NATIONAL RESEARCH UNIVERSITY

HIGHER SCHOOL OF ECONOMICS

Institute of Education

As a manuscript

Gracheva Daria

Establishing Comparability of Test Results for Performance-Based Assessment of Complex Constructs

SUMMARY OF THE THESIS

for the purpose of obtaining academic

degree Doctor of Philosophy in Education

Academic Supervisor:

Avdeeva Svetlana, PhD

Moscow – 2024

Table of contents

Concepts and abbreviations used

Latent construct - set of patterns and regularities of manifestation of a phenomenon defined by a developer or an expert on the basis of known theories, perceptions and ideas, taking into account the available constraints and tasks.

Complex construct - a construct comprising of multiple elements, attitudes, behaviors, or ways of acting and thinking, with a focus on their application to real-life contexts.

VPBA - Virtual Performance-Based Assessment, a testing format that involves assessing a respondent's abilities through behavioural indicators in a pre-modelled digital test environment.

Behavioural indicator - an observed action (behaviour) in the test environment, by which a conclusion is made about the expression of the latent construct in the test taker.

Scenario-based tasks (scenario-based tasks) are a type of VPBA in which behavioural indicators are combined by context.

ECD - Evidence-Centered Design, an evidence-based approach to test development.

CFA - confirmatory factor analysis.

Introduction

Justification of the relevance of the study

Modern education in Russia is focused not only on the acquisition of subject knowledge, but also on the formation of students' universal learning actions (ULA), which help to successfully apply new knowledge and skills in life situations and provide the opportunity to develop independently throughout life. The Federal State Educational Standards (FSES) establish requirements for the formation of ULAs (regular, cognitive, communicative) at different levels of general education[1] . In foreign literature instead of a set of ULAs there are terms "universal skills" or "21st century skills", which include critical and creative thinking, communication, cooperation and other skills[2][3]. The importance of developing universal skills of students for self-development and successful adaptation in adult life is emphasised by many foreign experts [Ananiadoui, Claro, 2009; Griffin, Care, 2014; Pellegrino, 2017].

According to the FSES of primary general education (FSES PGE), students master basic logical and research actions, which are part of universal learning cognitive actions and at the same time are behavioural indicators of critical thinking: "choose a source of information", "recognise reliable and unreliable information", "identify the lack of information", "analyse information", "formulate conclusions and support them with evidence". In the FSES of basic general education (FSES BGE), students in the process of learning "independently formulate conclusions", learn to "select, analyse, systematise" information, assess its reliability. Also, the universal learning communicative actions described in the FSES contain characteristics of communication and cooperation, such as "show respectful attitude to the interlocutor", "express emotions in accordance with the goals of communication" and others.

[1] Orders of the Ministry of Education of the Russian Federation on the approval of federal state educational standards of primary general education, basic general education, secondary general education. Website of the Ministry of Education of the Russian Federation. URL: https://fgosreestr.ru/educational_standard

[2] New Vision for Education. Unlocking the Potential of Technology. World Economic Forum report, 2016.

[3] Partnership for 21st Century Learning (2016). Framework for 21st century learning. URL: http://www.p21.org/our-work/p21-framework

The development of universal learning actions at school leads to the need to assess them. Researchers note the complexity of assessing the ULAs [Shkerina et al., 2019]. and universal skills [Care et al., 2018; Geisinger, 2016]. compared to individual subject knowledge. In the field of measurement, universal skills are referred to as complex constructs - "constructs comprising of multiple elements, attitudes, behaviors, or ways of acting and thinking, with a focus on their application to real-life contexts." [Ercikan, Oliveri, 2016]. Other authors emphasise the presence of multiple elements and connections between them as the main characteristic of a complex construct [Ridley et al., 2021].They note that such constructs are difficult to conceptualise and operationalise [Carneiro, Rocha, Silva, 2009; Gorin, Mislevy, 2013].

In foreign literature, the term "complex construct" is predominantly used to denote universal skills - creativity, critical thinking, communication, and collaborative problem-solving skills [Andrews-Andrews-Todd, Forsyth2020; Ercikan, Oliveri2016; Hyytinen et al., 2024]. However, there are references to other complex constructs, e.g., statistical literacy [Watson, Callingham, 2003], mathematical reasoning [Pitta-Pitta-Pantazi, Sophocleous, 2017] or digital literacy [Avdeeva, Tarasova, 2023]. In PISA study, the term "complex construct" is mentioned in conjunction with the construct "global competence" [Sälzer, Roczen, 2018]. In Russia, based on the methodology and results of PISA, a tool for assessing the complex construct "functional literacy" has been developed [Kovaleva, Kolachev, 2023].

Unlike individual knowledge and skills, the measurement of complex constructs implies going beyond traditional types of tasks, such as multiple-choice tasks and self-report questionnaires. An alternative testing format should take into account multiple components of such a construct and assess different models of respondents' behaviour in a life situation [Ercikan, Oliveri, 2016]. A suitable format for assessing complex constructs is *performance-based* tasks, where test takers can demonstrate the level of skill mastery through observable actions in a predetermined situation (test environment). Due to the development of technology, the term *virtual performance-based assessment* (VPBA) has emerged [Andrews-Todd et al., 2021].where the respondent's behaviour is recorded in a digital test environment.

The category of *virtual performance-based* includes *scenario-based tasks*, in which respondents' actions in the digital environment are united by the context (*scenario-based* tasks), so such tasks are usually called contextual tasks [Ruiz-Primo, Li, 2015]. The context "immerses" respondents in the task, bringing the testing environment closer to real tasks, which is especially important when measuring universal skills. Today, scenario-based tasks are realised in digital environments using game elements and simulations, for example, to assess critical thinking [Braun et al., 2020; Uglanova et al., 2022] or collaborative problem solving skills in PISA [Stadler et al., 2020].

Despite the advantages of the new assessment format, little attention has been paid to the possibility of developing parallel forms of scenario tasks. The limitations of using only one form are obvious. In the case of repeated testing with single tasks, progress in performance may be due to practice or learning effects. In addition, frequent use of same tasks reduces test takers' involvement in the proposed scenarios because all situations seem familiar. The use of multiple forms of scenario-based tasks will not only overcome the above problems, but will also open up opportunities for large-scale monitoring and comparative studies of universal skills and other complex constructs.

Experts emphasise that the comparability of the results obtained from the scenario task forms should be ensured throughout the entire assessment cycle, from the conceptualisation and development of the measurement tool to the data handling phase and the analysis of test consequences [He, Vijver van de, 2012; Kolen, 1999]. Comparability of test results is often considered in the context of fair assessment and ensuring equal opportunities for each participant to demonstrate their knowledge and skills [Gipps, Stobart, 2009; Kunnan, 2004].

The present study focuses on methodological issues of ensuring comparability of test results of complex constructs using scenario-based tasks. In the international literature, there are examples of studies that consider multiple forms of tasks to measure universal skills or other complex constructs [Rojas et al., 2021; Wang, Liu, Hau, 2022]. However, they do not detail approaches to developing parallel forms or justify the data analysis methodology to prove comparability of results. Developing scenario tasks in a digital environment, even in a single form, is a labour-intensive process [Uglanova, Brun, Vasin, 2018]. Therefore, the task of formalising approaches for developing parallel forms of such tasks arises.

Thus, the relevance of the study is conditioned by the importance of objective and fair measurement of complex constructs and the need to ensure comparability of test results between parallel forms. Violation of the comparability of test results makes it difficult to correctly compare and interpret the results and can negatively affect decision-making on the assessment results. The development of a theoretically grounded and empirically proven approach to ensuring the comparability of scenario task parallel forms will make it possible to obtain a more accurate comparison of testing results of complex constructs, which, in turn, will contribute to improving the quality of assessment.

Evaluating complex constructs in a scenario format imposes additional challenges to the process of ensuring comparability of results. Much attention in the present work is paid to the study of context as an integral part of the scenario task. Previous studies have developed parallel test forms by modifying contextual features, for example, in the case of measuring writing skills with essays [Cho, Rijmen, Novák, 2013] or competences in the field of medicine [Lievens, Sackett, 2007]. Experts are concerned about the problem of extrapolation (generalisation) of conclusions drawn from test results using scenario tasks with different contexts [Andrews-Todd et al, 2021]. There is evidence that changing the context of a task causes differences in test results [Nelson, Guegan, 2019; Schliemann, Magalhães, 1990]. In addition, contextualised tasks may require participants to use additional knowledge and skills that are not the purpose of assessment [Messick, 1994]. As a result, context can influence the construct being measured, causing problems with validity and reliability of measurement, comparability of results [Bond, Moss, Carr, 1996]. Thus, an approach to ensuring comparability of scenario tasks should be designed to take into account context effects - changes in test results (construct structure and/or psychometric characteristics of tasks) caused by changes in the contextual characteristics of a scenario task. In addition, it is important to assess the magnitude of differences in the test results of complex constructs using scenario tasks that are caused by a change in context.

Thus, the **purpose of** this dissertation research is to develop a methodological approach to ensure comparability of scenario tasks for measuring complex constructs throughout the assessment cycle.

The **research questions of** the paper are posed as follows:

How to ensure comparability of scenario-based task forms during the development phase of measurement tools?

What methodological approaches to working with data will allow to justify comparability of test results between forms of scenario tasks?

To what extent is the context of the scenario task related to differences in test performance on complex constructs?

