entangled and intertwined relations between the competencies.

2. Theoretic approaches to developing the U-TUCE

2.1. Making changes to the TUCE

For the pilot study, the tasks originally included in the TUCE were translated and adapted for the Russian language in accordance with the International Test Commission (ITC) Guidelines for Translation and Adaptation of Tests [International Test Commission, 2017]. However, in the process of modifying the TUCE, about 50% of its items were changed or replaced with newer ones to ensure that the test content is up-to-date. This was carried out by experts in the field of economics — leading professors of the Department of Theoretical Economics of the Faculty of Economic Sciences at the National Research University "Higher School of Economics" (NRU HSE). To make changes to the tasks and develop the new ones, we used the socio-cognitive theory of educational measurement [Mislevy, 2018] and its practical implementation, the evidence-centered design [Mislevy, Riconscente, 2011]. This approach breaks down the test development process into steps for argumentized selection of behavioral indicators necessary for the desired type of claim about the test-takers and eliminates alternative explanations for test results. This allowed us to make precise theoretically justified changes to the test context, providing systematized validity evidence of the intended test results interpretation.

2.2. Automatic Item Generation

To protect the U-TUCE from cheating and ensure a constant supply of new items, we developed algorithms for Automatic Item Generation (AIG). AIG is a rapidly growing area of psychometric research, responding to the need in developing a large number of items while simultaneously reducing the cost of their creation [Gierl, Lai, 2016]. Often, the development of such algorithms relies on the formation of cognitive models of items, in which two types of item elements are distinguished [Irvine, Kyllonen, 2013]: radicals (are assumed to influence the psychometric properties of items) and incidentals (are assumed to have no influence over the psychometric properties of items). For a given item model, radicals remain constant, while incidentals vary in a free or constrained manner to produce different versions of items with similar psychometric properties. Incidentals often include details of item context, answer options from a pre-compiled database of options, and specific numbers (according to the defined constraints) in items requiring computations. This allows item cloning, with item psychometric properties preserved or changed to a very limited extent.

New items were developed jointly with experts in economics, who assessed the quality of items at each stage of the process and, if necessary, made corrections. Thus, the experts were involved in (i) describing the cognitive processes expected from students for solving the item, (ii) identifying radicals and incidentals on, (iii) formulating the range of variation of incidentals, (iv) evaluating the resulting task variants, and (v) analyzing the expected psychometric properties of the new task variants. Additionally, 10 cognitive labs were conducted with students from the target test population in the form of deep interviews and think-aloud protocols to infer the reasoning process utilized by them while solving the items. In total, 30 out of 60 U-TUCE items were processed in this way (see Table 2).

Table 2. **Distribution of added/reworked items among topics**

Content topic	Number of items
The Basic Economic Problem	4
Markets and Price Determination	3
Theories of the Firm	2
Theory of Consumer Behavior	3
Factor Markets	3
The Role of Government in a Market Economy	1
Measuring Aggregate Economic Performance	3
Aggregate Supply and Aggregate Demand	1
Money and Financial Markets	4
Monetary and Fiscal policies	3
International Economics	3

Notably, despite the substantial number of changes we proposed to the TUCE when creating the U-TUCE, we kept them partially limited in order to preserve score comparability between the two instruments in the case of, for example, an international study. To do this, we included 30 items from the TUCE in the U-TUCE without changes to make sure that they can be used as anchor items for comparing scores.

2.3. The psychometric quality analysis of the U-TUCE and feedback development

To analyze the psychometric quality of the test and feedback development, we used the latent variable modelling paradigm, which assumes that observed item responses are manifestations of students' latent traits. We used Item Response Theory (IRT) [Van der Linden, 2018] and Cognitive Diagnostic Models (CDM) [Von Davier, Lee, 2019]. IRT is based upon the assumption that the underlying economic literacy reflected in the U-TUCE items is a continuous unbounded interval characteristic. The CDM, on the other hand, introduces a latent classification of subjects based on discrete latent cognitive states defined in terms of mastering/non-mastering of the sub-competencies reflected in the test specification. This is the psychometric basis for formative feedback that provides test takers with detailed information about which subcompetencies they should develop to further improve their economic literacy.

To avoid possible confusion between different forms of results, we use a multi-model strategy for psychometric analyses [Federiakin, Kardanova, 2020; Kanonire et al., 2020]. This strategy suggests that as long as different psychometric models fit the data well, it is acceptable to use their results for different purposes, provided they are not presented simultaneously to the same users in a contradictory or confusing way. Thus, we avoid any contradictions in the use of IRT and CDM results by suggesting the use of formative feedback by educational practitioners and test takers themselves, and summative feedback in managerial decision making and research.

2.4. The description of the final version of the U-TUCE

Both the TUCE and the U-TUCE, contain 60 dichotomous multiple-choice items each. The multiple-choice format was chosen for the U-TUCE since this item format is very well-known to the respondents and allows for preservation of comparability across TUCE and U-TUCE. The items are distributed equally between micro- and macroeconomic. The item order was determined randomly for each respondent, and there was no explicit indication that the item belongs to a particular topic. The order of the response options for each respondent was also randomized.

