ЕГЭ по истории и валидность результатов:
эксперты региональных предметных комиссий об открытых заданиях

О. Д. Федоров, К. Д. Веринчук

Статья поступила в редакцию в марте 2021 г.

Федоров Олег Дмитриевич — кандидат исторических наук, доцент, директор Сибирского института управления РАНХиГС (г. Новосибирск). Адрес: 630102, Новосибирск, ул. Нижегородская, 6. E-mail: fedorov-od@ranepa.ru (контактное лицо для переписки)

Веринчук Ксения Дмитриевна — учитель истории Европейской гимназии (г. Москва). Адрес: 107113, Москва, Сокольнический Вал, 28. E-mail: k.verinchuk@gmail.com

Аннотация

ЕГЭ является одновременно выпускным и вступительным экзаменом, при этом его валидность масштабно не исследовалась. Ярко выраженная эволюция контрольно-измерительных материалов экзамена в направлении включения все большего количества заданий с открытым ответом выдвигает на первый план вопрос о связи между этими изменениями и валидностью тестовых результатов.

Проведены глубинные полуструктурированные интервью с 36 экспертами ЕГЭ по истории, которые позволили выделить три основные угрозы валидности ЕГЭ: критерии оценки заданий № 24 и № 25, содержание самих заданий, субъективность в оценках экспертов. Собранные материалы интервью подвергнуты содержательному анализу, результаты которого представлены вместе с рекомендациями по совершенствованию процесса создания и проверки заданий.

Ключевые слова

ЕГЭ, тестирование, валидность, контрольно-измерительные материалы по истории, задания со свободно конструируемым ответом, субъективизм оценок.

Для цитирования

Федоров О. Д., Веринчук К. Д. (2021) ЕГЭ по истории и валидность результатов: эксперты региональных предметных комиссий об открытых заданиях // Вопросы образования / Educational Studies Moscow. № 3. С. 189–211. https://doi.org/10.17323/1814-9545-2021-3-189-211

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

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 Str., 630102 Novosibirsk, Russian Federation. E-mail: fedorov-od@ranepa.ru (corresponding author)

Ksenia Verinchuk, History Teacher, European Gymnasium. Address: 28 Sokolnicheskiy Val Str., 107113 Moscow, Russian Federation. E-mail: 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. D., Verinchuk K. D. (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

Единый государственный экзамен находился в разработке и апробации около десяти лет — с начала 2000-х, а в 2009 г. был принят как основная форма выпускного экзамена в школах, совмещенного с вступительным экзаменом в высшие учебные заведения.

ЕГЭ по истории не является обязательным экзаменом, тем не менее каждый год его сдают более 100 тыс. школьников для поступления в вузы гуманитарной направленности — будущие юристы, международники, историки, специалисты в области управления, педагоги, культурологи, искусствоведы и др.

Учитывая, насколько высоки ставки, требования к валидности результатов ЕГЭ также должны быть очень высокими. Под валидностью тестовых результатов подразумевается обоснованность интерпретации этих результатов в том или ином контексте [Messick, 1995].

Согласно спецификации ЕГЭ по истории контрольно-измерительные материалы (КИМ) экзамена позволяют «установить уровень освоения выпускниками знаний и умений по курсу истории в соответствии с требованиями Федерального компонента государственных стандартов среднего (полного) общего образования». Чтобы оценить валидность ЕГЭ, необходимо выяснить, поддерживают ли содержание и форма теста такой результат. Если баллы школьника в ЕГЭ действительно позволяют судить о степени усвоения школьной программы, валидность теста можно считать высокой.

После апробации ЕГЭ дополнительных валидизационных исследований не проводилось. Однако за десять лет с момента официального введения ЕГЭ по истории его содержание и форма существенно изменились. На протяжении нескольких лет можно наблюдать тенденцию к увеличению роли заданий со свободно конструируемым ответом, т. е. заданий, на которые учащийся отвечает письменно, развернуто и в свободной форме и которые проверяют эксперты региональных предметных комиссии совокупной численностью более 5 тыс. человек.

Настоящая статья посвящена изучению валидности контрольно-измерительных материалов ЕГЭ по истории в части заданий со свободно конструируемым ответом.

1. Эволюция КИМов ЕГЭ по истории и проблема валидности

После внедрения ЕГЭ в КИМах по истории произошли масштабные трансформации, обстоятельно описанные сотрудниками ФИПИ [Артасов, Мельникова, 2018]. В контексте данного исследования наиболее важными представляются несколько изменений.

- Во-первых, отмена в 2015 г. так называемой части А, которая содержала задания с выбором одного варианта ответа из четырех предложенных, серьезно усложнила экзамен, а также увеличила удельный вес баллов по заданиям со свободно конструируемым ответом в общей сумме баллов.
- Во-вторых, в 2016 г. в КИМы было введено историческое сочинение и произошла модификация других заданий открытого типа. В текущей спецификации ЕГЭ (2020 г.) заданий со свободно конструируемым ответом 5 из 25, и их вклад в первичный балл испытуемого — около 40%.
- В-третьих, эволюция ЕГЭ состояла в постепенном уходе от заданий с множественными вариантами ответа и увеличении числа заданий с открытым ответом.