In order to achieve the purpose, the following objectives are addressed:

- Formulate and justify an approach to developing scenario-based task forms to ensure comparable results between forms.
- Formulate and justify methodological approaches to work with data to test the comparability of scenario-based task forms, taking into account the effect of context.
- Develop scenario-based task forms according to the chosen development approach.
- Implement a data methodology to test the comparability of scenario-based task forms.
- To quantify the association of context with differences in test performance between scenario-based task forms.

The findings of the study are validated on tasks to measure critical thinking in primary schools' students.

Theoretical foundations of the study

The work draws on the conceptual framework of comparability of measurements [He, Vijver van de, 2012; Kolen, 1999] and fairness in assessment [Kunnan, 2004] which considers the comparability of test results throughout the assessment cycle from the early stages of instrument development to the handling of test data. Based on them, two main comparability criteria were identified: construct comparability and comparability of statistical (psychometric) characteristics at the level of the whole test and individual indicators.

Degree of development of the problem

In the case of using several forms of tasks, the challenge arises to ensure comparability of test results between forms. When measuring complex constructs in a scenario format, we additionally face the challenges of reproducing the features of the test environment, the nature of the construct and the links between all its components in several forms.

Comparability of test results between test forms should be ensured throughout the entire assessment cycle - from the early stages of conceptualisation and development of the measurement instrument to data handling and interpretation of results. This requires the development of a methodological approach to ensure comparability, including the development of methodological approaches to test development and quantitative data analysis.

Next, let us consider the degree of development of the research problem from two positions that are most relevant for this study: the definition of an approach to the development of comparable scenario-based task forms and the methodology of data analysis taking into account the effect of context.

Approaches to developing comparable test forms

Several approaches to developing comparable forms of the test have been described in the literature (Gracheva, Tarasova, 2022).

The subjective approach assumes that the comparability of test forms is ensured by experts. Test developers create forms of tasks based on the test specification and make a judgement on their comparability relying on practical experience.

The task bank approach consists of randomly selecting tasks from the task bank to create new task forms [Irvine, Kyllonen, 2013]. It is believed that random selection levelling out the differences between test forms, making them comparable.

The development of comparable task forms has been actively developed within the framework of automatic item generation (AIG) research. The idea of automatic item generation is to use computer technology to automate the test development process [Gierl, Lai, Tanygin, 2021]. Among the technologies for automating the test development process, the use of machine learning models, in particular, Natural Language Processing models, is gaining popularity. For example, there are works where machine learning models are used for generating forms of reading tasks [Attali et al., 2022].

Classical approaches to automating test development use task templates, which have been described in the works of various researchers [Bejar, 1991; Haladyna, Shindoll, 1989; Osburn, 1968]. The template-based development approach assumes that a "parent task" is selected for each aspect of the content area, which is then presented as a template. The idea of templates is used in the logical structures and key elements approach [Gierl, Lai, Tanygin, 2021], as well as in the radical-incidental approach [Irvine, Dann, Anderson, 1990].

In the radical-incidental approach, tasks are broken down into elements and then the elements are categorised into radical and incidental. Radical elements determine the difficulty of the task, i.e., changing them has an impact on the psychometric characteristics of the task. Changing the incidental elements has no significant effect on the psychometric characteristics. Thus, to create comparable forms of the test, only the incidental elements of the task are to be replaced. As a result, it is possible to obtain tasks that are as similar as possible. Such tasks are called clones in the literature, and the above-mentioned approaches based on task templates can be united by the common name - cloning approaches [Clause et al., 1998; Gierl, Haladyna, 2012].

The analysis of the literature has allowed us to highlight the limitations of the described approaches in relation to the development of scenario-type tasks forms. First of all, the choice of development approach should take into account the peculiarities of the test format and the construct being measured.

Developing scenario-based tasks is a labour-intensive process (Uglanova et al., 2018), so creating a bank of tasks is not feasible. Subjectivity of experts creates risks for comparability of results between forms, and uninterpretable machine learning methods cannot accurately reproduce the peculiarities of the test environment and the nature of a complex construct.

Cloning approaches allow obtaining more comparable test forms by structuring the elements of tasks. However, there is no consensus among experts as to which elements should be emphasised in the tasks, and which of them belong to radicals and which to incidentals [Lievens, Sackett, 2007; Williamson et al., 2002].

As a limitation, we note that the cloning approach is referred to as a simplistic approach to developing test forms that provokes "cramming". Reproducing most of the elements of the parent task leads to the creation of almost identical tasks. At the same time, it is noted that an important aspect of measuring high-order skills is the possibility of their assessment in a free test environment [Poddyakov, 2012]. A contradiction arises between the free environment necessary for the manifestation of a complex skill and fair assessment.

With the development of AIG technologies, cloning approaches have been actively used in various fields. For example, in creating simulations in the field of networking technologies [Fay, Levy, Mehta, 2018], medical situational tests [Lievens, Sackett, 2007], computerised tests for managers [Lievens, Anseel, 2007]. However, no examples of using cloning approaches to develop scenario-based tasks to measure complex constructs were found in the literature.

The Evidence-Centered Design (ECD) approach is used to create scenario tasks that evaluate complex constructs [Mislevy, Haertel, 2006], which considers the testing process as a process of gathering evidence to draw a reasonable conclusion about the test takers' abilities. The structure of the ECD consists of several models that integrate the basic processes in creating a measurement instrument. This evidential approach helps to ensure measurement validity throughout the assessment cycle.

Thus, cloning approaches have not been sufficiently explored in application to the creation of comparable forms of scenario tasks. The cloning approach in combination with ECD may be a promising approach for ensuring comparability of measurement results of complex constructs.

Among the primary challenges for work in this area are the following: identifying specific elements of scenario-based tasks and finding a balance between a free test environment and cloning rules to ensure comparable results between scenario forms.

However, studies show that providing comparability measures at the development stage is not a guarantee of comparability at the data level [Lee, Anderson, 2007]. Specific research is needed to prove that test results between scenario-type task forms are comparable.

Data manipulation techniques to ensure comparability of test forms

The data obtained from testing complex constructs using scenario-based tasks have features that determine the methods of analysis.

Firstly, complex constructs include several components that are the purpose of assessment.

Multivariate measurement models within the latent modelling approach are used to analyse complex constructs [Levy, 2013].. For example, multivariate IRT models [Reckase, 2006]. Alternatively, measurement models in structural equation modelling (SEM) methodology - multivariate confirmatory factor analysis models - can be used as an alternative [Brown, 2006].

The second difference between the assessment of complex constructs using scenario tasks and "traditional" assessment is the violation of the assumption of local independence of behavioural indicators. Scenario tasks differ from more traditional tasks, such as multiple-choice, in having a richer test environment, context and simulations that create dependencies between the observed actions of the respondent. One way to account for such contextual dependencies is to isolate the context factor through the construction of bifactor models [Levy, 2013; Rijmen, 2010].

Thus, the analysis of the results of measuring complex constructs using scenario tasks involves the use of multivariate latent modelling methodology, taking into account the contextual relationships between indicators.

In the methodology of structural equation modelling, comparability at the level of the whole test and psychometric characteristics of tasks can be assessed in the multigroup

CFA model. The proof of comparability is reduced to the sequential verification of the levels of invariance of the multigroup model: at the level of construct structures (configural invariance) and individual psychometric characteristics of indicators: discrimination and difficulty (metric and scalar invariance) [Brown, 2006].

Although measurement invariance methodology traditionally tests the comparability of a measurement instrument between different groups of respondents, there are studies that have tested instrument functioning across different versions (forms) of the test [Rojas et al., 2021].

This thesis emphasises the importance of investigating the context of a scenario task to ensure comparability of measurement. The effect of context in the forms of scenario-based tasks can cause differences in both the theoretical structure of the construct and the characteristics of individual indicators. There is a need not only to substantiate methodological approaches to ensuring comparability of scenario task forms taking into account the context component, but also to quantify the effect of context on test results.

In the literature, more attention has been paid to investigating the method effect - how test results depend on the measurement method [Eid, Geiser, Koch, 2016]. To investigate the method effect, a methodology based on the Multitrait-Multimethod Matrix (MTMM), which was proposed by D. Campbell and D. Fiske, is used [Campbell, Fiske, 1959].

Generalizability theory is another method within which the effect of the method can be studied. The basics of the Generalizability Theory are described in the articles by L. Cronbach [Cronbach, 1972].and were later supplemented in the works of R. Shavelson and R. Brennan [[Brennan, 1992; Shavelson, Webb, Rowley, 1992]. Within the framework of this thesis research, the foundations of the Generalizability Theory were described in the article [Gracheva, 2023 .

In articles that investigate method effects, the question of the interaction effect between the measurement method and the respondent is raised. It is suggested that the method effect may not be equivalent for all respondents [Kroehne et al., 2019; Shavelson, Baxter, Gao, 1993]. For example, some respondents perform better on scenario tasks in one context and worse in another context, and vice versa for other respondents. Generalizability Theory methods not only quantify the overall effect of context, which is the same for all respondents, but also the effect of the interaction between the test taker and the context.

Thus, to check comparability at the stage of working with data, it is proposed to use multigroup CFA models, which will allow us to simultaneously assess whether the construct and psychometric criterion of comparability of measurements will be achieved. At the same time, the use of multigroup models should take into account the peculiarities of complex construct assessment via scenario tasks: multidimensionality and consideration of contextual relations between indicators. The class of bifactor CFA models meets these needs. An equally important task is to quantify the effect of context on the results of testing complex constructs using scenario tasks.

