The USE Test in History and Its Validity: Experts of Regional Subject-Specific Committees Speculating on Free-Response Items

O. D. Fedorov, K. D. Verinchuk

Received in March 2021

Oleg Fedorov, Candidate of Sciences in History, Associate Professor, Director of the Siberian Institute of Management, Russian Presidential Academy of National Economy and Public Administration (RANEPA). Address: 6 Nizhegorodskaya St, 630102 Novosibirsk, Russian Federation. Email: fedorov-od@ranepa.ru (corresponding author)
Ksenia Verinchuk, History Teacher, European Gymnasium. Address: 28 Sokolnicheskiy Val St, 107113 Moscow, Russian Federation. Email: k.verinchuk@gmail.com

Abstract

The Unified State Exam (USE) in Russia is both an achievement and admission test, yet its validity has not been looked into on a large scale. The evolution of USE tests is distinctly marked by a growing number of constructed-response items, which might be affecting the validity of test results in many ways.

In-depth semi-structured interviews with 36 USE experts in History allow identifying three major threats to USE validity: assessment criteria for items 24 and 25, item content, and expert bias. Interview transcripts were analyzed using content analysis, the results of which are presented along with recommendations on how to further improve the processes of item design and evaluation.

Keywords

constructed-response items, free-response items, history performance assessments, subjectivity in assessment, testing, Unified State Exam (USE), validity.

For citing

Fedorov O., Verinchuk K. (2021) EGE po istorii i validnost' ego rezul'tatov: eksperty regional'nykh predmetnykh komissiy ob otkrytykh zadaniyakh [The USE Test in History and Its Validity: Experts of Regional Subject-Specific Committees Speculating on Free-Response Items]. *Voprosy obrazovaniya/Educational Studies Moscow*, no 3, pp. 189–211 https://doi.org/10.17323/1814–9545–2021–3–189–211

It took about a decade to design and try out the Unified State Exam (USE) before it was approved as the only form of school exit examinations and the main form of preliminary college entrance examinations in 2009.

The USE test in history is not obligatory, yet over 100,000 school leavers take it every year to apply for college programs in humanities, such as law, international relations, history, management, teaching, cultural studies, arts, etc.

Because the USE is a high-stakes exam, its validity should also be very high. Test validity is understood as the extent to which test score interpretations hold across settings [Messick 1995].

According to the USE History Test specifications, the test's items "measure school graduates' knowledge and skills in history in compliance with the requirements of the Federal Component of State Standards for Secondary Education". To measure USE validity, it is necessary to find out whether the test's form and content contribute to such measurement: if USE scores do reflect the level of course achievement, test validity can be regarded as high.

No additional validation studies were carried out after the USE tryout phase. However, the history test's content and form have changed a lot since it was formally introduced ten years ago. The recent years have seen an increase in the role of constructed-response (free-response) items, i. e. items which ask students to provide detailed written answers in free form and which are evaluated by over 5,000 raters of regional subject-specific committees.

The present article explores the validity of constructed-response items in the USE History Test.

1. Evolution of USE History Test Items and the Problem of Validity

After the USE had been introduced, history test items underwent major transformations, which were extensively described by researchers from the Federal Institute of Educational Measurements (FIEM) [Artasov, Melnikova 2018]. Within the scope of the present study, the following changes appear to be of the most importance.

First, the 2015 abolishment of the so-called Part A, which consisted of multiple-choice items with four answer options, dramatically increased test difficulty as well as the weight of constructed-response items in the total score.

Second, a history essay was introduced and a number of other constructed-response items were modified in 2016. In the current specifications (2020), 5 of the 25 items are free-response items, which account for 40% of the raw score.

Third, the USE evolved by gradually increasing the proportion of constructed-response items at the expense of multiple-choice items.

Multiple-choice items become less prevalent and more difficult, as they largely ask about historical nuances and details. This type of items is not equivalent to what is considered the basic level in the USE. Item's form gives no exact idea of its difficulty: it can be a "classical" item with four options, as in Section A before 2015, or a "hidden" multiple-choice task, as in B10-type items of the 2010 specification (Figures 1 and 2). Items 18 and 19 in the 2016–2019 USE specifications, which involve visual processing, are also examples of multiple-choice items.

Table 1 shows changes in the number of multiple-choice items over ten years, including percentages in the total number of items. However, the most important indicator is the contribution of these items

Figure 1. **A sample multiple-choice item from the 2009 USE demo test.**

A6 Which of the monuments listed below dates back to the 18th century?