3. Pilot studies of the U-TUCE and analysis of its psychometric quality

Piloting of the U-TUCE was conducted in two steps. The first pilot study used a small sample of students of the National Research University Higher School of Economics. The purpose of the first pilot was primary analysis of the psychometric quality of the U-TUCE (before cloning the tasks using AIG algorithms). The second, more extensive, pilot used a large sample of students from five universities. It aimed to investigate the psychometric quality of the U-TUCE in detail. Besides IRT-based analysis, the study utilized analysis of the U-TUCE in CDM to ensure the feasibility of formative feedback.

3.1. The first piloting

3.1.1. Sample and testing procedure

The sample of the first pilot study consisted of 502 Higher School of Economics students from different majors. The data was analyzed in IRT. The test was administered in a computer-based format in the autumn of 2021. Testing took place on personal computers in the university’s learning management system. The testing time was limited to 90 minutes. Each student was assigned the same version of the test. All items missed by respondents were coded as incorrect answers for analysis purposes.

3.1.2. Results of the analysis

The first version of the test was analyzed using the Rasch dichotomous model with Winsteps software[2] [Van der Linden, 2018].

Below is a summary of the results of the conducted psychometric analysis:

1) The U-TUCE is essentially unidimensional, which means that the test measures a single construct;
2) All test items have good psychometric properties and can be used in the test;
3) Measurement reliability is adequate for individual assessment (Cronbach's alpha = 0.92).

An important index characterizing the quality of measurement is the person separation index in the Rasch modelling paradigm, which in this case was equal to 3.15. This index allows us to determine a number of statistically distinguishable (allowing for the measurement error) groups of test takers into which the entire sample of participants can be divided. The index value of 3.15 means that the whole sample of the pilot study participants can be divided into at least four groups. In other words, the test allows us to differentiate students according to their level of the measured construct.

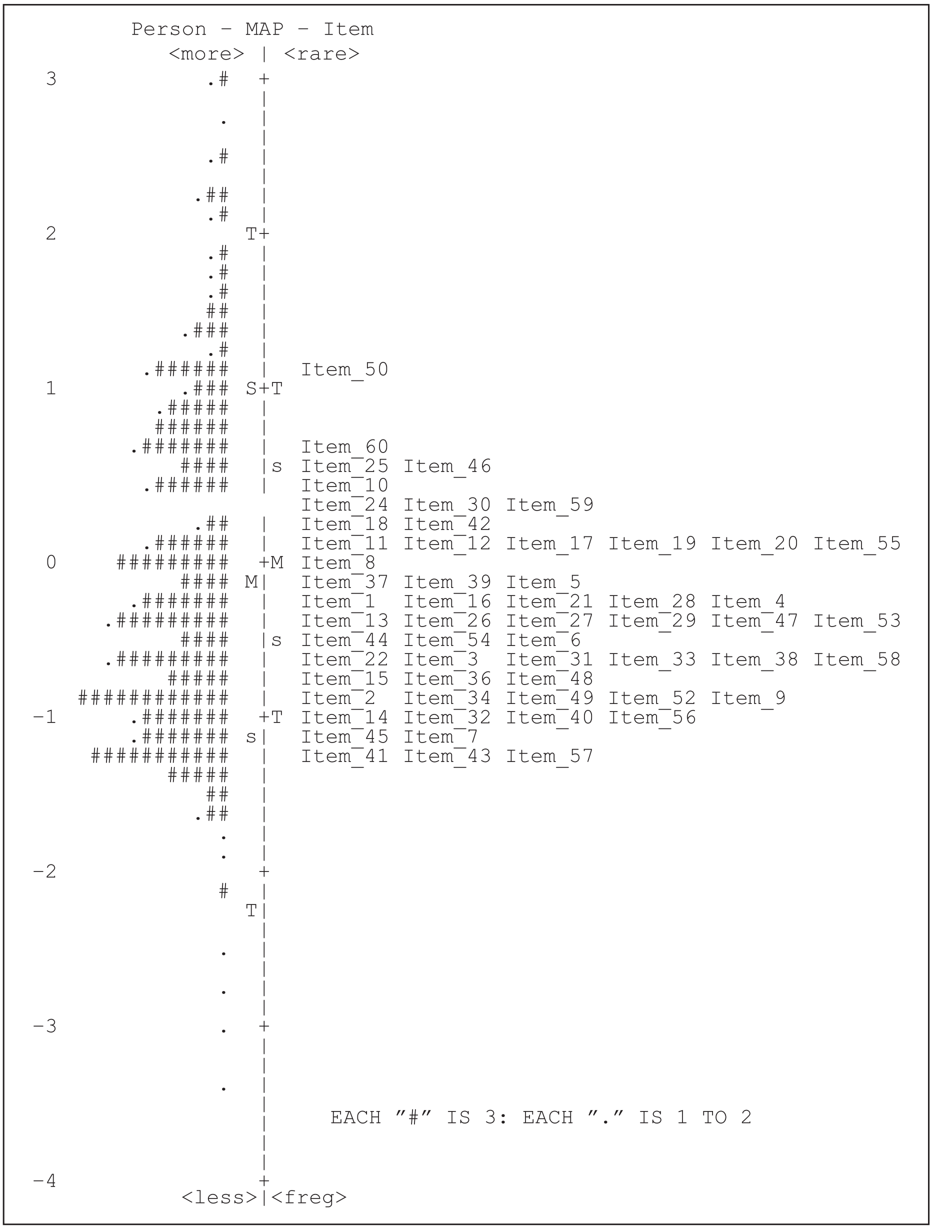
The variable map in Figure 1 shows the relative distribution of items and students in a common metric. The students and test items are on the left and right sides of the map, respectively. The more difficult items and higher-performing students are located in the upper part of the map, while the easier items and lower-performing students are placed in the lower one. The distribution of students is wide and represents a good differentiation between the higher and lower scoring students for measurement purposes. Moreover, the student sample is well located with respect to the items, the mean ability level is 0.18 logits below the mean of item difficulty, and the variance of the student ability measures is quite large (1.02 logits).