Research methodology and design

The empirical basis of the study is the data on the results of testing using the 4C tool for measuring critical thinking in primary school students, developed by the staff of the Laboratory of Measurement of New Constructs and Test Design of the Institute of Education of the HSE University.

The principles of Evidence-Centred Design (ECD) [Mislevy, Haertel, 2006] and automatic item generations [Gierl, Lai, Tanygin, 2021] were used to formulate an approach to developing comparable forms of scenario tasks.

To prove comparability at the data level, latent modelling methodology was applied and measurement invariance was tested using a multigroup confirmatory factor analysis (CFA) model, taking into account contextual relationships between indicators (bifactor CFA models). Because behavioural indicators are often categorical (dichotomous or polytomous), it is recommended to use confirmatory factor analysis for categorical variables (categorical confirmatory factor analysis (CCFA)) [Kim, Yoon, 2011]. Additionally, the study used the CFA-MTMM methodology to assess the degree of comparability of scenario task forms developed in the logic of the cloning approach. Generalizability Theory methods were used to assess the effect of context on the results [Cronbach, 1972].

A balanced within-group design was used to collect data on scenario task forms, where test forms are presented to one group of respondents in all possible orders (randomly). The choice in favour of the within-group design over the between-group design avoids alternative explanations of the results related to differences in respondent characteristics. The study was conducted on samples of 4th grade students (samples ranging from 381 to 1,096 students) who participated in assessment of 21st century skills.

Instrument description

The 4C tool is designed following ECD. This research uses two scenario-based tasks for measuring critical thinking in primary school: the "Aquarium" task and the "Dino" task.

According to the conceptual framework of the tool, critical thinking skill includes the skill of working with information in accordance with the goals and conditions of the task and the skill of formulating one's own conclusion using the results obtained at the analysis stage. The tool assesses the following components of the skill of analysing information: identifying valid (reliable) sources of information; identifying relevant information for the task. The conceptual framework of critical thinking is presented in more detail in the article [Uglanova et al., 2022].

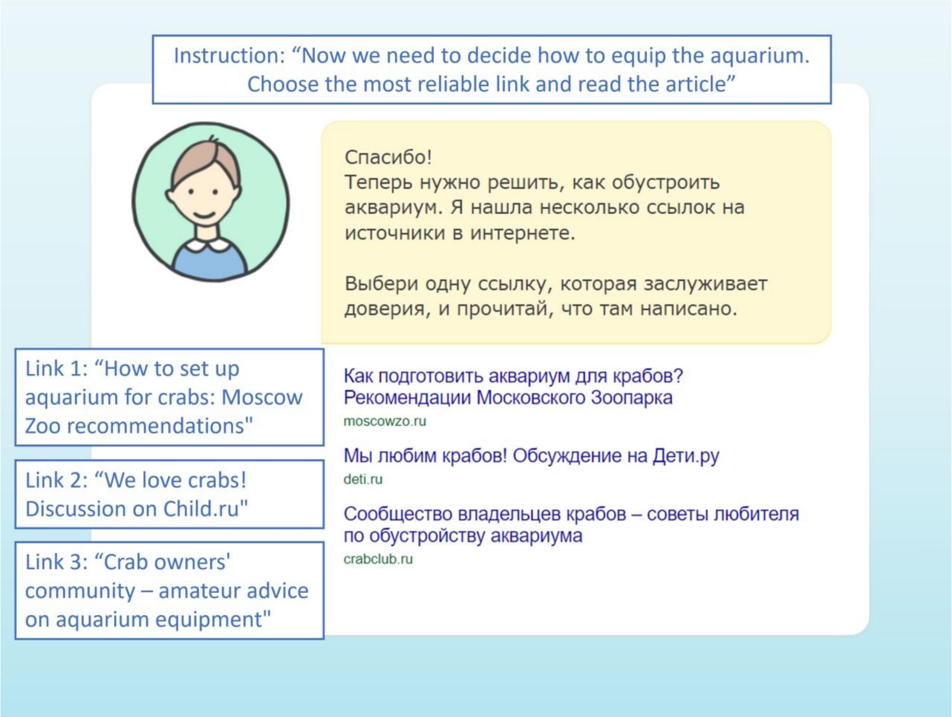
Below is a brief description of scenario-based tasks in two forms.

“Aquarium” scenario

The context of the scenario sets up a problem situation where the test taker needs to equip an aquarium for crabs.

The plot of the scenario assumes that the test taker first needs to identify a reliable source containing information on how to equip a crab aquarium. To do this, the task includes a simulation of an Internet browser where several links are presented, only one of which is reliable. When selecting the most reliable link, the test taker demonstrates the ability to identify a reliable source of information and receives 1 point (Figure 1).

Next, the test taker analyses the text on the selected link to find out what objects are needed to equip an aquarium (Figure 2). The instructions for this part of the task are as follows: "Highlight sentences with useful information that help you yo equip the aquarium." For each correctly highlighted (relevant) sentence in the text of the article the test taker gets 1 point. As a result of this part of the task, the test taker saves the sentences from the text of the article into a notebook (right side of Figure 2).



Instruction: “Now we need to decide how to equip the aquarium. Choose the most reliable link and read the article”

Спасибо!
Теперь нужно решить, как обустроить аквариум. Я нашла несколько ссылок на источники в интернете.

Выбери одну ссылку, которая заслуживает доверия, и прочитай, что там написано.

Link 1: “How to set up aquarium for crabs: Moscow Zoo recommendations"

Как подготовить аквариум для крабов? Рекомендации Московского Зоопарка
moscowzo.ru

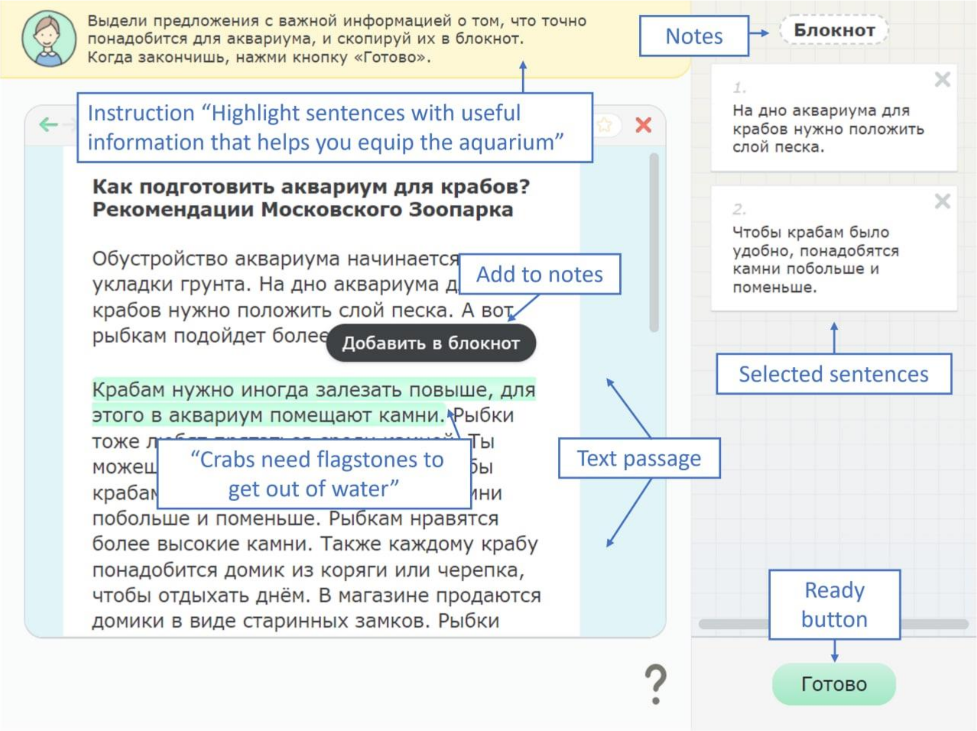
Link 2: “We love crabs! Discussion on Child.ru"

Мы любим крабов! Обсуждение на Дети.ру
deti.ru

Link 3: “Crab owners' community – amateur advice on aquarium equipment"

Сообщество владельцев крабов – советы любителя по обустройству аквариума
crabclub.ru

Figure 1 - Example of the screen of the scenario task "Aquarium" (reliability of the information source)



Выдели предложения с важной информацией о том, что точно понадобится для аквариума, и скопируй их в блокнот.
Когда закончишь, нажми кнопку «Готово».