1) Dormition Cathedral of the Moscow Kremlin;
2) Church of the Intercession on the Nerl River;
3) Peter and Paul Cathedral in Saint Petersburg;
4) Palace of Tsar Aleksey Mikhaylovich in Kolomenskoye.

Figure 2. **A sample multiple-choice item from the 2010 USE demo test.**

B10 Which three of the events listed below took place during Perestroika?
Please circle the figures corresponding to the right answers and write them down in the *table.*

1) Creation of the post of the President of the Soviet Union;
2) Repudiation of foreign and domestic debts (sovereign default);
3) Adoption of a new Constitution;
4) Proclamation of Russia's sovereignty;
5) Privatization;
6) GKChP's Declaration.

to the raw total score, which used to hover around 44% until the dramatic USE modification of 2016 brought it down to 17%. The most recent versions of the test offer only six multiple-choice items, and they are not those simple tasks that used to be part of Section A. Today, students are asked to select more than one correct answer options (Figure 3).

Moving away from multiple-choice items is easy to explain. Examinees may pick answers randomly. In a "classical" multiple-choice item above, the probability of guessing right is 25% (when choosing one out of four options). In Section A of the 2009–2014 specification or Section 1 of the 2015–2019 specification, chances of guessing are lower as multiple responses have to be selected. Being interested in minimizing the effects of guessing, test designers chose to reduce the number of multiple-choice items and increase their difficulty.

Figure 3. **A sample multiple-choice item from the 2017 USE demo test.**

3 All the terms below except *two* refer to events (phenomena) from the 19th century.

1) Free ploughmen; 2) Ministries; 3) The Decembrists;
4) Coup of June 3rd; 5) Magistrates' courts;
6) The Octobrist Party.

Please identify and write down the figures corresponding to the terms that refer to a different historical period..

Answer: [ ][ ]

Table 1. **The number of multiple-choice and constructed-response items in the USE History Test.**

	2009	2011	2013	2014	2015	Since 2016
Total items	50	49	37	40	40	25
Multiple-choice items						
Number of items	28 (A1–A27, B10)	30 (A1–A27, B2, B6, B12)	24 (A1–A21, B2, B4, B7)	24 (A1–A21, B2, B4, B7)	27 (1–21, 23, 25, 28, 32, 33–34)	6 (3, 7, 12, 16, 18–19)
Percent of total items	56%	61%	64%	64%	67%	24%
Maximum weight in the maximum total score	43%	44%	44%	44%	44%	17%
Constructed-response items						
Number of items	6 (C2–C7)	6 (C2–C7)	5 (C2–C6)	5 (C2–C6)	5 (36–40)	5 (21–25)
Percent of total items	12%	12%	12.50%	12.50%	20%	20%
Maximum weight in the maximum total score	28%	28%	28%	28%	28%	41%

Eliminating items of this type completely would hardly be reasonable—simple, easy-to-score tasks should also be present in the test. This is especially critical for the subject of history, where processes can barely be understood without knowing the fundamental facts. Situated at the beginning of the test, such items also have psychological effects, increasing students' self-efficacy and allowing low-performers to score at least the required minimum score.

1.1. Evolution of Constructed-Response Items over Ten Years

Some items in the USE History Test—items C1–C3 in the 2009–2014 versions, same as items 20–22 in the 2016–2019 versions—have not changed essentially. Those items imply analysis of a historical source

Figure 4. **Item 24 from the 2019 USE demo test.**

24 There are diverse and often controversial alternative views on historical events. One of such controversial views is presented below..

Alexander III's domestic policy contributed to progress in social and economic development.

Using your historical knowledge, give two arguments to support this standpoint and two arguments to oppose it.
While providing your arguments, make sure to use historical facts.
Please write down your response as follows:.

Supporting arguments:
1) ...
2) ...

Opposing arguments:
1) ...
2) ...

with the use of one's own historical knowledge, each of them yielding a maximum of two points.

Item C4, transformed into item 38 in 2015 and then into item 23 in 2016, underwent a number of modifications due to changes in the skill codifier. In 2009, this item measured "the set of knowledge and skills necessary to summarize and systematize historical materials" and was worth four points. In 2017, it assessed the "ability to apply the principles of structural functionalism and spatiotemporal analysis when examining facts, phenomena, and processes (problem-solving test item)", yielding three points. Structurally, the item has basically remained the same, asking students to write down three elements, e. g. three axes of foreign policy, three reasons to build a town on the bank of a river, etc.