It is worth noting that the unidimensionality of the test is an expected result [Federiakin et al., 2022]. Previously, economic literacy was defined, assessed, and promoted as an integral learning outcome. Correspondingly, at the content level, it is conceptualized in didactic terms as a unidimensional construct.

Thus, based on the initial analysis results, we conclude that the U-TUCE is a reliable measurement instrument that can be used for individual level assessment.

[2] https://www.winsteps.com/winsteps.htm

Figure 1. **Variable map for U-TUCE test**



```
          Person - MAP - Item
             <more> | <rare>
   3            .#  +
                    |
                 .  |
                    |
                .#  |
                    |
               .##  |
                .#  |
   2               T+
                .#  |
                .#  |
                .#  |
                ##  |
              .###  |
                .#  |
           .######  |  Item_50
   1          .### S+T
            .#####  |
            ######  |
          .#######  |  Item_60
              ####  |s Item_25 Item_46
           .######  |  Item_10
                       Item_24 Item_30 Item_59
               .##  |  Item_18 Item_42
           .######  |  Item_11 Item_12 Item_17 Item_19 Item_20 Item_55
   0     #########  +M Item_8
              #### M|  Item_37 Item_39 Item_5
          .#######  |  Item_1  Item_16 Item_21 Item_28 Item_4
        .#########  |  Item_13 Item_26 Item_27 Item_29 Item_47 Item_53
              ####  |s Item_44 Item_54 Item_6
        .#########  |  Item_22 Item_3  Item_31 Item_33 Item_38 Item_58
             #####  |  Item_15 Item_36 Item_48
      ############  |  Item_2  Item_34 Item_49 Item_52 Item_9
  -1      .#######  +T Item_14 Item_32 Item_40 Item_56
          .####### s|  Item_45 Item_7
       ###########  |  Item_41 Item_43 Item_57
             #####  |
                ##  |
               .##  |
                 .  |
                 .  |
  -2                +
                 #  |
                   T|
                    |
                 .  |
                    |
                 .  |
                    |
  -3             .  +
                    |
                    |
                 .  |
                    |
                    |    EACH "#" IS 3: EACH "." IS 1 TO 2
                    |
                    |
  -4                +
              <less>|<freg>
```

3.2. The second piloting

3.2.1. Sample and instrument

The study involved 4430 1st–4th year students at five Russian universities: three federal universities (1,817 students in total), one classical university (271 students), and one research university (NRU HSE; 2,334 students). The sample of NRU HSE consisted of students not majoring in economics, but taking the course "Economics" as a generic discipline. The course included 30 hours of class contact hours and 122 hours of homework. The samples of the other universities were formed by the participant universities themselves.

For the larger pilot study of the U-TUCE, five test forms were developed, differing in the included item clones, but overlapping in the anchor items. The item-writers from the faculty of Economic Sciences at NRU HSE examined all five U-TUCE forms and confirmed their content quality.

3.2.2. Testing procedure

The test was administered in a computer-based format in the spring of 2022. Testing took place in the computer labs of the universities under the supervision of the staff members of the participating universities. The testing time was limited to 90 minutes. Each student was randomly assigned one of five test forms in order to create equivalent subsamples and to avoid the effects of differences in ability between the subsamples [Hinkelmann, Kempthorne, 2007]. All tasks missed by respondents were coded as incorrect answers for analysis purposes.

3.2.3. Methodology of data analysis

Analysis of the psychometric quality of the U-TUCE was conducted in several steps. Firstly, the degree to which different versions of the U-TUCE are parallel was analyzed. We analyzed this via Differential Item Functioning (DIF) tests [Holland, Wainer, 1993] on the format-specific items with the scales equated via anchor items. If the items did not exhibit any DIF, we anchored their difficulty to the same values across all the respective formats. The main purpose of this stage was to verify the functioning of item clones.

Secondly, the dimensionality of the test was investigated using IRT to ensure that the test allows extraction of one common factor of economic literacy. To this end, the AIC [Akaike, 1974] and BIC [Schwarz, 1978] information criteria of univariate, multivariate and higher-order Rasch models and the likelihood ratio test [Paek, Wilson, 2011] were compared.

Thirdly, the psychometric properties of individual test items were investigated to confirm their quality. Here, item fit statistics were studied in detail to make sure that item scores are well predicted by latent ability.

Finally, the appropriateness of Cognitive Diagnostic Models for formative feedback was studied.