Instruction “Highlight sentences with useful information that helps you equip the aquarium”

Как подготовить аквариум для крабов? Рекомендации Московского Зоопарка

Обустройство аквариума начинается укладки грунта. На дно аквариума д крабов нужно положить слой песка. А вот рыбкам подойдет более

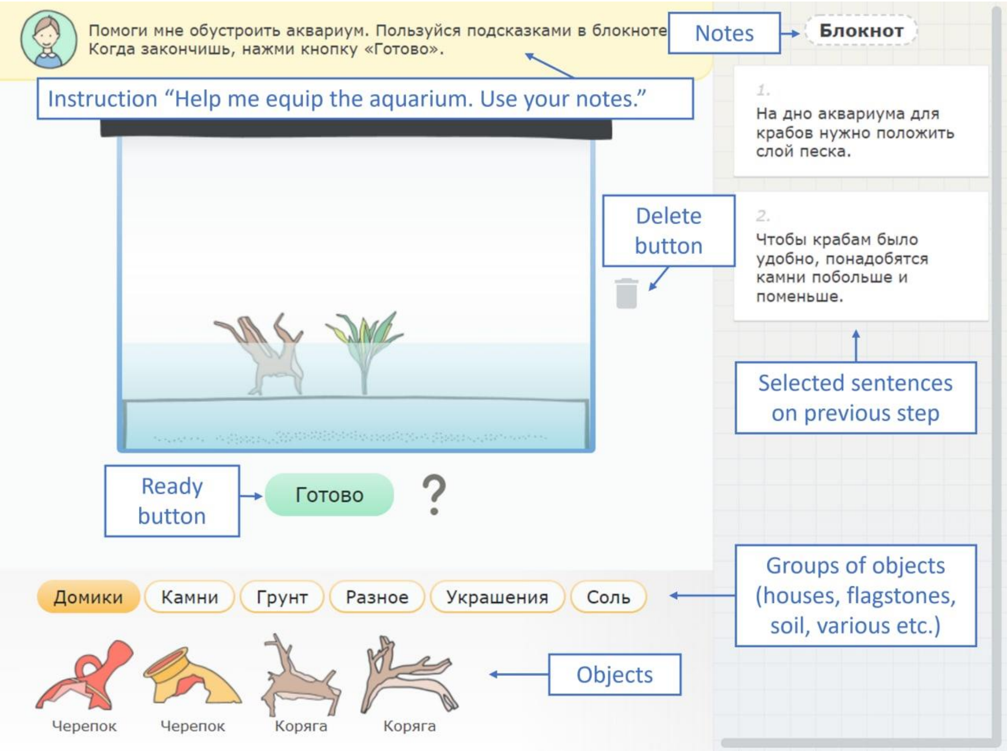
Крабам нужно иногда залезать повыше, для этого в аквариум помещают камни. Рыбки тоже л Ты можеш бы крабам ни побольше и поменьше. Рыбкам нравятся более высокие камни. Также каждому крабу понадобится домик из коряги или черепка, чтобы отдыхать днём. В магазине продаются домики в виде старинных замков. Рыбки

“Crabs need flagstones to get out of water”

1. На дно аквариума для крабов нужно положить слой песка.

2. Чтобы крабам было удобно, понадобятся камни побольше и поменьше.

Figure 2 - Example of the Aquarium scenario task screen (relevance of information)



Помоги мне обустроить аквариум. Пользуйся подсказками в блокноте
Когда закончишь, нажми кнопку «Готово».

1. На дно аквариума для крабов нужно положить слой песка.

2. Чтобы крабам было удобно, понадобятся камни побольше и поменьше.

Notes
Блокнот
Instruction “Help me equip the aquarium. Use your notes.”
Delete button
Selected sentences on previous step
Ready button
Готово
?
Groups of objects (houses, flagstones, soil, various etc.)
Домики
Камни
Грунт
Разное
Украшения
Соль
Objects
Черепок
Черепок
Коряга
Коряга

Figure 3 - Example of the screen of the scenario task "Aquarium" (making conclusion)

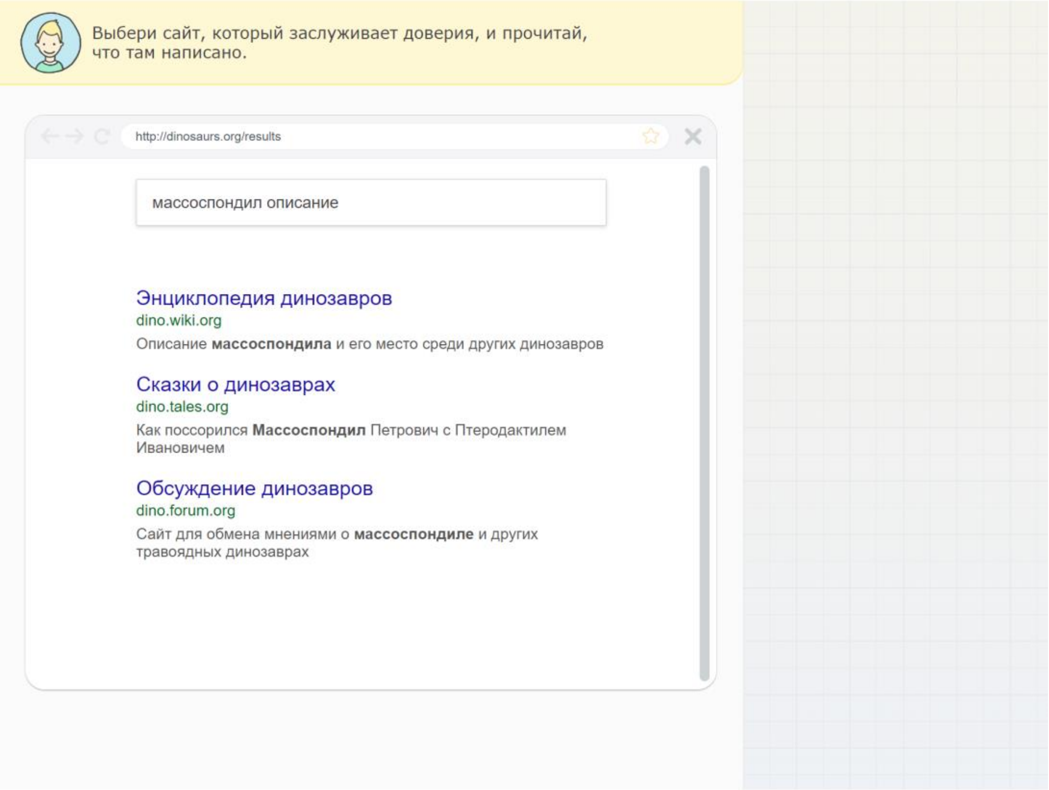
Finally, based on the analysed text, the test taker equips an aquarium from the objects in the simulation. The simulation contains objects that were mentioned in the text of the article and those that were not mentioned, grouped into categories: houses, rocks, soil, miscellaneous, decorations, salt (Figure 3).

According to the analysed information, the test taker must decide which objects should be placed in the aquarium and which should not. The test taker receives 1 point for each correctly placed object.

“Dino” scenario

The context of the scenario sets up a problem situation where the test taker is asked to help prepare a report on a non-existent Massospondylus dinosaur to answer the key question of how many legs this dinosaur walked on.

To achieve the goal, the test taker must select the most reliable link that contains reliable information about this type of dinosaur (Figure 4). The test taker gets 1 point for choosing the most reliable link.



Выбери сайт, который заслуживает доверия, и прочитай, что там написано.

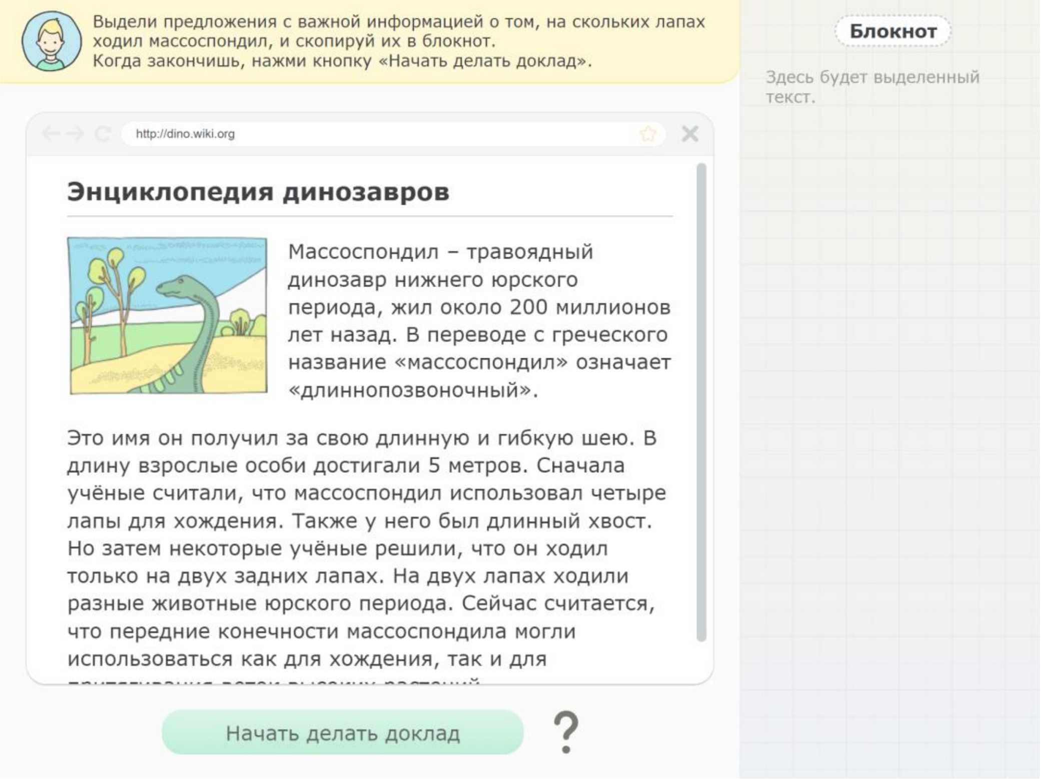
массоспондил описание

Энциклопедия динозавров
dino.wiki.org
Описание **массоспондила** и его место среди других динозавров

Сказки о динозаврах
dino.tales.org
Как поссорился **Массоспондил** Петрович с Птеродактилем Ивановичем

Обсуждение динозавров
dino.forum.org
Сайт для обмена мнениями о **массоспондиле** и других травоядных динозаврах

Figure 4 - Example of the screen of the scenario task "Dinosaur" (reliability of the information source)



Выдели предложения с важной информацией о том, на скольких лапах ходил массоспондил, и скопируй их в блокнот.
Когда закончишь, нажми кнопку «Начать делать доклад».