Similar changes occurred to item C5, same as item 39 in the 2015 version and item 24 in the 2016–2020 versions. In any of its forms, this is a reasoning item. In the original version (2009–2012), examinees were asked to select one of the two standpoints and provide three arguments to defend it. Since 2013, the item has offered only one standpoint, asking students to give two arguments to support it and two arguments to oppose it (Figure 4). In that more complicated configuration, the item could yield a maximum of four points instead of three. According to FIEM reports, this item remains the most difficult one — year after year, regardless of the form.

Figure 5. **Item 25 from the 2020 USE demo test.**

25 Please write a history essay on ONE of the following periods in Russian history:

(1) 1019–1054; (2) March 1801—May 1812;
(3) October 1917—October 1922.

Please make sure to:

- Name at least two significant events (phenomena, processes) relevant to this historical period;
- Name two historical figures related to such events (phenomena, processes) and describe their roles in such events (phenomena, processes) using your knowledge of historical facts;

Note:
When describing the role of each historical figure you have named, you should specify exactly which of their actions essentially affected the course and/or outcome of the named events (phenomena, processes).

- Specify at least two causal relations that describe what caused the events (phenomena, processes) during that period;
- Evaluate the effects of the events (phenomena, processes) from that period on further historical developments in Russia using your knowledge of historical facts and/or historians' opinions.

It is essential to ensure adequate use of the historical terms and concepts relevant to the period.

Item C6 has changed the most radically of all. In the 2009–2011 test versions, it represented analysis of a historical situation with a maximum raw score of four points. In 2012–2015, students were asked to analyze the contributions of one of the three historical figures proposed (five points maximum). Since 2016, the item requires writing a history essay on one of the three historical periods (Figure 5). Detailed scoring criteria were developed for this new item format, allowing a maximum raw score of 11 points. The introduction of history essay and detailed scoring criteria became the key factor of change. As a result, item 25 in the current version has the highest weight in the test.

Item C7, which existed in the test specifications up to 2013, implied comparing two historical situations and filling out a table with common features and differences. This item was worth a maximum of four points. It explicitly measured students' ability to compare and under-

stand historical contexts, but for some reason it was removed from the test. As a result, the USE History Test became essentially less effective as a measure of competencies, as item C7 had involved identifying dimensions for comparison and performing the comparison. No similar task has been introduced in the more recent versions, which means that the USE has become a more knowledge-based assessment.

Therefore, the role of constructed-response items has changed considerably following the introduction of history essay on a particular period of Russian history. Items of this type are often used to achieve deeper measurements in educational assessment, as they allow capturing a broader scope of the construct measured. In this case, we can see attempts of test developers to define the scope of construct measurement through detailed assessment criteria for item 25, which were introduced in 2016, got extended in 2017, and have been elaborated yearly ever since.

All the items in Section II are scored independently by two raters. In case of essential rater disagreement (two or more points for items 20–24 and on some specific criteria for item 25), a third rater is called to take one of the sides or give a verdict of their own. In our study, the main focus is placed on items 24 and 25 in the current USE specification (2020): first, they account for the highest proportion of the raw score, largely predicting the scaled score, and second, only these two items demand writing an essay. Items 20–23 of Section II also imply free responses, yet shorter ones than in items 24 and 25.

1.2. Possible Effects of Item Change on Test Validity

Ample research has revealed considerable differences between multiple-choice and constructed-response items in how they measure the same construct [Traub, Fisher 1977; Ward, Frederiksen, Carlson 1980; Thissen, Wainer, Wang 1994; Lissitz, Hou, Slater 2012], especially in disciplines (such as history) where writing skills of the examinee matter [Traub 1993].

Those differences mostly derive from the specific characteristics of free-response items. First of all, they make guesswork nearly impossible, often happen to be more difficult than multiple-choice items, and are normally scored by raters based on specifically developed criteria [Haladyna, Rodriguez 2013]. Second, they are believed to allow deeper and more comprehensive measurements and thus increase content validity [Dennis, Newstead 1994], i. e. they can measure complex, multicomponent skills and competencies. Because of these qualities, constructed-response items are used in educational assessment despite the prevalence of multiple-choice items [Maris, Bechger 2006]. At the same time, almost all free-response items require using specific, clearly defined scoring criteria that are rather laborious to design, and should be scored by raters trained to apply such criteria. With such an assessment method, bias and ambiguity are inevitable, potentially increasing the costs of measurement [Arffman 2015]. Problems with scoring constructed-response items are particularly acute for the USE

History Test: historical events often receive diametrically opposed interpretations from different schools of thought, and raters may have divergent opinions on the test content. Under such circumstances, free-response item scoring criteria should be indisputable and unambiguous, yet guidelines for raters allow giving points for other meaningful responses on almost every item. These factors exacerbate the problem of item validity essentially.