Заданий с множественными вариантами ответа становится меньше, а сами они становятся сложнее, поскольку ориенти-

Рис. 1. **Пример задания с выбором ответа в демоверсии ЕГЭ 2009 г.**

А6 Какой из перечисленных архитектурных памятников относится к XVIII в.?

1) Успенский собор Московского Кремля
2) церковь Покрова на Нерли
3) Петропавловский собор в Санкт-Петербурге
4) царский дворец в Коломенском

Рис. 2. **Пример задания с выбором ответа в демоверсии ЕГЭ 2010 г.**

В10 Какие три из перечисленных событий произошли в период перестройки?
Обведите соответствующие цифры и запишите их в *таблицу*

1) учреждение должности Президента СССР
2) отказ государства от оплаты внешних и внутренних долгов (дефолт)
3) принятие новой Конституции
4) провозглашение суверенитета России
5) проведение приватизации
6) выступление ГКЧП

руются в большей степени на исторические тонкости и детали. Тип заданий с выбором ответа не идентичен тому, что в ЕГЭ называется базовым уровнем. Сама форма задания не говорит однозначно о его трудности: это может быть как «классическое» задание с четырьмя вариантами ответа в части А до 2015 г., так и «скрытое» задание с выбором ответа в заданиях типа В10 в части В спецификации 2010 г. (рис. 1 и 2). Задания № 18 и № 19 в спецификациях ЕГЭ 2016–2019 гг., которые предполагают работу с визуальными источниками, — это тоже примеры задания с выбором ответа.

В табл. 1 представлены изменения в количестве заданий с выбором ответа на протяжении 10 лет. Приведены как коли-

Рис. 3. **Пример задания с выбором ответа в демоверсии ЕГЭ 2017 г.**

3 Ниже приведен список терминов. Все они, за исключением *двух*, относятся к событиям (явлениям) XIX в.

1) вольные хлебопашцы; 2) министерства;
3) декабристы; 4) третьеиюньский переворот;
5) мировые судьи; 6) октябристы.

Найдите и запишите *порядковые номера* терминов, относящихся к другому историческом у периоду.

Ответ: ☐☐

чество заданий, так и их доля от общего числа. Но наиболее важный показатель — это вклад, который выполнение этих заданий вносит в первичный итоговый балл. До 2016 г. он оставался более или менее постоянным — около 44%. После кардинального изменения спецификации ЕГЭ по истории в 2016 г. вклад заданий с выбором ответа в итоговый балл сократился до 17%. В последних версиях ЕГЭ заданий с выбором ответа осталось всего шесть, и это уже не те простые задания, которые присутствовали в части А раньше. Теперь в каждом случае испытуемый должен выбрать более одного правильного ответа из предложенных (рис. 3).

Уход от заданий с выбором ответа объясним. Испытуемый может выбирать ответ наугад. В случае «классического» задания, приведенного выше, шанс наугад ответить правильно составляет 25% (при выборе одного из четырех вариантов ответа). В заданиях части В 2009–2014 гг. или части I 2015–2019 гг. вероятность угадать уже меньше, так как выбрать нужно несколько вариантов. Составители тестов, будучи заинтересованными в минимизации угадывания, пошли по пути сокращения количества заданий с выбором ответа и усложнения оставшихся вопросов.

Совсем отказываться от заданий такого типа вряд ли рационально, так как тест нуждается и в простых, легко проверяемых заданиях. Особенно это важно для такого предмета, как история, где едва ли возможно понимание процессов без знания базовой фактологии. Расположенные в начале теста, такие задания имеют и психологическое значение: они повышают уверенность экзаменуемого и позволяют школьникам с довольно низким уровнем знаний набрать хотя бы минимальный балл.

Таблица 1. **Количество заданий с выбором ответа и заданий со свободно конструируемым ответом в ЕГЭ по истории**

	2009 г.	2011 г.	2013 г.	2014 г.	2015 г.	С 2016 г.
Общее количество заданий	50	49	37	40	40	25
Выбор ответа						
Количество заданий	28 (А1–А27, В10)	30 (А1–А27, В2, В6, В12)	24 (А1–А21, В2, В4, В7)	24 (А1–А21, В2, В4, В7)	27 (1–21, 23, 25, 28, 32, 33–34)	6 (3, 7, 12, 16, 18–19)
Доля от общего числа заданий, %	56	61	64	64	67	24
Доля максимального балла за эти задания от общего максимального балла, %	43	44	44	44	44	17
Открытая форма						
Количество заданий	6 (С2–С7)	6 (С2–С7)	5 (С2–С6)	5 (С2–С6)	5 (36–40)	5 (21–25)
Доля от общего числа заданий, %	12	12	12,5	12,5	20	20
Доля максимального балла за эти задания от общего максимального балла, %	28	28	28	28	28	41

1.1. Эволюция открытых заданий за десять лет

В составе КИМов есть задания, которые не претерпели серьезных изменений: это задания С1–3 в версиях ЕГЭ 2009–2014 гг., они же задания № 20–22 в версиях ЕГЭ 2016–2019 гг. Эти задания предполагают анализ исторического источника с привлечением собственных знаний. За каждое из них максимально начисляется 2 балла.