Notably, while the first three steps are somewhat routine in psychometric research and practice and well-described in the scientific literature, the last one, the use of CDMs for the formative feedback, is relatively new, so we will focus on it in detail.

3.2.4. Cognitive Diagnostic Models

Cognitive Diagnostic Models (CDMs) [Rupp et al., 2010] provide detailed feedback on a person's cognitive profile in terms of skills, sub-competences, and cognitive operations that must be mastered in order to successfully solve test items. This cognitive profile of a person consists of zeros and ones, where a one indicates acquisition of a sub-competence and a zero indicates its absence. Thus, an individual cognitive profile is a vector of zeros and ones encoding the interpretation of the latent class to which the respondent belongs. This provides diagnostic information for the student and facilitates aggregation of information at the sample or subsample level.

A crucial component of the CDM methodology is the so-called Q-matrix [Tatsuoka, 1983], which relates test items to a set of theoretically defined sub-competencies. The Q-matrix encodes the information on whether or not a particular sub-competency is needed to solve a particular item in the form of ones and zeros. The Q-matrix is traditionally compiled by experts in the respective subject matter and test developers. It is worth noting that compiling a Q-matrix can be a non-trivial research task in itself [Bley, 2017; Tjoe, de la Torre, 2013; Köhn, Chiu, 2018; de la Torre, Minchen, 2014; Roussos et al., 2007b].

There is a huge variety of CDMs [Liu et al., 2023] that can be classified using very different criteria. For example, there are models for polytomous [Chen, de la Torre, 2018; Tu et al., 2010] and dichotomous [de la Torre, 2011; Henson et al., 2009] data. There are also models allowing skills to have dichotomous [Henson et al., 2009] or polytomous [Helm et al., 2022; Chen, de la Torre, 2013] scores. There are CDMs that describe local task dependence [Zhan et al., 2015] or models that take response time into account [Zhan et al., 2018]. However, one of the most important criteria for model classification is the assumption of compensatory or non-compensatory nature of the sub-competencies. If, when solving an item, mastery of some sub-competencies can compensate for the absence of others, such models are called compensatory [Templin, Henson, 2006; de la Torre, 2011]. An example of such an item is a language spelling task in which extensive vocabulary can compensate for ignorance of certain spelling rules. If, however, a model assumes that one must possess all the sub-competencies involved in an item in order to solve it, such a model is called a non-compensatory model [Tatsuoka, 1983; Junker, Sijtsma, 2001; Hartz, 2002]. An example of such an item is a mathematical problem requiring several consecutive calculations, where the result of the previous calculation is supplied as a given value into the next one. In this case, a mistake in one of the actions will propagate the error along the calculation chain, and the presence of any sub-competence cannot compensate for the absence of any other.

At the same time, many models are special cases of more general frameworks [de la Torre, 2011; Henson et al., 2009; von Davier, Lee, 2019]. By imposing special constraints on the parameters of these models, many special models can be obtained. This is an important aspect of studies on CDM properties. The development of frameworks allows one to create generalized software estimating parameters of entire groups of models. This enables comparison of the model fit of competing models since they can be estimated using the same algorithms [Deonovic et al., 2019].

3.2.5. Selection of the CDMs for the U-TUCE

When modelling the U-TUCE, individual topics of the test (12 in total) and types of knowledge from its taxonomy (3) were regarded as sub-competences. Eight cognitive diagnostic models were considered for the U-TUCE, each described in terms of three dichotomous criteria ($2^3 = 8$). The first criterion is the model structure. Here, Compensatory and Non-Compensatory Reparametrized Unified Models (C-RUM [Rupp et al., 2010] and NC-RUM [Roussos et al., 2007a]) were used. These models assume that acquisition of every additional sub-competence required by an item leads to a higher probability of the correct answer. Moreover, subcompetences differ in their influence on the probability of the correct response, and items differ in their sensitivity to each sub-competence. However, while in C-RUM a lack of one competence can be compensated for by acquisition of another one, in NC-RUM there is no such compensation. We did not consider other cognitive diagnostic models [e.g., Junker, Sijtsma, 2001] because they are either (1) too limiting, or (2) do not fit the theoretical assumptions about the response process, or (3) are special cases of other CDM frameworks that cannot be compared directly with models from the G-DINA framework due to different parameter estimation algorithms.

The second criterion for model description is the structure of the Q-matrix that was supplied to the model. We used nested and non-nested Q-matrices that define how the knowledge types from the U-TUCE taxonomy relate to each other. Tables 3 and 4 show the exemplary structure of non-nested and nested Q-matrices used for the U-TUCE test calibration in CDM, respectively.

Table 3. **Example structure of a non-nested Q-matrix**

Item knowledge type	Knowledge type		
	DK	PK	FK
DK	1	0	0
PK	0	1	0
FK	0	0	1

Table 4. **Example structure of a nested Q-matrix**

Item knowledge type	Knowledge type		
	DK	PK	FK
DK	1	0	0
PK	1	1	0
FK	1	1	1

Thus, in the case of a nested Q-matrix, all items load the DK type. Also, in the case of the nested Q-matrix, the sub-competences of the knowledge types are completely separate: each type is the "added value" of its cognitive state, and in order to solve the tasks for the "higher" types, it is necessary to have mastered all the lower types of knowledge as well. The non-nested Q-matrix describes each "higher" type of knowledge as a composite of "lower" ones and its own "added value".