Энциклопедия динозавров

Массоспондил – травоядный динозавр нижнего юрского периода, жил около 200 миллионов лет назад. В переводе с греческого название «массоспондил» означает «длиннопозвоночный».

Это имя он получил за свою длинную и гибкую шею. В длину взрослые особи достигали 5 метров. Сначала учёные считали, что массоспондил использовал четыре лапы для хождения. Также у него был длинный хвост. Но затем некоторые учёные решили, что он ходил только на двух задних лапах. На двух лапах ходили разные животные юрского периода. Сейчас считается, что передние конечности массоспондила могли использоваться как для хождения, так и для

Figure 5 - Example of the screen of the scenario task "Dinosaur" (relevance of information)

Next, the test taker is asked to analyse the text of the electronic article, highlighting only the important information about how many legs the massospondylus walked on (Figure 5). The article test contains relevant sentences that contain three different points of view regarding the question asked (the dinosaur used two legs to walk, four legs to walk, or it is not known exactly). Finally, the test taker is asked to make a final conclusion about how many legs the Massospondylus dinosaur walked on, based on the analysed information.

Results of the study

The results of the study are presented according to the research questions.

The first stage of the work answered the question - **how to ensure comparability of scenario-based task forms during the development phase of measurement tools?**

As part of this thesis research, a cloning approach was formed to develop comparable forms of scenario-based tasks. The main results of this stage of the research are published in the article [Gracheva, Tarasova, 2022] and expanded in the text of the thesis research.

The literature review concluded that the cloning approach is the most appropriate approach for developing comparable scenario-based task forms measuring complex latent constructs. The cloning approach involves creating a task structure based on a "parent task" and reproducing this structure when developing task forms. To create such a structure, the task must be divided into elements that can be radical, influencing the difficulty of the task, or incidental.

The result of the research was the description of the cloning approach for scenario tasks. It is proposed to distinguish elements of scenario tasks on two levels. The upper level elements include the context and content of the scenario task. The research proposed definitions of context and content as separate complex elements that are an integral part of the scenario-based assessment format.

"The *context of* a scenario task is the stimulus material that defines the main problem situation of the scenario task and the development of the situation (plot) - the sequence of actions, relations between the stages of the task, characters, etc.".

According to this definition, we can think of context as a complex element of a task that can be modified either completely, or partially, as part of the cloning procedure.

The same situation can have different thematic *content*. Therefore, we additionally introduce the concept of scenario task content - a characteristic of the stimulus material that determines the thematic content of the scenario. We will also consider the content as a complex element of the scenario task, which can be changed.

Lower-level items included stimulus items that motivate respondents to perform actions that reflect the target construct (e.g., highlighting relevant information in the text), thematic items that are modified by new content, mechanistic items that reflect the test taker's interaction with the test environment (e.g., highlighting, selecting, dragging and dropping, etc.), and text structure items.

The development of comparable forms of scenario tasks implied the division of the selected elements into radical and incidental. The context of the task was chosen as the main rqadical element, and the content was chosen as a incidental element, the change of which does not significantly affect the psychometric characteristics of the task. The lower level elements could refer to both radical and incidental elements.

For the "Aquarium" scenario task, the scenario context defines a problem situation where the test taker needs to equip a space for a pet using a set of objects. The context of the scenario task "Aquarium" is the equipment of an aquarium for crabs. When developing a parallel forms of the scenario task "Aquarium" new content was used - equipment of a terrarium for geckos. Further clone of the "Aquarium" scenario task will be called "Terrarium".

For the scenario task "Dino" the scenario context sets a problem situation where the test taker is asked to help analyse information about some object/subject to prepare a report and answer the key question of the report. The context of the scenario task "Dino" is preparing a report about a non-existent dinosaur Massospondylus to answer the key question how many legs this dinosaur walked on. In developing the parallel form, new content was used where the test taker is asked to analyse information about why hedgehogs rub their needles on objects (scenario task "Hedgehog").

In the article [Gracheva, Tarasova, 2022] it was proposed to distinguish strict and non-strict cloning approaches depending on which elements are considered radicals. The description of strict and non-strict cloning directions was extended and supplemented in the present study.

In the strict cloning approach, sentence structure is a radical element of the task that cannot be changed. In previous studies, grammatical bases of sentences have also been a radical element in cloning of situational judgment tests [Lievens, Sackett, 2007].

In the non-strict cloning approach, there is no requirement to follow sentence structures. The sentences in the texts of scenarios may also contain stimulus elements, mechanics elements, and thematic elements that can be used by developers to guide the creation of an alternative task forms. However, the focus of the non-strict cloning approach is to develop a test environment for the new forms that will help to reproduce the test taker's behaviour without regard to strict adherence to the sentence structures in the texts.

Strict cloning was implemented on a pair of scenarios "Aquarium" / "Terrarium", non-strict cloning on a pair of scenarios "Dino" / "Hedgehog" in terms of tasks for assessing information analysis as a component of critical thinking. Detailed templates of cloning tasks are given in the text of the thesis research.

The following hypotheses were posed in the study:

- forms of scenario tasks created according to the strict cloning approach will be more comparable to each other than forms of scenario tasks developed in the logic of non-strict cloning.
- forms of scenario tasks that have a common context but different content will be more comparable to each other than scenario tasks developed in different contexts.

Using the CFA-MTMM method, it was proved that there is no statistically significant difference in the consistency of test forms created according to different cloning approaches. At the same time, forms created in the same context show greater comparability than scenario-based tasks with different contexts (e.g., "Aquarium" and "Dino"). That is, both (strict and non-strict) cloning approaches can be used to develop forms of scenario-based tasks that measure complex constructs. This opens the door for developers to exercise greater freedom in developing comparable task forms and avoids the problem of excessive similarity of cloned tests. As a limitation of this finding, it is worth noting that the parts of the scenario tasks for assessing information analysis skills are linearly structured and focus more on working with electronic articles and searching for sources in a browser simulation. The test environment significantly limits the behaviour that the test taker can demonstrate in a similar situation in real life.

The results of the study should be accepted subject to limitations. In the study, cloning approaches were tested on different scenarios (in terms of number of indicators, length, context, etc.), therefore, to confirm the conclusions, a repeated study can be conducted using several forms of the same scenario task, which are developed in the logic of strict and non-strict cloning.

However, in order to make a judgement about comparability of measurements, it is also necessary to prove the comparability of the construct structure and the psychometric properties of the individual indicators.

The second stage of the work answered the question - **what methodological approaches to working with data will allow to justify comparability of test results between forms of scenario tasks?**

When working with scenario-based tasks, there is a risk that test results are explained not only by the latent trait being measured, but also by the context component. It becomes necessary to "cleanse" the test results from the context effect.

This dissertation study proposed to extend the methodology for testing comparability of test results by using bifactor models, which allow us to replicate the multidimensional structure of a complex construct and account for contextual relationships between behavioural indicators. Bifactor models assume that differences in responses to a task can be explained by a general factor (latent construct) and specific factors that do not relate to it (contextual factors).

The results from the article [Gracheva, 2022] shows that the structure of the measurement model taking into account the contextual relationships between the indicators fits the data well. To obtain reliable results about the comparability of the structure of critical thinking and the characteristics of individual indicators, the measurement invariance of the measurement tool was tested on the model taking into account contextual relationships (context factors in the CFA bifactor model).

The justification of the test development approach and methodology of data analysis allows the implementation of a quantitative analysis methodology to ensure the comparability of scenario-based task forms developed in accordance with the chosen test development approach. The following hypothesis was posed as part of the research:

- forms of scenario tasks created using the cloning approach will have equivalent construct structures and psychometric characteristics of the indicators.

The analysis of comparability at the data level includes two steps: checking the measurement invariance of the instrument (comparability of construct structures and characteristics of individual indicators) and comparing average results across test forms within the framework of latent modelling methodology.

The measurement invariance between scenario forms was tested using the example of the "Aquarium" scenario, which measures two components of critical thinking: analysing information and making a conclusion. Indicators of the ability to make a conclusion in this task are measured in a freer test environment: based on the read material of an electronic article, the test taker equips an aquarium from a set of objects (Figure 3). For the alternative scenario, the names of the objects and the interface were cloned.

The measurement invariance analysed in the article [Gracheva, 2022] showed that both forms reproduce the theoretically expected structure of critical thinking (analysis and conclusion factors were distinguished separately), psychometric characteristics (difficulty and discrimination) of indicators are equivalent. It is concluded that the cloning approach allows obtaining comparable psychometric quality of scenario format measurement tools.

Next, the average results (on a scale of factor scores that are estimated by the CFA model) were compared across forms. Such comparison of results is considered more accurate if the structure of the construct and the characteristics of indicators are equivalent. On average, the test takers' results on the skill of analysing information did not differ across forms. However, statistically significant differences were found in the average results for the skill of making a conclusion.