Задание С4, преобразованное в 2015 г. в задание № 38 и ставшее в 2016 г. заданием № 23, подвергалось неоднократной переработке в связи с изменением кодификатора умений. В 2009 г. это задание проверяло «комплекс знаний и умений к заданию на обобщенную характеристику, систематизацию исторического материала» и оценивалось 4 баллами, а в 2017 г. оно выявляло «умение использовать принципы структурно-функционального, временно́го и пространственного анализа при рассмотрении фактов, явлений, процессов (задание-задача)» и измерялось 3 баллами. Форма задания мало изменилась: в ответе нужно указать три элемента, например три направления внешней политики, три причины строить город на берегу реки.

Похожие изменения произошли с заданием С5, оно же задание № 39 в версии 2015 г. и задание № 24 в версиях 2016–2020 гг.

Рис. 4. **Задание № 24 из демоверсии ЕГЭ по истории 2019 г.**

24 По историческим вопросам высказываются различные, часто противоречивые точки зрения. Ниже приведена одна из противоречивых точек зрения.

Внутренняя политика Александра III способствовала прогрессивному развитию социальной и экономической сфер общественной жизни.

Используя исторические знания, приведите два аргумента, которыми можно подтвердить данную точку зрения, и два аргумента, которыми можно ее опровергнуть. При изложении аргументов обязательно используйте исторические факты.
Ответ запишите в следующем виде.

Аргументы в подтверждение:
1) …
2) …

Аргументы в опровержение:
1) …
2) …

В любой форме это задание на аргументацию. В первоначальном варианте (2009–2012 гг.) испытуемый должен был выбрать одну из двух предложенных точек зрения и обосновать ее, приведя три аргумента в защиту своей позиции. С 2013 г. экзаменуемому давалась одна точка зрения и предлагалось привести два аргумента в ее защиту и два аргумента в опровержение (рис. 4). Усложненное таким образом задание уже оценивалось максимально не тремя баллами, а четырьмя. Судя по публикуемым ФИПИ материалам, из года в год это задание в любой форме остается одним из самых трудных для испытуемых.

Наиболее радикальным изменениям подверглось задание С6. В версиях ЕГЭ 2009–2011 гг. оно представляло собой анализ исторической ситуации, который максимально оценивался в 4 первичных балла. В 2012–2015 г. школьникам предлагалось проанализировать деятельность одного из трех предложенных исторических деятелей (максимальный балл — 5), а с 2016 г. требуется написать историческое сочинение по одному из трех исторических периодов (рис. 5). К этому новому формату задания разработаны подробные критерии оценки, по которым макси-

Рис. 5. **Задание № 25 в демоверсии ЕГЭ 2020 г.**

25 Вам необходимо написать историческое сочинение об ОДНОМ из периодов истории России:

1) 1019–1054 гг.; 2) март 1801 — май 1812 г.;
3) октябрь 1917 — октябрь 1922 г.

В сочинении необходимо:

– указать не менее двух значимых событий (явлений, процессов), относящихся к данному периоду истории;
– назвать две исторические личности, деятельность которых связана с указанными событиями (явлениями, процессами), и, используя знание исторических фактов, охарактеризовать роли названных Вами личностей в этих событиях (явлениях, процессах);

Внимание!
При характеристике роли каждой названной Вами личности необходимо указать конкретные действия этой личности , в значительной мере повлиявшие на ход и (или) результат указанных событий (процессов, явлений).

– указать не менее двух причинно-следственных связей, характеризующих причины возникновения событий (явлений, процессов), происходивших в данный период;
– используя знание исторических фактов и (или) мнений историков, оценить влияние событий (явлений, процессов) данного периода на дальнейшую историю России.

В ходе изложения необходимо корректно использовать исторические термины, понятия, относящиеся к данному периоду.

мально можно получить 11 первичных баллов. Именно введение исторического сочинения и детализированных критериев его оценки стало ключевым фактором перемен. В данный момент № 25 является самым значительным по весу среди заданий теста.