The third basis for classification is whether the latent space of person parameters was restricted or not [Ma et al., 2023]. If knowledge types indeed constitute a hierarchy, only the cognitive profiles shown in the hierarchical part of Table 5 are possible, and the non-hierarchical ones are degenerate. Otherwise, the non-hierarchical cognitive profiles are possible, and the models with non-restricted person parameter space are going to fit significantly better than the model with a restricted person parameter space.

Table 5. **Full list of possible sub-competence profiles**

Cognitive profile types	Possible cognitive profile	DK	PK	FK
Hierarchical	1	0	0	0
	2	1	0	0
	3	1	1	0
	4	1	1	1
Non- hierarchical	5	0	1	0
	6	0	1	1
	7	0	0	1
	8	1	0	1

Thus, eight models were selected to analyze the U-TUCE (see Table 5). The global model fit of these models with the data was analyzed using relative and absolute fit indices. Among the indices of relative fit, we used AIC and BIC. The other two indices were used to analyze the absolute model fit: the item Root Mean Square Error of Approximation (RMSEA) [Steiger, 1990], together with its standard deviation) and the Standardized Root Mean Residual (SRMR

[Hu, Bentler, 1999]). In all cases, the lower value of an index indicates a better model fit.

3.2.6. The results of the U-TUCE analysis in the IRT

The results show that the U-TUCE results are mostly comparable across five automatically generated U-TUCE forms. For the sake of brevity, we do not provide an extensive description of these analyses. However, we note that these results were obtained via Differential Item Functioning analysis methods across groups defined by different U-TUCE forms [Holland, Wainer, 1993]. Overall, 90% of items were comparable across every pair of the U-TUCE forms. However, for the purpose of the analysis presented here, the incomparable automatically generated items were ignored and treated as the same item.

In general, the results showed that the test scores are of high quality and can be used and reported on individual level. The dichotomous Rasch model from Item Response Theory [Rasch, 1993] was used to obtain these results. Item fit analyses (the value of In-Fit item statistics [Wright, Stone, 1979] varied from 0.88 to 1.17) revealed that the Rasch model fits the data well. The EAP-reliability [Adams, 2005] was equal to 0.821. This implies that we can hold the results to be highly trustworthy.

3.2.7. The results of the U-TUCE analysis in the CDM

The results of comparing the global model fit are presented in Table 6.

Table 6. **Comparison of global model fit**

Model	Q-matrix	Person space	AIC	BIC	Item RMSEA (SD)	SRMR
Comp.	Nested	Restricted	201598.9	297597.8	0.269 (0.094)	0.027
		Non-Restricted	170936.1	173313.7	0.070 (0.019)	0.024
	Non-nested	Restricted	202656.5	298309.9	0.215 (0.063)	0.030
		Non-Restricted	171152.5	173184.7	0.077 (0.020)	0.026
Non-comp.	Nested	Restricted	202113.2	298112.0	0.242 (0.080)	0.029
		Non-Restricted	171012.3	173389.9	0.083 (0.018)	0.025
	Non-nested	Restricted	202496.2	298149.2	0.260 (0.095)	0.031
		Non-Restricted	171512.9	173545.9	0.068 (0.019)	0.028

Table 6 shows that AIC and BIC are in conflict. While AIC indicates that the second model (a compensatory one with the nested Q-matrix and unrestricted person parameter space) is a better fit to the data, BIC highlights the fourth model (a compensatory one with the non-nested Q-matrix and unrestricted person parameter space). We focus on AIC for model selection, as there is evidence that BIC oversimplifies the data-generating model in general in statistics [Evans, 2019] and in IRT in particular [Robitzsch, 2022]. Moreover, the absolute model fit indices demonstrate that the second model does fit the data better than the fourth one (albeit statistically insignificant in the case of RMSEA). Thus, for the further use, we select the second model.

Accordingly, we conclude that to solve U-TUCE items correctly, one must have at least one sub-competence of those required in an item (although the presence of additional sub-competences additionally increases the probability of solving it correctly). The nested Q-matrix means that the "lower" types of knowledge are required to solve the "higher" types of knowledge items. In other words, the estimates from CDM types of knowledge do not constitute any type of amalgamation of "lower" and "higher" types, and "lower" levels are controlled for when the "higher" ones are estimated. The unrestricted person parameter space means that knowledge types do not actually form a hierarchy — that is, it is possible to have, for example, the FK without DK, which is consistent with the definition of economic literacy. As follows from our analysis of the global model fit, the selected model fits data well. However, it is worth noting that all models with an unrestricted person parameter space fit the data better than their restricted counterparts, which means that the hierarchy of knowledge types is too restrictive assumption for the U-TUCE.