The study used a balanced within-group data collection design, where both forms of the scenario task were completed by all test takers, and the forms were presented in random order. Therefore, the differences obtained may be related to differences in content (thematic content of the scenario) rather than to the effect of learning in solving similar tasks or experience of interaction with the computer interface. An alternative explanation could be the format of the tasks, which implies greater freedom of the test taker and an element of interactivity. It can be concluded that tasks realised in a more interactive environment are at greater risk of incomparability of test results.

The study has limitations. Comparability analyses were conducted on a single pair of scenario tasks created in a strict cloning approach, using critical thinking as an example of a complex construct. The design and methodology of the study can be applied to analyse the comparability of other scenarios and constructs to validate the findings of the study.

The third stage of the work answered the question **- to what extent is the context of the scenario task related to differences in test performance on complex constructs?** The purpose of the third study was to assess the effects of context and content on test scores of complex constructs using scenario tasks. Both effects were quantified within the Generalizability Theory methodology. The results of the study were published in the article [Gracheva, 2023].

The article uses data obtained in the autumn of 2021 during the testing of 4th grade students who participated in the study of universal skills (critical thinking, creativity, communication, cooperation) using the "4C" tool. To assess critical thinking, students were asked to complete three initial scenario tasks ("Aquarium", "Dino", "Journey").

Additionally, students were asked to complete alternative forms of the scenarios following a within-group balanced design. However, due to time constraints, it was not possible to use all scenario tasks in two forms within one testing session. Therefore, test takers were randomly divided into groups. The first group took the "Aquarium" and "Terrarium" scenarios (998 respondents, approximately 2/5 of the whole sample), the second group took the "Dino" and "Hedgehog" scenarios (466 respondents, approximately 1/5 of the whole sample), the third group took the "Journey" and "Labyrinth" scenarios (1096 respondents, approximately 2/5 of the whole sample). The Journey/Labyrinth and Aquarium/Terrarium scenarios were administered to more respondents because they are key tasks for assessing critical thinking, communication and cooperation skills in the 4C tool (they contain more indicators).

The research design of the study yielded data on all scenario-based tasks to measure critical thinking in two forms.

The context effect was assessed by comparing different scenario tasks (with different contexts) measuring critical thinking. The effect of content was assessed by comparing comparable forms of scenario tasks developed in the logic of the cloning approach. As a result of the analysis, it was found that the effect of changing the content on test results is lower than the effect of changing the context, which confirms the logic of the cloning approach, where the context is classified as a radical element of the scenario task, and the content is classified as a incidental element.

Generalizability Theory methods allowed us to assess not only the overall effect of context and content, but also these effects in interaction with the test taker. As a result of the analysis, it was found that the interaction effect of context and test taker was higher than the overall effect of context. Similar results were obtained for the content effect analysis. In other words, for one test taker, the context of one scenario was easier than the context of another scenario and vice versa for another test taker. It can be concluded that ensuring comparability of scenario task contexts at the development stage of the measurement tool will minimise the overall effect of context, but contextual tasks are inevitably subject to test taker-context interaction effects. In previous studies of performance-based tasks, the test taker-task interaction effect was predominant [Shavelson, Baxter, Gao, 1993]. Also, studies emphasise that when contextual tasks are used, the degree to which the test taker is aware of the context (or content) can influence test scores [Ahmed, Pollitt, 2007].

To minimise effects that are irrelevant to the construct being measured, testing of complex skills should include scenario tasks with different contexts. In this study, it is empirically proven that critical thinking tests should be administered in at least two contexts in order to achieve satisfactory reliability. Increasing the number of contexts will not only increase the reliability of the measurement, but also the validity of the conclusions that are drawn from critical thinking testing.

The research conducted has limitations. Firstly, critical thinking is considered as a unidimensional construct. Analysis in terms of components is possible using the methodology of multidimensional Generalizability Theory [Keller, Clauser, Swanson, 2010]. Secondly, this study uses classical Generalizability Theory methods based on raw test data, but there are extensions of this approach in structural equation modelling methodology [Jorgensen, 2021] or Bayesian networks [Jiang, Skorupski, 2018].

Conclusion

This study analyses previous research on measurement comparability and fair assessment. It was found that existing conceptual frameworks of measurement comparability consider comparability throughout the assessment cycle [He, Vijver van de, 2012; Kolen, 1999; Kunnan, 2004]. Thus, the decision on measurement comparability should be based on the different evidences of comparability collected during the design and implementation phases of a measurement instrument. Based on this, the study presents a methodological approach to ensure comparability of scenario task forms for measuring complex constructs (cloning approach) from the measurement tool development stage to the handling of test data, linking the principles of ECD (at the stages of construct domain modelling, task model creation and measurement model creation) and the principles of automatic item generation.

When forming the cloning approach, the peculiarities of measuring complex constructs using scenario tasks were taken into account. The main feature of the scenario-based task is the presence of a context that combines behavioural indicators of the construct, brings the task closer to real life and motivates the respondent to perform actions reflecting the latent construct. The study suggests separating the context (the main problem situation of the scenario and its development) and the content (thematic content) of the scenario task. In order to create comparable forms of scenario tasks, it is proposed to change the thematic content while preserving the scenario context.

Based on the proposed approach, two scenario tasks were developed to measure critical thinking in primary school students. The tasks from the 4C tool were used as initial scenario tasks.

In terms of working with data, it is proposed to expand the methodology of testing results comparability verification through the use of bifactor models, which allow to reproduce the multidimensional structure of the complex construct and take into account the contextual relationships between behavioural indicators. Empirically, a critical thinking assessment model based on CFA bifactor models was found to be consistent with data measuring complex constructs in a scenario format. Data analyses using this model showed that task forms created based on the proposed cloning approach demonstrate equivalent construct (critical thinking) structures, and psychometric characteristics of indicators (difficulties, discriminations).

In this study, the task was to quantify the relationship between context and differences in test results between scenario task forms. Using the Generalizability Theory methodology, the effect of scenario task context and the effect of context and respondent interaction on critical thinking test results were evaluated separately. It was found that the effect of test taker and context interaction was higher than the effect of the overall context of the scenario task, i.e. one context may be easier for one respondent and more difficult for another respondent. It has been empirically proved that testing complex constructs in several contexts (using several scenario tasks) allows to reduce contextual effects and increase the reliability of the measurement.

The analysis allowed us to compare the context effects obtained when comparing different scenario tasks and the content effects obtained when comparing comparable forms of scenario-based tasks developed in the logic of the cloning approach. It is found that the effect of content on the test scores of complex constructs is lower than the effect of scenario context. The obtained result confirms that the development of comparable variants of scenario-type tasks can take place by changing the content (thematic content), preserving the main scenario context, which makes more differences in the measurement results.

The theoretical significance of the work in the area of educational measurement lies in the presentation of a methodological approach to ensuring comparability of scenario tasks for the assessment of complex constructs from the development of a measurement tool to the handling of test data. The work proposed definitions of scenario task context and content that can serve as a basis for further development of methodological approaches to developing comparable forms of scenario tasks.

The **practical significance of** the work consists in the fact that the results of the study can be used to develop comparable forms of scenario-based tasks measuring complex constructs in the logic of ECD. In the context of Russian education, a set of universal learning actions, which are fixed in the FSES, can be considered as complex constructs. The forms of scenario tasks of the "4C" tool for assessing critical thinking in primary school students, considered in this study, can be recognised as comparable and used in future comparative and monitoring studies of critical thinking, responding to the demand for fair assessment. New empirical evidence is obtained on how changing the context and content of scenario-based tasks in critical thinking assessments is related to test scores and measurement reliability, which may be useful to developers when designing critical thinking assessments. In addition, the proposed development approach will reduce the time and human resources required to create forms of scenario tasks without sacrificing their quality.

The study has limitations. First, the proposed cloning approach was developed for instruments measuring complex constructs using universal skills as an example. Extrapolation of the approaches to specialised skills and subject knowledge requires further adaptation of the cloning model and an empirical comparability study. A related limitation is that the empirical part of the study was conducted only on data from the 4C instrument on a sample of primary school students. In future studies, the proposed methodological approaches can be tested on other instruments, complex constructs and age group of respondents.

Secondly, methodological approaches to ensure comparability were developed for one testing format - *scenario-based* tasks in a digital environment as a kind of *performance-based* tasks. Unlike traditional tests with a set of independent tasks, *performance-based* tests do not have a clear structure. In this study, we draw on the work of [Andrews-Todd et al., 2021] who introduced the term *virtual-performance-based assessment* and consider three types of VPBAs: simulations, scenario-based tasks and game-based assessment tasks. However, in real practice there is no unified understanding of VPBA types on the part of researchers, and completely different test formats can be found under the term *performance-based assessment*. This "terminological confusion" is a limitation to the application of the research findings. In addition, as technology advances, scenario-based tasks may have varying degrees of interactivity and simulation complexity. The results of this study showed that tasks that involve greater freedom of action (assembling an object from items) are at greater risk of non-comparability of results. Ensuring comparability of freer test environments may require new methodological approaches and may be a continuation of this study.

Third, implementing the cloning approach to develop comparable versions of scenario tasks requires immersion of the researcher or developer in the theoretical model of the construct. Although the cloning approach is based on the principles of automatic item generation and is intended to simplify the development process, the role of the expert in the development process is still key. A promising direction for future research is to combine machine learning technologies and pre-designed cognitive models and job templates for cloning. Studies are already emerging where generative artificial intelligence models have been trained to develop reading comprehension assessment tasks based on cognitive task models [Sayin, Gierl, 2024]..