Задание С7, существовавшее в спецификации ЕГЭ по истории до 2013 г., предполагало сравнение двух исторических ситуаций и заполнение таблицы с общими элементами и раз-

личиями. Максимально за это задание можно было получить 4 балла. Оно в явном виде проверяло навыки сравнения и понимания исторического контекста, но от него по каким-то причинам было решено отказаться. В результате тест существенно потерял в ориентации на проверку компетенций: это задание предполагало самостоятельное определение линий сравнения, а также осуществление данного сравнения. Аналогичного задания в более поздних версиях не появилось, т. е. в ЕГЭ усилилась знаниевая ориентация.

Таким образом, роль заданий со свободно конструируемым ответом существенно изменилась благодаря введению исторического сочинения по конкретному периоду российской истории. Задания такого типа часто применяются для более глубокого измерения в образовательном тестировании, они позволяют шире «охватить» выбранный конструкт. В данном случае мы видим попытки разработчиков определить сферу измерения конструкта через подробные критерии задания № 25, которые были введены в 2016 г. и расширены в 2017 г., а затем ежегодно уточнялись.

Все задания части II проверяют независимо друг от друга два эксперта, при наличии существенного расхождения между их оценками (более 1 балла в заданиях № 20–24, а также по определенным критериям задания № 25) работа передается третьему эксперту, который соглашается с позицией одного из экспертов либо выносит собственный вердикт. В нашем исследовании основное внимание уделяется заданиям № 24 и № 25 в текущей спецификации ЕГЭ (2020 г.): во-первых, они вносят наибольший вклад в первичный балл экзаменуемого, во многом определяя итоговый балл; во-вторых, только в этих заданиях требуется развернутый письменный ответ ученика. Задания № 20–23 части II тоже предполагают открытый ответ, но более краткий, чем требуется в заданиях № 24 и № 25.

1.2. Возможные последствия изменений КИМов для валидности интерпретации результатов

Многочисленные исследования выявили значительную разницу в результатах измерения одного и того же конструкта посредством заданий с выбором ответа и заданий со свободно конструируемым ответом [Traub, Fisher, 1977; Ward, Frederiksen, Carlson, 1980; Thissen, Wainer, Wang, 1994; Lissitz, Hou, Slater, 2012], особенно в тех дисциплинах, где значимы письменные навыки испытуемого, в частности в истории [Traub, 1993].

Главным образом эти различия обусловлены особенностями заданий с открытым ответом. Во-первых, они почти исключают возможность угадывания, часто бывают сложнее заданий с выбором ответа и, как правило, оцениваются экспертами по специально созданным критериям [Haladyna, Rodriguez, 2013]. Во-вторых, принято считать, что они помогают сделать изме-

рение в тесте более основательным и глубоким, повышая контентную валидность [Dennis, Newstead, 1994], т. е. с помощью таких заданий можно измерять сложные, многосоставные навыки и умения. Именно из-за этих свойств задания со свободно конструируемым ответом используются в образовательном тестировании, несмотря на доминирование заданий с множественными вариантами ответа [Maris, Bechger, 2006]. С другой стороны, почти все подобные задания подразумевают наличие специфических, точно прописанных критериев оценки, которые довольно сложны в разработке. Для их оценки нужны эксперты, обученные работать с этими критериями. При таком способе оценивания неизбежна субъективность, неоднозначность результатов тестирования, и цена такого оценивания потенциально более высока [Arffman, 2015]. Проблемы оценивания свободно конструируемых ответов особенно остры именно в ЕГЭ по истории — дисциплине, в которой присутствуют научные школы, зачастую диаметрально противоположно трактующие одни и те же исторические события, а эксперты, оценивающие тестовые задания, могут расходиться во мнениях относительно содержания экзаменационного материала. В таких условиях критерии оценки заданий со свободно конструируемым ответом должны быть непреложны и не иметь двояких толкований, однако в рабочих материалах эксперта практически ко всем заданиям для проверки содержится указание, что могут засчитываться и иные верные по смыслу ответы. Эти обстоятельства существенным образом обостряют проблему валидности КИМов.

Высокий уровень валидности теста подразумевает, что мы наиболее эффективным образом используем достоинства разных форматов заданий для максимизации релевантной дисперсии тестовых результатов [Messick, 1993]. Это означает, что задания должны быть подобраны таким образом, чтобы наиболее полно измерить предлагаемый конструкт, сведя уровень «шума» к нулю.

Чтобы задания со свободно конструируемым ответом хорошо работали и положительно влияли на валидность тестовых баллов, нужно не только обратить внимание на конструктную валидность самих заданий, но и избежать трех основных типов ошибок в работе с ними [Popham, 1990]:

- связанных с критериями оценивания (слишком общие или слишком подробные критерии, непонятные и т. п.);
- связанных с процедурой оценивания (эксперты проверяют несколько дней без отдыха, пытаются учесть слишком много критериев одновременно и т. п.);
- связанных с самими экспертами (не следуют критериям, слишком строгие или слишком снисходительные).