A high degree of test validity suggests that advantages of various task formats are used in the most efficient way to minimize "construct-irrelevant variance" [Messick 1993]. It means that items should be selected and organized to measure the construct in the most comprehensive way by eliminating "noise".

For free-response items to be effective and contribute to test validity, it is necessary not only to pay attention to content validity of items as such but also to avoid three major types of mistakes when working with them [Popham 1990]:

- Problems associated with scoring criteria (too general, too specific, ambiguous, etc.);
- Problems associated with assessment procedures (raters reviewing tests for dozens of hours without rest, trying to consider too many criteria at the same time, etc.);
- Rater effects (raters not using the scoring criteria, being too stringent or too lenient).

Researchers distinguish between two major threats to construct validity: construct underrepresentation and construct-irrelevant variance [Messick 1995]. The latter arises from systematic error and makes the test measure not exactly (or anything but) what it was designed to measure (for a detailed taxonomy of such systematic errors in high-stakes testing, see [Haladyna, Downing 2005]).

The present article seeks to answer the following questions:

1. How did the USE History Test items transform, and how does their transformation affect test validity?
2. How is validity of the USE History Test related to constructed-response items?
3. What threats to validity are there in the test?
4. How does the test, in its current specification, deal with those threats?

Since 2016, when the current USE specification was introduced, no data on test validity has been published, and functioning of items has only been evaluated in part. Relevant literature is limited to test manuals and a few articles on the evolution of USE specifications published in *Pedagogicheskie Izmereniya/Educational Measurements* and describing the test content without providing a critical analysis of specific items or

scoring criteria. Given that transition to a new model of the USE is projected for 2022 to meet the new education standards, it would make sense to investigate validity of the existing items and give recommendations on how the test format could be improved.

2. Research Data and Methods

Any attempt to evaluate the USE History Test from a psychometric perspective is challenged by the lack of open data. Only some of the USE results are published in a very limited format, while previous versions and technical reports are not published at all. However, the FIEM website provides documents that allow looking through the test content. Available materials can be divided into the following categories: USE History Test specifications, USE demo tests in history, codifiers for the USE History Test, methodological recommendations for history teachers, and manuals for chairs and raters of regional subject-specific committees.

Specifications are reviewed on a yearly basis and provide the most valuable information on the current version of the test. This document provides information about the purpose and content of the test, the number of items and their difficulty, the scoring schemes, the duration of testing, and some other aspects.

A demo test is a sample test in its current version with responses. It is reviewed every year and gives an idea of how the actual test presented to examinees will look like and what kinds of tasks it will contain.

A codifier is a highly generalized list of topics that may be addressed in test items. In case of the history test, the codifier represents a chronological list of historical events, figures, periods, and phenomena covered by school history curriculum, basically a distilled summary of history curriculum guidelines from the previous generation.

Methodological recommendations for history teachers, issued by FIEM yearly after getting the test results, contain some information on item difficulty and test reliability (Cronbach's alpha) as well as a detailed analysis of the test results: how exactly students performed on specific items, which topics they stumbled upon, what mistakes prevailed this year, and other facts to which the test developers find it important to draw teachers' attention.

Manuals for chairs and raters of regional subject-specific committees are developed by FIEM counselors to guide raters on how to make a more effective use of the scoring criteria. In addition, manuals provide definitions of concepts involved in assessment (e. g. "historical event").

The great variety of test validity evaluation methods makes it hard to stick to any single standardized validation procedure—nearly any information about the test can be used for that purpose [Messick 1995]. Meanwhile, the type of arguments that can be given by researchers depends primarily on the type of the test evaluated [Haladyna 2006]. The USE History Test, in effect, combines two types of tests, being

used as a school exit examination and a college entrance examination at the same time.

Data on threats to test validity can be collected by scrutinizing the procedures of item scoring. Raters' scoring performance can be assessed through rater agreement indexes, participation in rater training, and third-party assessment of compliance with item scoring criteria.

We found it reasonable to interview the experts who have actually scored the USE History Test. Experts often play an important role in validation studies. As a rule, they are invited to judge item content and scoring criteria [Kane 2006]. It is especially important to involve experts who did not participate directly in item development, as their judgments will be much more objective than the opinions of test designers.

Experts taking part in scoring the USE History Test (raters) possess deep subject-specific and procedural knowledge and can be an important and reliable source of information about threats to validity. Their assistance is especially valuable in the absence of many other sources traditionally used in validation studies such as rater agreement indexes, test scores, and technical reports describing the stages of test content development.