We used indices of accuracy, consistency [Johnson & Sinharay, 2018] and EAP-trustworthiness of classification to analyze the reliability of formative feedback. EAP-trustworthiness was estimated as the average individual probability of a person's belonging to the class into which he/she was classified by the CDM:

$$rel(EAP_{pk}) = \begin{cases} EAP_{pk}, \text{ if } EAP_{pk} \geq 0.5 \\ 1 - EAP_{pk}, \text{ if } EAP_{pk} < 0.5 \end{cases},$$

where $rel\,(EAP_{pk})$ is the individual reliability of respondent p classification according to sub-competence k, EAP_{pk} — the EAP-estimate of the probability that respondent p has mastered sub-competence k. Table 7 summarizes the reliability of the classification.

As can be seen from Table 7, all reliability estimates are very high. Note that because each item measured more than one topic and at least one type of knowledge, the total sum of the items in the corresponding column in Table 7 amounts to a number greater than 60 since each item was counted more than once. Overall, the high classification reliability scores (Table 7) together with the good model fit (Table 6) imply a high degree of confidence in the modelling results and allow the classification results to be used at the individual level to provide formative feedback.

Table 7. **Reliability of classification from the best fitting CDM**

Area	Topic	Number of items	Reliability of classification across entire sample		
			Accuracy	Consistency	EAP-trustworthiness
Microeconomics	The Basic Economic Problem	4	0.942	0.922	0.944
	Markets and Price Determination	7	0.938	0.908	0.942
	Theories of the Firm	6	0.894	0.839	0.899
	Theory of Consumer Behavior	5	0.825	0.788	0.831
	Factor Markets	4	0.750	0.756	0.758
	The Role of Government in a Market Economy	4	0.805	0.723	0.817
Macroeconomics	Measuring Aggregate Economic Performance	4	0.969	0.954	0.972
	Aggregate Supply and Aggregate Demand	7	0.928	0.896	0.931
	Money and Financial Markets	5	0.907	0.844	0.915
	Monetary and Fiscal Policies	6	0.955	0.937	0.957
	Limitations of Macroeconomic Policies	3	0.914	0.853	0.921
	International Economics	4	0.908	0.876	0.911
Knowledge types					
DK		55	0.924	0.867	0.931
PK		43	0.923	0.866	0.930
FK		27	0.992	0.986	0.993

4. Conclusion

The demands of the labor market and the rapidly changing socio-economic conditions of the modern world require new types of literacies from university graduates to remain successful in the highly competitive labor market. One of such types of literacy is the economic literacy. Modern employers are increasingly interested in employees who are able to successfully solve economic problems in everyday life. This does not go unnoticed by stakeholders in education. However, this creates a new type of demand within the education system itself: apart from special courses and educational programs that develop economic literacy, it requires assessment instruments that can measure this construct in an effective, valid, and reliable manner.

For the purposes of this study, we define economic literacy as a composite construct reflecting the ability of higher education students to notice basic concepts and principles of economic functioning at the micro- and macro-level and apply methods of economic analysis to find solutions to practical problems faced by households, firms, and the state. To measure this construct in line with this definition, we propose to use the Test of Understanding in College Economics (U-TUCE) based on the well-known TUCE test aimed at measuring economic literacy. However, the latest version of the TUCE was proposed in 2006, which identifies some problems related to its application. Among them are the following issues: (1) some of the topics covered are outdated, (2) the test is publicly available, which significantly lowers its security, (3) the test does not provide the formative feedback for the practitioners and the test-takers, which limits its use to mainly academic and administrative purposes.

Correspondingly, the changes we have made to the TUCE to develop the U-TUCE define the content of this paper. We have modified the theoretical framework of the test, highlighting that the test we propose measures different types of mastery of economic literacy. Besides, we have updated the content of the test by replacing several topics with more relevant ones. We have also replaced or redesigned 50% of the original test items to reflect the economic context that has changed since 2006. Furthermore, we developed algorithms for automatic item generation, which allow us to create new versions of the test, while ensuring meaningful comparability of the new version of the U-TUCE with older versions of the TUCE and increasing the level of protection against cheating. In doing so, all changes to the test are made in such a way as to maintain comparability with the previous versions of the TUCE test if necessary. Further, we conducted two approbations to ensure that the developed U-TUCE functions as expected in accordance with the international testing standards.

To analyze the test data, we adopted a multimodal data analysis strategy, where different models are applied to produce types of results. Particularly, we used both kinds of information: about the components of the global construct, which is of interest to educational practitioners and students themselves, and about the global construct as a whole, which is of interest to researchers, administrators, and policy makers. The results of this data analysis strategy show that the test can be employed for both purposes simultaneously. The IRT results on the global construct, and CDM results on each of the construct components provide high quality information to the respective groups of test results users. Further research, however, is required to improve the quality of the feedback for each user group, particularly, to formulate specific phrases and wording that will help users more easily understand the content of the responses and make more informed decisions.

This study has a number of limitations, which may be the subject of future research. Thus, we do not present the results and conclusions from the comparability analysis of automatically generated item variants. Our analyses have shown that not all item models generate comparable item variants. In particular, some of them generate variants whose difficulty is statistically significantly different from the difficulty of the original item and that of the remaining variants. For the purposes of the analyses presented in this paper, we ignored such cases. However, a more correct methodology for analyzing such data would be to use these items as if they were presented to only one subsample so that they do not participate in comparative analysis of the results between groups, being analyzed only within the corresponding subsample. This also determines why we do not talk about comparison results in this paper. Ignoring non-comparable items can lead to the distortion of the assessment results, reducing the validity of their interpretation. On the other hand, adding these results to the paper would dilute its focus and increase its scope beyond all reasonable limits. Also, we do not consider “external” validity evidence for these test results, for example, the correlation of its scores (calculated by different methods) with, for example, students’ grades. This could strengthen the validity of the U-TUCE.