В литературе выделяются два типа угроз валидности: неполная представленность измеряемого конструкта в тесте (*construct underrepresentation*) и дисперсия оценки, не связанная с измеряемым конструктом (*construct-irrelevant variance*) [Messick, 1995]. Последняя представляет собой систематическую ошибку тестирования, которая приводит к тому, что созданный тест измеряет не совсем (или совсем не) то, что подразумевали разработчики (подробную таксономию таких систематических ошибок для тестов с высокими ставками см. в [Haladyna, Downing, 2005]).

В данной работе мы отвечаем на следующие исследовательские вопросы.

1. Каким образом происходила трансформация КИМов ЕГЭ по истории и как она влияет на валидность теста?
2. Как связана валидность ЕГЭ по истории с заданиями со свободно конструируемым ответом?
3. Какие угрозы валидности присутствуют в экзамене?
4. Как экзамен в его текущей спецификации справляется с этими угрозами?

За период действия текущей спецификации ЕГЭ (с 2016 г.) не публиковалось данных об исследовании валидности этого теста, а функционирование заданий в нем оценивалось лишь частично. Из публикаций по теме доступны только методические рекомендации создателей, а также несколько статей в журнале «Педагогические измерения» об эволюции спецификации экзамена, описывающие содержание заданий без критического анализа этих заданий или критериев их оценки. Учитывая, что в 2022 г. предполагается переход на новую модель ЕГЭ в связи с новыми образовательными стандартами, представляется целесообразным исследовать валидность текущих заданий и дать рекомендации для улучшения тестового формата.

2. Данные и методы исследования

Любая попытка оценить ЕГЭ по истории в психометрической перспективе сталкивается со сложностями, обусловленными нехваткой открытых данных об этом экзамене. Результаты сдачи ЕГЭ публикуются только частично и в очень ограниченном формате, варианты прошлых лет и технические отчеты официально не публикуются вообще. Однако сайт ФИПИ предлагает документы, позволяющие ознакомиться с содержанием теста. Доступные материалы можно разделить на следующие категории: спецификации ЕГЭ по истории, демоверсии ЕГЭ по истории, кодификаторы ЕГЭ по истории, методические рекомендации для учителей истории, методические рекомендации для председателей и экспертов региональных предметных комиссий.

Спецификация дает самую важную информацию о текущей версии теста и обновляется каждый год. Этот документ содержит информацию о цели экзамена, его содержании, количестве и уровне сложности заданий, системе начисления баллов, времени проведения экзамена и некоторых других характеристиках теста.

Демоверсия представляет собой пробный бланк текущей версии экзамена с ответами на задания и обновляется каждый год. Она дает представление о том, как выглядит тест, предъявляемый испытуемым, и о содержании заданий в тесте.

Кодификатор — это перечень тем, которые могут содержаться в заданиях, в крайне обобщенных формулировках. Для экзамена по истории это хронологический список событий, личностей, эпох и явлений, предлагаемых к изучению, фактически краткая выжимка из примерной основной образовательной программы по истории предыдущего поколения.

В методических рекомендациях для учителей истории, которые выпускает ФИПИ каждый год после проверки текущего экзамена, содержится частичная информация о трудности заданий и надежности теста (альфа Кронбаха) и подробный разбор итогов экзамена: как именно справились экзаменуемые с теми или иными заданиями, какие темы для них оказались сложными, какие ошибки были наиболее распространенными в этом году; приводятся и другие факты, на которые составители считают нужным обратить внимание педагогов.

Методические рекомендации для председателей и экспертов региональных предметных комиссий составляют методисты ФИПИ. В них экспертам разъясняют, как лучше работать с критериями. Дополнительно в рекомендациях приводятся определения тех или иных понятий, которые требуются для проверки (например, «исторические событие»).

Многообразие способов оценки валидности интерпретации тестовых баллов не позволяет однозначно выделить какую-либо стандартизированную процедуру валидизации. Практически любая информация о тесте может быть использована для этой цели [Messick, 1995]. При этом выбор типа доказательств, которые может привести исследователь, зависит в первую очередь от того, с каким тестом он работает [Haladyna, 2006]. В ЕГЭ по истории совмещены два типа теста, поскольку он одновременно является выпускным испытанием в школе и вступительным экзаменом в вуз.

Данные о том, какие угрозы валидности присутствуют в экзамене, могут быть собраны путем детального изучения таких процедур, как проверка заданий и начисление баллов. Деятельность экспертов по начислению баллов можно оценить через индекс согласия экспертов, через знакомство с процеду-

рой обучения и тренировки экспертов, через внешнюю оценку критериев задания.

Мы сочли целесообразным обратиться напрямую к экспертам, участвующим в проверке ЕГЭ по истории. Эксперты часто играют важную роль в валидизационных исследованиях. Как правило, их привлекают для оценки содержания заданий, а также для оценки критериев проверки этих заданий [Kane, Crooks, Cohen, 2005]. Особенно важно работать с экспертами, которые непосредственно не принимали участия в составлении заданий, так как их оценка будет более объективной, чем мнение разработчиков теста.