Within the 2019/20 academic year, 36 interviews were carried out with experts (11 men and 25 women) who had participated in the activities of regional subject-specific committees in eight regions of Russia: Moscow, Saint Petersburg, Moscow Oblast, Kostroma Oblast, Novosibirsk Oblast, Chelyabinsk Oblast, Belgorod Oblast, and Kemerovo Oblast. The raters were on average 48 years of age, ranging from 28- to 68-year-olds, and had from six to 40 years of teaching experience. Only seven of the 36 raters had participated in USE scoring for less than five years. Many had been involved in scoring since the tryout phase, and eight were chairs of subject-specific committees in their regions.

Most raters were school history teachers, many of them were instructional designers and faculty members, and only five exclusively taught university courses on history teaching methodology. Of the 36 experts, 32 had a degree in teaching, 12 were candidates and doctors of sciences in education, and five were candidates and doctors of sciences in history.

Every year, raters should take a short qualification course and pass a test to be able to participate in USE scoring. Apart from that, raters attend webinars and offline meetings with test developers, in which they discuss different aspects of rating. The goal of such events is to reach agreement in rating approaches, but in practice rater disagreement remains at the same level year after year.

As part of our study, every rater completed a short online questionnaire about their age, education, and years of teaching, and then took part in a 30-minute audio-recorded Skype interview. Interviews were semi-structured and based on a guide of 15 questions. All the interview transcripts were analyzed using thematic content analysis which

involved coding every single mention of threats to validity. Taking cue from the available literature, five categories of threats to validity were defined, depending on whether they arise from inadequate scoring criteria, inadequate scoring procedures, rater effects, problematic responses, or item content.

3. Results and Discussion

Thematic content analysis of interview transcripts revealed 217 mentions of threats to validity. Threats associated with inadequate scoring criteria prevailed (97 mentions, or 44.7% of all mentions), being followed by threats arising from item content (62 mentions, or 28.5%). Threats related to rater effects were third by the frequency of mention (58 mentions, or 26.7%). Scoring procedures and problematic responses were mentioned rarely and did not play a significant role in this analysis.

3.1. Threats to Validity Arising from Scoring Criteria

Every rater complained about problematic scoring criteria for items 24 or 25, and most of them referred to scoring criteria as the main challenge in their work.

The scoring criteria for item 25 appear to be the most disputable, accounting for 61 of all the 97 mentions of inadequate criteria. All in all, there are seven item scoring criteria:

- Naming events from the historical period;
- The role of a historical figure in the period's events;
- Causal relations between events in the period;
- The impact of events on further historical developments;
- Adequate use of historical terms;
- No factual errors;
- Narrative style.

The latter two criteria are only evaluated if at least four points are given for the previous five. Only two criteria, "naming events from the historical period" and "narrative style", were reported by raters as unproblematic and easy to score.

Problems reported by raters are not always associated with how specific scoring criteria are formulated. Many raters pointed out that the criteria had been reviewed every year since the introduction of history essay, and not always in good time. As one of the interviewees said, "one of the problems is that the criteria are extended every year. <...> It would be great to work for at least three years without any additions or modifications." However, another modification was applied to the scoring criteria as well as to the format of this item in 2021.

Raters are unanimous in their negative attitude toward the "role of a historical figure in the period's events" criterion, which requires not just describing the role of a figure but specifying the figure's particular actions and how they changed the course of history. Item 25's requirements are regarded by raters as extremely stringent: "If a child

writes that Marshal Georgy Zhukov commanded the Assault on Berlin, they won't get a point for that, as they do not specify what his contribution was exactly. You don't know what he actually did there—maybe he just sat there sipping his tea. What should be written is that he designed a combat operation, issued an order to assault Berlin, and so on and so forth." This criterion is described in a number of interviews as overly specific, artificial, and absurd: "It requires too much details. If a student writes that Kutuzov commanded the Russian corps in the Battle of Borodino, does it mean they have a bad knowledge of history? No historian would ever apply criteria like that."

Obviously, when scoring item 25, raters need to examine the role of a historical figure and events in a very specific dimension, which they believe makes it impossible to assess a student's level of subject knowledge objectively. According to one of the raters, "a student can be drilled on that, but it will not reflect their knowledge of history, rather their ability to cut corners by using the clichés built in the criterion".

The way historical figures and their impact on events are interpreted within this item features "too many restrictions that make no allowance for intentional causality, being based exclusively on Soviet practices and objective causes, even though we all know that there's always the subjective factor in history and an individual's desire to do something is a cause, too".