Acknowledgements

The article was prepared in the framework of a research grant funded by the Ministry of Science and Higher Education of the Russian Federation (grant ID: 075-15-2022-325).

References

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

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

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

Bley S. (2017) Developing and Validating a Technology-Based Diagnostic Assessment Using the Evidence-Centered Game Design Approach: An Example of Intrapreneurship Competence. *Empirical Research in Vocational Education and Training*, vol. 9, no 1, Article no 6. https:// doi. org/ 10. 1186/ s40461- 017- 0049-0

Chen J.S., de la Torre J. (2013) A General Cognitive Diagnosis Model for Expert-Defined Polytomous Attributes. *Applied Psychological Measurement*, vol. 37, no 6, pp. 419–437. http://dx.doi.org/10.1177/0146621613479818

Chen J.S., de la Torre J. (2018) Introducing the General Polytomous Diagnosis Modeling Framework. *Frontiers in Psychology*, vol. 9, Article no 1474. http://dx.doi.org/10.3389/fpsyg.2018.01474

De la Torre J. (2011) The Generalized DINA Model Framework. *Psychometrika*, vol. 76, March, pp. 179–199. https://doi.org/10.1007/s11336-011-9207-7

De la Torre J., Minchen N. (2014) Cognitively Diagnostic Assessments and the Cognitive Diagnosis Model Framework. *Psicología Educativa*, vol. 20, no 2, pp. 89–97. http://dx.doi.org/10.1016/j.pse.2014.11.001

Deonovic B., Chopade P., Yudelson M., de la Torre J., von Davier A.A. (2019) Application of Cognitive Diagnostic Models to Learning and Assessment Systems. *Handbook of Diagnostic Classification Models* (eds M. von Davier, Y.-S. Lee), Cham: Springer, pp. 437–460. https://doi.org/10.1007/978-3-030-05584-4

Evans N.J. (2019) Assessing the Practical Differences between Model Selection Methods in Inferences about Choice Response Time Tasks. *Psychonomic Bulletin & Review*, vol. 26, no 4, Article no 10701098.

Federiakin D., Kardanova E. (2020) Psychometrical Modeling of Components of Composite Constructs: Recycling Data Can Be Useful. *Informatization of Education and E–learning Methodology: Digital Technologies in Education* (eds M. Noskov, A. Semenov, S. Grigoriev), Krasnoyarsk, pp. 157–171.

Federiakin D., Zlatkin-Troitschanskaia O., Kardanova E., Kühling-Thees C., Reichert-Schlax J., Koreshnikova Y. (2022) Cross-National Structure of Economic Competence: Insights from a German and Russian Assessment. *Research in Comparative and International Education*, vol. 17, no 2, pp. 225–245. http://dx.doi.org/10.1177/17454999211061243

Fels R. (1967) A New Test of Understanding in College Economics. *American Economic Review*, vol. 57, no 2, pp. 660–666.

Gierl M.J., Lai H. (2016) Automatic Item Generation. *Handbook of Test Development* (eds S. Lane, M.R. Raymond, T.M. Haladyna), New York, NY: Routledge, pp. 410–429.

Hartz S.M. (2002) *A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities: Blending Theory with Practicality*. Champaign, IL: University of Illinois at Urbana-Champaign.

Helm C., Warwas J., Schirmer H. (2022) Cognitive Diagnosis Models of Students' Skill Profiles as a Basis for Adaptive Teaching: An Example from Introductory Accounting Classes. *Empirical Research in Vocational Education and Training,* vol. 14, no 1, Article no 9. http://dx.doi.org/10.1186/s40461-022-00137-3

Henson R.A., Templin J.L., Willse J.T. (2009) Defining a Family of Cognitive Diagnosis Models Using Log-Linear Models with Latent Defining a Family of Cognitive Diagnosis Models Using Log-Linear Models with Latent Variables. *Psychometrika*, vol. 74, November, pp. 191–210. https://doi.org/10.1007/s11336-008-9089-5

Hinkelmann K., Kempthorne O. (2007) *Design and Analysis of Experiments. Vol. 1: Introduction to Experimental Design*. New York, NY: John Wiley & Sons.

Holland P.W., Wainer H. (eds) (1993) *Differential Item Functioning*. Hillsdale: Lawrence Erlbaum Associates.

Hu L., Bentler P.M. (1999) Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 6, no 1, pp. 1–55.

International Test Commission (2017) *The ITC Guidelines for Translating and Adapting Tests* (Second edition). Available at: https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf (accessed 20 August 2024).

Irvine S.H., Kyllonen P.C. (2013) *Item Generation for Test Development*. New York, NY: Routledge.