Future research directions include examining the effects of context and content on the test scores of complex constructs in educational and psychological testing. This study was able to establish the presence of context effects at the general level and in interaction with the respondent, but the reasons for these effects require further investigation. For example, context effects can be further investigated under the concept of knowledge transfer from one context to another [Barnett, Ceci, 2002]. A separate study could be devoted to analysing the relationship of context characteristics (e.g., abstractness, proximity to real-world tasks) to test scores on complex constructs.

In conclusion, the evidence of comparability collected predominantly relied on quantitative research methods. Creating guidelines for conducting cognitive labs with respondents to explain the effects of context and possible differences in test scores between options would be an important addition to this study.

Conclusions submitted for defence

1. The approach to developing comparable forms of scenario tasks proposed in this study, which links the stages of construct domain modelling and task model development in the evidence-centered design paradigm, allowed us to ensure equivalence of the structure of the complex construct and psychometric characteristics of behavioural indicators between the forms.

2. The data methodology to ensure comparability of scenario task forms was extended to include methods based on bifactor models to account for the multidimensional nature of the complex construct and the relationships between behavioural indicators due to the presence of a contextual component.

3. The proposed methodology allowed us to identify the effect of scenario task context on test scores and found that the effect of context and respondent interaction on scores was higher than the effect of overall scenario task context.

4. To reduce the effect of context on test results and increase measurement reliability when assessing complex constructs, multiple scenario tasks with different contexts should be used.

Approbation and implementation of the research results

List of publications of the author of the thesis, which reflect the main scientific results of the research:

- Gracheva, D.A. The Role of Context in Scenario-Based Tasks for Measuring Universal Skills: The Use of Generalizability Theory / D.A. Gracheva // Educational Studies. - 2023. - № 3. - С. 221-230;
- Gracheva D. A. Analysis of Task Comparability in Digital Environment by the Case of Metacognitive Skills // Psychological Science and Education. - 2022. - № 6 (27). - С. 57-67;
- Gracheva, D. A. Approaches to the Development of Scenario-Based Task Forms Within the Framework of Evidence-Centered Design / D. A. Gracheva, K. V. Tarasova // Domestic and foreign pedagogy. - 2022. - № 3 (1). - С. 83-97.

Additional publications with the author's participation on the topic:

- Uglanova I. Computer-based performance approach for critical thinking assessment in children / I. Uglanova, E. Orel, D. Gracheva, K. Tarasova // British Journal of Educational Psychology. Gracheva, K. Tarasova // British Journal of Educational Psychology. - 2023. - №93. - P. 531-544.

List of scientific conferences where the results of the study were presented:

- Quantitative Research Methods Conference (QRM). Paper: Testing measurement invariance across alternative test forms? 14-15 June 2021, online conference.
- Conference "13th Annual International Conference on Education and New Learning Technologies (EDULEARN21)". Paper: Investigating the effect of context on comparability of computerised performance-based tasks, 5-6 July 2021, online conference.
- Conference "22nd Annual Meeting of the Association for Educational Assessment - Europe (AEA-Europe). Assessment for Changing Times: Opportunities and Challenges". Paper: Comparability of computerised performance-based assessment for measuring critical thinking, 9-12 November 2021, Dublin, Ireland (online presentation).
- Conference "24th Annual Meeting of the Association for Educational Assessment - Europe (AEA-Europe 2023). Assessment reform journeys: intentions, enactment and evaluation" Paper: The application of generalisability theory to the scenario-based performance assessment of 21st century skills: analysis of task context effects. 6-9 November 2023, Malta.

References

1. Avdeeva, S.M. Digital Literacy Assessment: Methodology, Conceptual Model and Measurement Tool / S.M. Avdeeva, K.V. Tarasova // Educational Studies. – 2023. – № 2. – C. 8-32.
2. Gracheva, D.A. The Role of Context in Scenario-Based Tasks for Measuring Universal Skills: The Use of Generalizability Theory / D.A. Gracheva // Educational Studies. – 2023. – № 3. – C. 221–230.
3. Gracheva D. A. Analysis of Task Comparability in Digital Environment by the Case of Metacognitive Skills // Psychological Science and Education. – 2022. – № 6 (27). – C. 57-67.
4. Gracheva, D. A. Approaches to the Development of Scenario-Based Task Forms Within the Framework of Evidence-Centered Design / D. A. Gracheva, K. V. Tarasova // Domestic and foreign pedagogy. – 2022. – № 3 (1). – C. 83–97.
5. Kovaleva, G. S. Functionality of the project "Monitoring the formation of functional literacy of students" / G. S. Kovaleva, N. I. Kolachev // Domestic and foreign pedagogy. – 2023. – T. 2. № 1 (90). – C. 9-32.
6. Poddyakov, A.N. Solving complex problems in PISA–2012 and PISA–2015: interaction with complex reality / A.N. Poddyakov // Educational Policy. – 2012. – № 6 (62). – C. 34-53.
7. Uglanova, I.L. Evidence-Centered Design method for measuring complex psychological constructs / I.L. Uglanova, I.V. Brun, G.M. Vasin // Modern Foreign Psychology. – 2018. – № 3 (7). – C. 18-27.
8. Shkerina, L. V. The meta-disciplinary Olympiad for school students: the new approach to assessing the meta-disciplinary universal educational actions of students / L. V. Shkerina, O. V. Berseneva, N. A. Zhuravleva, M. A. Cave // Perspectives of Science & Education. – 2019. –№ 2 (38). – C. 194-211.
9. Ahmed A. Improving the quality of contextualized questions: an experimental investigation of focus / A. Ahmed, A. Pollitt // Assessment in Education: Principles, Policy & Practice. – 2007. – № 2 (14). – P. 201–232.

10. Ananiadoui K., M. Claro. 21st Century Skills and Competences for New Millennium Learners in OECD Countries [Электронный ресурс]. Режим доступа: https://doi.org/10.1787/19939019 (дата обращения: 05.07.2024).
11. Andrews-Todd J. Virtual Performance-Based Assessments Methodology of Educational Measurement and Assessment // Computational psychometrics: New methodologies for a new generation of digital learning and assessment: With examples in R and Python / под ред. A. A. Von Davier, R. J. Mislevy, J. Hao. Springer. – 2021. – P. 45–60.
12. Andrews-Todd J. Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task / J. Andrews-Todd, C. M. Forsyth // Computers in human behavior. 2020. – № 104. – P. 105759.
13. Attali Y. The interactive reading task: Transformer-based automatic item generation / Y. Attali [et al.] // Frontiers in Artificial Intelligence. 2022. – №5. – P. 903077.
14. Barnett S. M. When and where do we apply what we learn?: A taxonomy for far transfer / S.M. Barnett, S.J. Ceci // Psychological bulletin. – 2002. – № 4 (128). – P. 612–637.
15. Bejar I. I. A generative approach to psychological and educational measurement // ETS Research Report Series. – 1991. – № 1. – P 1–54.
16. Bond L., Moss P., Carr P. Fairness in large-scale performance assessment // Technical issues in large-scale performance assessment. – 1996. – P. 117–140.
17. Braun H. I. Performance assessment of critical thinking: Conceptualization, design, and implementation / H.I. Braun [et al.] // Frontiers in Education. – 2020. – №5. URL: https://www.frontiersin.org/articles/10.3389/feduc.2020.00156/full
18. Brennan R. L. Generalizability theory / R.L. Brennan // Educational Measurement: Issues and Practice. 1992. – № 4 (11). – P. 27–34.
19. Brown T. A. Confirmatory factor analysis for applied research. / T.A. Brown. – New York: The Guilford Press, 2006.
20. Campbell D. T., Convergent and discriminant validation by the multitrait-multimethod matrix. / D.T. Campbell, D.W. Fiske // Psychological bulletin. – 1959. – № 2 (56). – P. 81–105.

21. Care E. Education System Alignment for 21st Century Skills: Focus on Assessment / E. Care, H. Kim, A. Vista, K. Anderson // Center for Universal Education at The Brookings Institution. – 2018.
22. Carneiro J. Proposal of a validation framework for a new measurement model and its application to the export performance construct / J. Carneiro, A. Rocha, J. F. Silva // BAR-Brazilian Administration Review. – 2009. – № 6. – P. 331–353.
23. Cho Y. Investigating the effects of prompt characteristics on the comparability of TOEFL iBTTM integrated writing tasks / Y. Cho, F. Rijmen, J. Novák // Language Testing. – 2013. – № 4 (30). – P. 513–534.
24. Clause C. S. Parallel test form development: A procedure for alternate predictors and an example / C.S. Clause // Personnel Psychology. – 1998. – № 1 (51). – P. 193–208.
25. Cronbach L. J. The Dependability of Behavioral Measurements. / L.G. Cronbach [et al]. Hoboken, NJ: Wiley, 1972.
26. Eid M. Measuring Method Effects: From Traditional to Design-Oriented Approaches / M. Eid, C. Geiser, T. Koch // Current Directions in Psychological Science. – 2016. – № 4 (25). – P. 275–280.
27. Ercikan, K. In search of validity evidence in support of the interpretation and use of assessments of complex constructs: Discussion of research on assessing 21st century skills. / K. Ercikan, M.E. Oliveri // Applied Measurement in Education. – 2016. – № 29(4). – P. 310-318.
28. Fay D. M. Investigating Psychometric Isomorphism for Traditional and Performance-Based Assessment / D.M. Fay, R. Levy, V. Mehta // Journal of Educational Measurement. – 2018. – № 1 (55). – P. 52–77.
29. Geisinger K. F. 21st Century Skills: What Are They and How Do We Assess Them? / K. F. Geisinger // Applied Measurement in Education. – 2016. – Vol. 29. № 4. – P. 245–249.
30. Gierl M. J., Haladyna T. M. Using weak and strong theory to create item models for automatic item generation: Some practical guidelines with examples // Automatic Item Generation / под ред. M. J Gierl, T. M. Haladyna. Routledge, 2012. P. 36–49.