Therefore, excessive fragmentation and stringency of the "role of a historic figure in the period's events" criterion leads to rater disagreement and the target construct being measured only partially.

Raters also complain a lot about the "causal relations between events in the period" criterion. The main problem with using it, as they claim, is that causal relations should be evaluated independently of the role of historical figures and any other events, which makes it difficult to decide on which criterion to give points. "Causal relations should be used to describe the role of a historical figure, but they are also used when analyzing historical events. There is no clear differentiation among these three criteria." Raters are required to strictly follow the order of criteria when scoring the items, which means that they have to reread responses over and over. This is how one of the raters puts it: "We have to raise this issue at our seminars again and again. There is always the question of whether these causal relations should be scored on the first, second, or third criterion." Therefore, the same sentence written by a student can serve as the grounds for giving a point on each of the three criteria at the same time, and the choice of methods to solve this problem is left to raters.

Another rater gives a real-life example: "Most often, students tend to associate the role of a historical figure with events that had some effects—and that makes causal relations. 'Olga carried out a tax reform <…> introduced *uroki* and *pogosti* to regularize the taxation system.' Where should this sentence be attributed? To the role of a historical fig-

ure or to causal relations?" Therefore, in raters' opinion, the insufficient differentiation of these three criteria poses serious threats to validity.

A different situation can be observed with the "adequate use of historical terms" and "no factual errors" criteria, on which there is no consensus among the raters: while some refer to them as problematic, others either consider them easy to evaluate or never mention them at all. The problem with these two is related to disagreement about how historical term and historical error should be understood.

Item 24 is judged by the majority of raters as difficult to score, accounting for 47 of the 97 mentions of inadequate scoring criteria. Scoring criteria for this item involve exact response formulations. First thing, the committee examines item 24 and scoring criteria in every test variant. Next, the committee works on "extending" the criteria in accordance with the instruction "other arguments are allowed" by suggesting various responses that students could provide as correct. Because item 24 may have no detailed meaningful scoring criteria in some variants, elaboration of those criteria is left to the discretion of regional committees, which may have a negative impact on rating outcomes, when identical scores in different regions indicate different levels of knowledge.

Raters also find it somewhat difficult to distinguish between facts and arguments provided by examinees: in compliance with the scoring criteria, facts alone are not enough for giving points. The lack of agreement about what should be regarded as a fact or an argument leads to divergences in scoring. According to one of the raters, "some facts are considered self-explanatory and don't have to be confirmed, while others should be linked to a theory that needs to be supported by evidence. This is also a game in a sense. If we build an argument, then we need arguments, and if we need facts, then we ask for facts."

A number of interviewees raised the question of inconsistency between the number of arguments required by item 24 and the number of points to be given for them. This item can yield a maximum of four points, but in case only two supporting or only two opposing arguments are provided, the student will be given only one point. Some raters, however, agree with this scoring scheme and argue that it serves well the purpose of the task, which is to measure the ability to see both sides of an issue. Others, meanwhile, find the scheme illogical and believe that each of the four arguments that the item asks for should be scored one point.

3.2. Threats to Validity Arising from Item Content

Nearly half of the raters pointed out that the title of item 25—History Essay—is inconsistent with the task formulation. As one of the raters said, "any essay is subjective, but this item is not a history essay. It's a description of a historical period, or an overview of a period—anything but an essay." Some of the raters believe that a history essay should demonstrate a student's personal attitude, opinion, and reasoning, which are not included in the scoring criteria for the current

version of item 25, so the genre of this item could rather be defined as description that follows a list of specific criteria.

Inconsistencies between instructional content and test content was mentioned by the majority of raters in some form or other. In particular, they maintain that the scoring criteria for item 25—the "role of a historical figure in the period's events" and "causal relations between events in the period"—are excessively specific. For instance, one of the raters said, with regard to the "role of a historical figure in the period's events" criterion: "Children read books, encyclopedias, and textbooks that describe roles of historical figures in a different manner. So, we get a double standard here: textbooks describe a historical figure in one way, but the test asks them to do it differently."

Many of the raters complain that hours allocated for teaching history are not enough to cover the whole history course comprehensively: "The historical-cultural standard provides for a huge number of didactic units. Way too many, to be honest. A lot of them just can't be crammed into the number of lessons—this is a basic exam, so only two lessons a week."