Эксперты, принимающие участие в проверке ЕГЭ по истории, обладают глубокими предметными и процедурными знаниями и могут служить важным и надежным источником информации об угрозах валидности. Их помощь представляется особенно ценной в отсутствие многих других источников, традиционно используемых в валидизационных исследованиях, таких как индекс согласия экспертов, результаты испытуемых, технические отчеты с описанием этапов содержательной разработки теста.

В течение 2019/2020 учебного года проведены 36 интервью с экспертами (11 мужчин и 25 женщин), принимающими участие в работе региональных предметных комиссий 8 российских регионов: Москвы, Санкт-Петербурга, Московской, Костромской, Новосибирской, Челябинской, Белгородской и Кемеровской областей. Педагогический стаж привлеченных экспертов — от 6 до 40 лет, средний возраст — 48 лет с разбросом от 28 до 68 лет. Только 7 из 36 экспертов принимали участие в проверке экзамена менее пяти лет, многие были вовлечены в проверку еще на стадии апробации ЕГЭ, некоторые являются руководителями предметных комиссии в своих регионах (8 человек).

Большинство экспертов — школьные преподаватели истории, многие из них являются методистами и преподают в вузах, и только пять экспертов занимаются исключительно преподаванием методики преподавания истории в университете. Среди 36 экспертов педагогическое образование имеют 32 человека, 12 — кандидаты и доктора педагогических наук, 5 — кандидаты и доктора исторических наук.

Каждый год эксперты обязаны пройти небольшие квалификационные курсы и сдать зачет, чтобы иметь возможность принять участие в оценке ЕГЭ. Помимо этого, для экспертов проводятся вебинары и очные встречи с разработчиками заданий, на которых они обсуждают различные аспекты проверки. Перед этими мероприятиями стоит цель достичь консенсуса в подходах к проверке, однако на деле рассогласованность экспертов из года в год остается на одном и том же уровне.

В рамках нашего исследования каждый эксперт заполнил краткий онлайн-опросник о своем возрасте, образовании и педагогическом стаже, а затем поучаствовал в 30-минутном скайп-интервью с обязательной аудиозаписью. Интервью носили полуструктурированный характер и проводились по гайду из 15 вопросов. Расшифровки каждого интервью подверглись тематическому анализу, в ходе которого кодировались все упоминания об угрозах валидности. В соответствии с литературными данными выделены пять категорий угроз валидности:

1) связанные с некорректными критериями проверки заданий;
2) связанные с некорректными процедурами проверки заданий;
3) связанные с экспертами;
4) связанные с проблематичными ответами учеников;
5) связанные с предметным содержанием заданий.

3. Результаты и их обсуждение

Тематический анализ расшифровок выявил 217 упоминаний об угрозах валидности. Из них наиболее частотной темой оказались угрозы, связанные с некорректными критериями оценки заданий (97 упоминаний, т. е. 44,7% всех упоминаний). На втором месте оказались упоминания угроз, связанных с содержанием заданий (62 упоминания, 28,5%). Третьими по частоте упоминания оказались угрозы, связанные с работой самих экспертов (58 упоминаний, 26,7%). Процедуры и проблематичные ответы учеников упоминались редко и не играли существенной роли в данном анализе.

3.1. Угрозы валидности, связанные с критериями оценки заданий

Все без исключения эксперты упоминали проблемные критерии оценки заданий № 24 или № 25, и большая часть экспертов указали критерии проверки в качестве основной сложности, с которой они сталкиваются в работе.

Больше всего вопросов вызывают критерии оценки задания № 25: из 97 упоминаний некорректных критериев к нему напрямую относится 61 упоминание. Всего критериев семь:

- указание событий периода;
- роль исторической личности в событиях периода;
- причинно-следственные связи между событиями периода;
- оценка влияния событий на дальнейшую историю;
- корректное использование исторической терминологии;
- отсутствие фактических ошибок;
- форма изложения.

Последние два критерия оцениваются только в том случае, если по остальным критериям набрано не меньше 4 баллов. Только

два критерия были оценены большинством экспертов как легкие в оценке и не представляющие сложностей: указание событий периода и форма изложения.

Проблемы, которые отмечают эксперты, не всегда связаны с формулировкой конкретных критериев. Многие эксперты обратили внимание на то, что критерии уточняются каждый год с момента ввода исторического сочинения, и не всегда своевременно. По словам одного из респондентов, «одна из сложностей заключается в том, что каждый год в критерии вносятся какие-то дополнения. <…> Очень бы хотелось поработать хотя бы года три, чтобы в критерии не вносились никакие изменения и никакие дополнения». Однако в 2021 г. критерии подверглись очередной трансформации, как и формат данного задания.