31. Gierl M. J. Using Automatic Item Generation to Create Solutions and Rationales for Computerized Formative Testing / M.J. Gierl, H. Lai // Applied Psychological Measurement. – 2018. – № 1 (42). – P. 42–57.
32. Gipps C., Stobart G. Fairness in Assessment // Educational assessment in the 21st century / под ред. C. Wyatt-Smith, J. J. Cumming, Dordrecht: Springer Netherlands, 2009. P. 105–118.
33. Gorin J. S. Inherent measurement challenges in the next generation science standards for both formative and summative assessment / J. S. Gorin, R. J. Mislevy // Invitational research symposium on science assessment. – Citeseer, 2013.
34. Griffin P., Care E. Assessment and teaching of 21st century skills: Methods and approach / P. Griffin, E. Care, Springer, 2014.
35. Haladyna T. M. Item Shells: A Method for Writing Effective Multiple-Choice Test Items / T.M. Haladyna, R.R. Shindoll // Evaluation & the Health Professions. – 1989. – № 1 (12). – P. 97–106.
36. He J., Vijver F. van de Bias and equivalence in cross-cultural research / J. He, F. van de Vijver // Online readings in psychology and culture. – 2012. – № 2 (2). – P. 2307–0919.
37. Hyytinen H. How do self-regulation and effort in test-taking contribute to undergraduate students' critical thinking performance? / H. Hyytinen [et al.] // Studies in Higher Education. – 2024. – Vol. 49. № 1. – P. 192–205.
38. Irvine S. H., Dann P. L., Anderson J. D. Towards a theory of algorithm-determined cognitive test construction / S.H. Irvine, P.L. Dann, J.D. Anderson // British Journal of Psychology. – 1990. – № 2 (81). – P. 173–195.
39. Irvine S. H., Kyllonen P. C. Item generation for test development, Routledge, 2013.
40. Jiang Z. Bayesian approach to estimating variance components within a multivariate generalizability theory framework / Z. Jiang, W.A. Skorupski // Behavior Research Methods. – 2018. – № 6 (50). – P. 2193–2214.
41. Jorgensen T. D. How to estimate absolute-error components in structural equation models of generalizability theory / T.D. Jorgensen // Psych. – 2021. – № 2 (3). – P. 113–133.

42. Keller L. A. Using multivariate generalizability theory to assess the effect of content stratification on the reliability of a performance assessment / L.A. Keller, B.E. Clauser, D.B. Swanson // Advances in health sciences education. – 2010. – №15. – P. 717–733.
43. Kim E.S. Testing Measurement Invariance: A Comparison of Multiple-Group Categorical CFA and IRT / E.S. Kim, M. Yoon // Structural Equation Modeling: A Multidisciplinary Journal. 2011. – № 2 (18). – P. 212–228.
44. Kolen M. J. Threats to score comparability with applications to performance assessments and computerized adaptive tests / M.J. Kolen // Educational Assessment. – 1999. – № 2 (6). – P. 73–96.
45. Kroehne U. Construct Equivalence of PISA Reading Comprehension Measured With Paper-Based and Computer-Based Assessments / U. Kroehne [et al.] // Educational Measurement: Issues and Practice. – 2019. – № 3 (38). – P. 97–111.
46. Kunnan A. J. Test fairness / A.J. Kunnan // European language testing in a global context. – 2004. – №18. – P. 27–48.
47. Lee H.-K., Anderson C. Validity and topic generality of a writing performance test / H.-K. Lee, C. Anderson // Language testing. – 2007. – № 3 (24). – P. 307–330.
48. Levy R. Psychometric and Evidentiary Advances, Opportunities, and Challenges for Simulation-Based Assessment / R. Levy // Educational Assessment. – 2013. – № 3 (18). – P. 182–207.
49. Lievens F. Creating Alternate In-Basket Forms Through Cloning: Some preliminary results / F. Lievens, F. Anseel // International Journal of Selection and Assessment. – 2007. – № 4 (15). – P. 428–433.
50. Lievens F. Situational judgment tests in high-stakes settings: Issues and strategies with generating alternate forms / F. Lievens, P. R. Sackett // Journal of Applied Psychology. – 2007. – № 4 (92). – P. 1043–1055.
51. Messick S. The Interplay of Evidence and Consequences in the Validation of Performance Assessments / S. Messick // Educational Researcher. – 1994. – № 2 (23). V P. 13–23.

52. Mislevy R. J. Implications of Evidence-Centered Design for Educational Testing / R. J. Mislevy, G. D. Haertel // Educational Measurement: Issues and Practice. – 2006. – № 4 (25). – P. 6–20.
53. Nelson J. “I’d like to be under the sea”: Contextual cues in virtual environments influence the orientation of idea generation / J. Nelson, J. Guegan // Computers in Human Behavior. – 2019. – № 90. – P. 93–102.
54. Osburn H. G. Item Sampling for Achievement Testing / H.G. Osburn // Educational and Psychological Measurement. – 1968. – № 1 (28). – P. 95–104.
55. Pellegrino, J. W. Teaching, learning and assessing 21st century skills [Электронный ресурс].–2017. – P. 223 – 251. Режим доступа: https://doi.org/10.1787/9789264270695-12-en
56. Pitta-Pantazi D. Higher order thinking in mathematics: a complex construct / D. Pitta-Pantazi, P. Sophocleous // The 10th Mathematical Creativity and Giftedness. – Nicosia, Cyprus, 2017.
57. Reckase M. D. 18 Multidimensional Item Response Theory Psychometrics // Handbook of statistics / под ред. C. R. Rao, S. Sinharay, Elsevier, 2006. – № 26. – P. 607–642.
58. Ridley C. R. Multicultural Counseling Competence: A Construct in Search of Operationalization / C. R/ Ridley [et al.] // The Counseling Psychologist. 2021. – Vol. 49. – № 4. – P. 504–533.
59. Rijmen F. Formal Relations and an Empirical Comparison among the Bi-Factor, the Testlet, and a Second-Order Multidimensional IRT Model / F. Rijmen // Journal of Educational Measurement. 2010. – № 3 (47). – P. 361–372.
60. Rojas M. Assessing collaborative problem-solving skills among elementary school students / M. Rojas [et al.] // Computers & Education. – 2021. – №175. – P. 104313.
61. Ruiz-Primo M. A. The Relationship between Item Context Characteristics and Student Performance: The Case of the 2006 and 2009 PISA Science Items / M.A. Ruiz-Primo, M. Li // Teachers College Record. 2015. – № 1 (117) – P. 1–36

62. Sälzer C. Assessing global competence in PISA 2018: Challenges and approaches to capturing a complex construct / C. Sälzer, N. Roczen // International journal of development education and global learning. – 2018. – Vol. 10. – № 1. P 5–20.

63. Sayin A. Using OpenAI GPT to Generate Reading Comprehension Items / A. Sayin, M. Gierl //Educational Measurement: Issues and Practice. – 2024. – Vol. 43. – №. 1. – P. 5-18.

64. Schliemann A. D. Proportional reasoning: From shopping, to kitchens, laboratories, and, hopefully, schools / A.D. Schliemann, V.P. Magalhães // Oaxtepec Mexico, 1990. – P. 67–73.

65. Shavelson R. J. Sampling variability of performance assessments / R.J. Shavelson, G.P. Baxter, X. Gao // Journal of educational Measurement. – 1993. – № 3 (30). – P. 215–232.

66. Shavelson R. J., Webb N. M., Rowley G. L. Generalizability theory // Methodological issues & strategies in clinical research / под ред. E. Kazdin, American Psychological Association, 1992.

67. Stadler M. The assessment of collaborative problem solving in PISA 2015: An investigation of the validity of the PISA 2015 CPS tasks / M. Stadler [et al.] // Computers & Education. – 2020. – №157. – P. 103964.

68. Uglanova I. Computer-based performance approach for critical thinking assessment in children / I. Uglanova [et al.] // British Journal of Educational Psychology. – 2023. – №93. – P. 531–544.

69. Wang D. Automated and interactive game-based assessment of critical thinking / D. Wang, H. Liu, K.-T. Hau // Education and Information Technologies. – 2022. – № 4 (27). – P. 4553–4575.

70. Watson J. Statistical literacy: A complex hierarchical construct / J. Watson, R. Callingham // Statistics Education Research Journal. – 2003. – Vol. 2. – № 2. – P. 3–46.

71. Williamson D. M. Hierarchical IRT examination of isomorphic equivalence of complex constructed response tasks. / D.M. Williamson [et al.] // April Paper

presented at the annual meeting of the American Educational Research Association New Orleans, LA, 2002.