The revised Federal State Education Standard (FSES) for middle and high school education places a special focus on students' skills and competencies, including the ones in history. However, 29 of the 36 raters believe that the USE History Test only measures knowledge: "memory", "facts", "factual knowledge", etc. A lot of raters are concerned about the test being focused on measuring examinees' knowledge of historical facts instead of their competencies (universal learning activities and skills), for example in working with historical sources.

Overly specific USE requirements for knowledge about some historical periods leave a number of raters asking questions: "Which criteria are used for selection? Some periods are described in broad strokes, like Kievan Rus' before the 16th century, while the 1881–1893 events should be memorized nearly minute by minute. So, my questions are, why is that, and what's the purpose?"

Despite criticism of USE content, all the raters perceived positively the 2008–2020 changes to the specifications, first of all changes in the types of items and test materials: the introduction of visual sources, the elaboration of items and scoring criteria, and the abolishment of simple multiple-choice items.

3.3. Threats to Validity Arising from Rater Effects

Raters recognize subjectivity of their judgments and often report disagreement with colleagues as well as differences in rating approaches. All the raters agree that the same criteria can be interpreted in different ways, which leads to divergent scoring. Rater effects have several typical manifestations in USE scoring: stringency, lack of subject knowledge, leniency, and rater bias caused by the desire to avoid third-party scoring.

Some raters approach scoring too stringently. This is how one of the raters describes the use of the "no factual errors" criterion in item

25: “Some pay attention to inaccuracies, while others blame them on age and just write them off. I mean first of all stylistic mistakes, inaccuracies that may affect meaning.” Similar difficulties are experienced when applying the “role of a historical figure in the period’s events” criterion: “One rater will say the topic is covered well enough, and some other rater will not agree that two sentences suffice to describe Suvorov’s personality.”

Every rater being an expert in their own field, they cannot always work effectively with some topics: “It [item 24] allows other formulations, which means that every rater will decide for themselves whether or not to give a point based on their knowledge and outlook.” Raters often cannot remember all the nuances and details of historical material when scoring the tests—that is when students happen to be informed better than raters.

Most raters mentioned that they “have to score what they [students] wrote, but in an objective manner”, without “overinterpreting” the responses. However, some raters explicitly admit that they find it difficult to maintain such an attitude when working with test materials. Of the 36 raters, 30 repeatedly said in the interview that assessment should be learner-centered. They openly expressed concern for students and agreed that the latter should be given support in spite of the objectivity requirement. Raters involved in appeal processes were worried about the tacit ban on changing scores to the benefit of students.

The number of third-party scorings performed in case of essential rater disagreement on an item is regarded as an indicator of committee performance. If third-party experts are invited too often, requirements for the committee are toughened while its suggestions for improving the scoring criteria are not taken into consideration. Two thirds of the raters consider trying to avoid third-party scoring to be a barrier on the way toward objective evaluation: “Such concentration on third-party scoring leads to rater bias. In an effort to reach more agreement, one may either underscore or overscore.”

4. Conclusions and Recommendations

Thematic content analysis of interviews with USE raters revealed important threats to validity of the USE test in history.

In the first place, such threats arise from poorly formulated scoring criteria for free-response items. For example, the “role of a historic figure in the period’s events” and “causal relations between events in the period” criteria for item 25 were judged negatively by almost every rater. These criteria are often elaborated to minimize the probability of third-party scoring, but it only makes them overly stringent or formalized. “Attempts to increase rater agreement by using more objective scoring criteria will often lead to a narrowing of the factors included in the scoring, thereby increasing the risk posed by this threat to validity” [Crooks, Kane, Cohen 1996]. In our view, this is exactly what is happening to the “role of a historical figure in the period’s events” and “caus-

al relations between events in the period" criteria. Excessively detailed scoring criteria for item 25 and specifically worded criteria for item 24 result in non-objective evaluation of many students.

Raters are also highly critical of the test's content validity. In spite of the requirements set out in the Federal State Education Standard for high school education, history test items are still largely focused on measuring knowledge of historical facts, not competencies. The content and expected learning outcomes of school education have undergone substantial changes over the past five years, including the adoption of a new FSES version and the introduction of a historical-cultural standard. The USE History Test in its 2020 specification has not been adapted yet to meet all the new requirements. Test developers promise that this issue will be solved in the next specification that will focus more on competencies. However, even a brief analysis of the 2021 documents shows that threats to validity are still there.

Interviews also allowed identifying the main threats to validity arising from rater effects: excessive stringency or leniency, lack of subject knowledge, interpretation in favor of students, and bias caused by the desire to avoid third-party scoring. Such rater effects have been extensively studied in literature (for a detailed review of publications on rater effects, see, for instance, [Myford, Wolfe 2003]).