Единогласно отрицательно оценивают эксперты критерий «роль исторической личности в событиях периода», согласно которому требуется не просто охарактеризовать роль личности, но и описать ее в конкретных действиях и раскрыть влияние этих действий на ход событий. Требования к этому заданию эксперты называют очень строгими: «Если ребенок пишет, что маршал Г. К. Жуков руководил штурмом Берлина, то это не оценивается баллом, так как это не конкретный его вклад в событие. Ведь не указано, чем конкретно он там занимался. Может быть, он просто сидел и чай пил. А нужно писать, что он разработал операцию, издал приказ о штурме Берлина, и т. д., и т. п.». Данный критерий во многих интервью описывается как слишком подробный, искусственный, абсурдный: «Присутствует излишняя детализация. Если ученик написал, что Кутузов командовал русскими войсками в Бородинской битве, разве это значит, что ученик плохо знает историю? Ни один историк такими критериями не будет оперировать».

Очевидно, что в оценке задания № 25 от эксперта требуется рассматривать роль личности и события в очень специфическом измерении, которое, по мнению экспертов, не позволяет объективно оценить предметный уровень ученика. По словам одного из экспертов, «ученика можно натаскать на это, но это говорит не о его знании истории, а об умении вывернуться, то есть использовать заложенные в критерии языковые клише».

В трактовке исторических личностей и их влияния на события, которая принята для этого задания, «слишком много ограничений, которые не учитывают целевую причинность, а только наработки советской эпохи и объективные причины, хотя мы знаем, что субъективный фактор в истории никто не отменял и желание отдельного человека что-то совершить — тоже причина».

Таким образом, чрезмерная дробность и строгость критерия «роль исторической личности в событиях периода» приво-

дят к расхождениям между экспертами, а предполагаемый к измерению конструкт оценивается лишь частично.

Критерий «причинно-следственные связи» также вызывает многочисленные претензии экспертов. Основная сложность в использовании этого критерия заключается, по их мнению, в том, что для начисления баллов по нему причинно-следственные связи оцениваются отдельно от роли личности и от связи с другими событиями. Возникает затруднение, за какой критерий начислять баллы: «Чтобы охарактеризовать роль личности, нужно использовать причинно-следственные связи. Точно так же причинно-следственные связи есть и при оценке события. Эти три критерии не очень четко между собой дифференцированы». Эксперты обязаны проверять задания строго по порядку следования критериев, и, следовательно, многократно перечитывать ответ. Один из экспертов выразился так: «Все время на семинарах это приходится обсуждать. Всегда возникает вопрос, где учитывать эти причинно-следственные связи: в первом, во втором, третьем критерии?». Таким образом, одна и та же фраза ученика может являться основанием для начисления балла сразу по трем критериям. Как решать эту проблему при проверке — выбор эксперта.

Другой эксперт приводит конкретный пример: «Чаще всего ребята роль личности связывают с теми событиями, которые на что-то повлияли, а это причинно- следственные связи. «Ольга провела налоговую реформу <...> ввела уроки и погосты, чтобы упорядочить систему налогообложения». Вот куда это предложение считать? В роль личности или в причинно-следственные связи?» Таким образом, три рассмотренных критерия, с точки зрения экспертов, недостаточно четко отделены друг от друга, что создает серьезные угрозы валидности.

В оценке критериев «корректное использование исторической терминологии» и «отсутствие фактических ошибок», в отличие от перечисленных выше, у экспертов нет единодушия: одни вовсе не упоминают эти критерии или называют их простыми в оценке, а другие описывают как проблемные. Здесь сложности связаны с расхождениями в том, что понимать под историческим термином и под исторической ошибкой.

Задание № 24 большинство экспертов оценили как сложное в проверке. Из 97 упоминаний о некорректных критериях к нему относятся 47. Критерии для проверки этого задания содержательные, т. е. к ним прописаны точные формулировки. В начале работы комиссия обязательно изучает задание № 24 во всех вариантах и критерии оценки во всех вариантах, затем работает над «расширением» критериев в соответствии с указанием для экспертов «могут быть приведены другие аргументы»: эксперты предлагают разные варианты, которые ученики

могут привести как правильные. Поскольку задание № 24 в некоторых вариантах может не иметь детализированных содержательных критериев оценки, доработка этих критериев остается на усмотрение каждой региональной комиссии, что может негативно влиять на результаты оценивания: одни и те же баллы в разных регионах могут соответствовать разным уровням подготовки экзаменующихся.

Определенные затруднения у экспертов вызывает необходимость отличить приведенные учениками факты от аргументов: в соответствии с критериями оценивания приведения только фактов недостаточно для начисления баллов. Отсутствие согласованной позиции по тому, что считать фактом, а что — аргументом, приводит к расхождениям в оценках. По словам одного из экспертов, «какие-то факты считаются как „говорящие“ факты, которые не требуют подтверждения, а какие-то предлагается связывать с тем теоретическим положением, которое нужно доказать. Это тоже своего рода игра. Если мы аргумент выстраиваем, то нужны аргументы, а если нужны факты — просим факты».