Johnson M.S., Sinharay S. (2018) Measures of Agreement to Assess Attribute-Level Classification Accuracy and Consistency for Cognitive Diagnostic Assessments. *Journal of Educational Measurement*, vol. 45, no 4, pp. 635–664. http://dx.doi.org/10.1111/jedm.12196

Junker B.W., Sijtsma K. (2001) Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory. *Applied Psychological Measurement*, vol. 25, no 3, pp. 258–272. http://dx.doi.org/10.1177/01466210122032064

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

Köhn H.-F., Chiu C.-Y. (2018) How to Build a Complete q-Matrix for a Cognitively Diagnostic Test. *Journal of Classification*, vol. 35, no 2, pp. 273–299. https://doi.org/10.1007/s00357-018-9255-0

Liu Y., Zhang T., Wang X., Yu G., Li T. (2023) New Development of Cognitive Diagnosis Models. *Frontiers of Computer Science*, vol. 17, no 1, Article no 171604. http://dx.doi.org/10.1007/s11704-022-1128-3

Ma C., Ouyang J., Xu G. (2023) Learning Latent and Hierarchical Structures in Cognitive Diagnosis Models. *Psychometrika*, vol. 88, no 1, pp. 175–207.

Mislevy R.J. (2018) *Sociocognitive Foundations of Educational Measurement*. New York, NY: Routledge.

Mislevy R.J., Riconscente M.M. (2011) Evidence-Centered Assessment Design. *Handbook of Test Development* (eds S. Lane, M.R. Raymond, T.M. Haladyna), New York, NY: Routledge, pp. 75–104.

Paek I., Wilson M. (2011) 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*, vol. 71, no 6, pp. 1023–1046. http://dx.doi.org/10.1177/0013164411400734

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

Robitzsch A. (2022) Four-Parameter Guessing Model and Related Item Response Models. *Mathematical and Computational Applications*, vol. 27, no 6, Article no 95. http://dx.doi.org/10.3390/mca27060095

Roussos L.A., DiBello L.V., Stout W. et al. (2007a) The Fusion Model Skills Diagnosis System. *Cognitive Diagnostic Assessment for Education: Theory and Applications* (eds J.P. Leighton, M. Gierl), Cambridge; New York: Cambridge University, pp. 275–318.

Roussos L.A., Templin J.L., Henson R.A. (2007b) Skills Diagnosis Using IRT-Based Latent Class Models. *Journal of Educational Measurement*, vol. 44, no 4, pp. 293–311. http://dx.doi.org/10.1111/j.1745-3984.2007.00040.x

Rupp A.A., Templin J., Henson R.A. (2010) *Diagnostic Measurement: Theory, Methods, and Applications Methodology in the Social Sciences*. New York, NY: Guilford.

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

Steiger J.H. (1990) Structural Model Evaluation and Modification: An Interval Estimation Approach. *Multivariate Behavioral Research*, vol. 25, no 2, pp. 173–180. http://dx.doi.org/10.1207/s15327906mbr2502_4

Tatsuoka K.K. (1983) Rule Space: An Approach for Dealing with Misconceptions Based on Item Response Theory. *Journal of Educational Measurement*, vol. 20, no 4, pp. 345–354. http://dx.doi.org/10.1111/J.1745-3984.1983.TB00212.X

Templin J.L., Henson R.A. (2006) Measurement of Psychological Disorders Using Cognitive Diagnosis Models. *Psychological Methods*, vol. 11, no 3, pp. 287–305. http://dx.doi.org/10.1037/1082-989X.11.3.287

Tjoe H., de la Torre J. (2013) Designing Cognitively-Based Proportional Reasoning Problems as an Application of Modern Psychological Measurement Models. *Journal of Mathematics Education*, vol. 6, no 2, pp. 17–26.

Tu D.B., Cai Y., Dai H.Q., Ding S.L. (2010) A Polytomous Cognitive Diagnosis Model: P-DINA Model. *Acta Psychologica Sinica*, vol. 42, no 10, pp. 1011–1020. http://dx.doi.org/10.3724/SP.J.1041.2010.01011

Van der Linden W.J. (ed.) (2018) *Handbook of Item Response Theory*. Three volume set. New York, NY: Chapman and Hall/CRC. https://doi.org/10.1201/9781315119144

Von Davier M., Lee Y.S. (2019) *Handbook of Diagnostic Classification Models*. Cham: Springer International.

Walstad W.B., Rebeck K. (2008) The Test of Understanding of College Economics. *American Economic Review*, vol. 98, no 2, pp. 547–551. http://dx.doi.org/10.1257/aer.98.2.547

Walstad W.B., Watts M.W., Rebeck K. (2007) *Test of Understanding in College Economics: Examiner’s Manual*. New York, NY: Council for Economic Education.

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

Zhan P., Li X., Wang W.-C., Bian Y., Wang L. (2015) The Multidimensional Testlet-Effect Cognitive Diagnostic Models. *Acta Psychologica Sinica*, vol. 47, no 5, pp. 689–701. http://dx.doi.org/10.3724/SP.J.1041.2015.00689

Zhan P., Jiao H., Liao D. (2018) Cognitive Diagnosis Modelling Incorporating Item Response Times. *British Journal of Mathematical and Statistical Psychology*, vol 71, no 2, pp. 262–286. http://dx.doi.org/10.1111/bmsp.12114