To summarize, the analysis performed in this study revealed some essential threats to validity of the USE test in history. To minimize those threats, it is necessary to improve the scoring criteria for certain items and pay attention to correspondence between test content and instructional content. How effectively these threats will be reduced in the new USE specification remains a question to be answered in the future.

5. Limitations

This study attempts to analyze threats to validity of test scores using thematic content analysis of interviews with test raters. Naturally, there is a number of limitations to this method.

First, all inferences are based exclusively on opinions expressed by raters, and the development of criteria for thematic analysis of interview transcripts, although based on relevant literature, essentially constituted a researchers' subjective action. The small size of the sample does not allow making large-scale inferences. At the same time, respondents' qualifications and experience support validity and reliability of the conclusions made.

Second, a few short interviews with raters are not enough to make strong inferences about test validity. A full-scale validation study of the USE test would require much more resources and a much broader range of methods including psychometric analysis of test items, factor analysis of USE scores in history, and probably an analysis of rater objectivity. It would make a lot of sense if test developers addressed this challenging responsibility in the future.

References

Arffman I. (2015) Threats to Validity when Using Open-Ended Items in International Achievement Studies: Coding Responses to the PISA 2012 Problem-Solving Test in Finland. *Scandinavian Journal of Educational Research*, vol. 60, no 6, pp. 609–625. doi:10.1080/00313831.2015.1066429.

Artasov I. A., Melnikova A. N. (2018) Evolyutsiya ekzamenatsionnykh modeley KIM EGE po istorii [The Evolution of History USE Models]. *Pedagogicheskie izmereniya*, no 2, pp. 48–56.

Crooks T. J., Kane M. T., Cohen A. S. (1996) Threats to Valid Use of Assessment. *Assessment in Education*, vol. 9, no 3, pp. 265–285. doi:10.1080/0969594960030302.

Dennis I., Newstead S. E. (1994) The Strange Case of the Disappearing Sex Bias. *Assessment & Evaluation in Higher Education*, vol. 19, no 1, pp. 49–56. https://doi.org/10.1080/0260293940190105.

Haladyna T. M., Downing S. M. (2005) Construct-Irrelevant Variance in High-Stakes Testing. *Educational Measurement: Issues and Practice*, vol. 23, no 1, pp. 17–27. doi:10.1111/j.1745–3992.2004.tb00149.x.

Haladyna T. M., Rodriguez M. (2013) *Developing and Validating Test Items*. London: Routledge. doi:10.4324/9780203850381.

Lissitz R., Hou X., Slater S. (2012) The Contribution of Constructed Response Items to Large Scale Assessment: Measuring and Understanding Their Impact. *Journal of Applied Testing Technology*, vol. 13, no 3, pp. 1–52.

Messick S. (1993) *Foundations of Validity: Meaning and Consequences in Psychological Assessment. ETS Research Report no RR-93–51*. http://dx.doi.org/10.1002/j.2333–8504.1993.tb01562.x.

Messick S. (1995) Validity of Psychological Assessment: Validation of Inferences from Persons Responses and Performances as Scientific Inquiry into Score Meaning. *American Psychologist*, vol. 50, no 9, pp. 741–749.

Miller M. D., Linn R. L., Gronlund N. E. (2009) *Measurement and Assessment in Teaching*. Upper Saddle River, NJ: Merrill/Pearson.

Myford C. M., Wolfe E. W. (2003) Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. *Journal of Applied Measurement*, vol. 4, no 4, pp. 386–422.

Popham W. J. (1990) *Modern Educational Measurement: A Practitioner's Perspective*. Englewood Cliffs, NJ: Prentice Hall.

Thissen D., Wainer H., Wang X. (1994) Are Tests Comprising Both Multiple-Choice and Free-Response Items Necessarily Less Unidimensional Than Multiple-Choice Tests? An Analysis of Two Tests. *Journal of Educational Measurement*, vol. 31, no 2, P. 113–123.

Traub R. E., Fisher C. W. (1977) On the Equivalence of Constructed-Response and Multiple-Choice Tests. *Applied Psychological Measurement*, vol. 1, no 3, pp. 355–369.

Traub S. L. (1993) Evaluating Medical Tests: Objective and Quantitative Guidelines. *American Journal of Health-System Pharmacy*, vol. 50, no 11, pp. 2440–2443.

Ward W. C., Frederiksen N., Carlson S. B. (1980) Construct Validity of Free-Response and Machine-Scorable Forms of a Test. *Journal of Educational Measurement*, vol. 17, no 1, pp. 11–29.