Во многих интервью поднимается вопрос о несоответствии количества аргументов, приводимых в задании № 24, и количества начисляемых баллов. Максимальная оценка за задание — 4 балла, но, если испытуемый привел два аргумента только в опровержение или только в доказательство теоретического положения, ему начисляется только 1 балл. Некоторые эксперты, впрочем, согласны с таким порядком начисления баллов и утверждают, что он отвечает цели задания — проверить умение видеть обе точки зрения. Другие эксперты считают такой порядок начисления баллов нелогичным и полагают, что каждый из четырех требующихся в задании аргументов должен оцениваться в 1 балл.

3.2. Угрозы валидности, связанные с содержанием заданий

Примерно половина всех экспертов обратили внимание на то, что название задания № 25 — «историческое сочинение» — не соответствует формулировке задания. По словам одного из экспертов, «любое сочинение субъективно, только это задание — не историческое сочинение. Это описание периода, характеристика периода, все, что хотите, но не сочинение». По мнению части экспертов, в историческом сочинении должны проявляться личная позиция ученика, его мнение и собственные рассуждения, которые в текущей версии задания № 25 критериями не учитываются, поэтому жанр этого задания можно охарактеризовать, скорее, как описание по плану, пунктами которого являются позиции критериев.

Большинство экспертов так или иначе упоминали несоответствие между содержанием школьной программы и содержа-

нием экзамена. В частности, они считают, что критерии оценки задания № 25 — «роль исторической личности в событиях периода» и «причинно-следственные связи между событиями периода» — излишне детализированы. Например, в отношении критерия «роль исторической личности в событиях периода» эксперт высказался следующим образом: «Дети читают книги, читают энциклопедии, читают учебники, где роль личности характеризуется иначе. Получается двойной стандарт: в учебниках по-одному характеризуют личность, а в экзамене спрашивают совсем по-другому».

Многие отмечают, что часов, выделенных на освоение программы по истории, не хватает для овладения всем материалом: «Есть историко-культурный стандарт, там гигантское количество дидактических единиц. Честно говоря, зашкаливает. Учителя в рамках урока — а экзамен же базовый, два часа в неделю — многие дидактические единицы физически отработать не успеют».

Обновленные требования ФГОС для основного и среднего общего образования уделяют особое внимание навыкам и умениям учащихся, в том числе по истории. При этом из 36 экспертов 29 считают, что ЕГЭ по истории проверяет исключительно знания — «память», «фактаж», «фактологию». Сфокусированность теста на проверке знания исторических фактов, а не компетенций (универсальных учебных действий, навыков) учащихся, например в работе с историческими источниками, вызывает беспокойство у многих экспертов.

Излишняя детализация знаний о некоторых периодах истории, заложенная в тестах ЕГЭ, вызывает вопросы у многих экспертов: «По какому критерию осуществляется отбор? Одни события у нас показываются размашисто, крупно, знаете, крупный мазок такой. Например, Русь до XVI века. А события с 1881 по 1893 год там нужно знать чуть ли не по минутам. У меня возникают вопросы, почему и зачем».

Несмотря на критику содержания ЕГЭ, все без исключения эксперты оценили изменения спецификаций с 2008 по 2020 г. как положительные. Прежде всего это касается изменения типа заданий и используемых материалов — введения визуальных источников, уточнения заданий и критериев их оценки, отмены простых заданий с выбором ответа.

3.3. Угрозы валидности, связанные с деятельностью экспертов

Эксперты отдают себе отчет в субъективности своих суждений по поводу проверяемых работ, они отмечают нередкие расхождения во мнениях с коллегами, разные подходы к проверке. Все опрошенные эксперты отмечают, что одни и те же критерии могут интерпретироваться по-разному и вызывают расхождения в оценках. Субъективность экспертов имеет не-

сколько типичных проявлений при проверке заданий ЕГЭ: излишняя строгость, недостаточная подготовка к работе с конкретной темой, толкование ответов в пользу экзаменующихся, предвзятость оценок в стремлении избежать третьей проверки.

Некоторые эксперты слишком строго подходят к оценке заданий. Например, один эксперт рассказал о работе с критерием «отсутствие фактических ошибок» в задании № 25: «Кто-то обращает внимание на неточность, а кто-то списывает на возраст ученика. Я [говорю] прежде всего о стилистических ошибках, неточностях, которые могут искажать содержание». Те же трудности можно встретить при работе с критерием «роль исторической личности в событиях периода»: «Один эксперт считает, что все раскрыто, а кто-то считает, что двумя предложениями личность Суворова раскрыть нельзя».

Каждый эксперт является специалистом в определенной области и не всегда может полноценно работать с той или иной темой: «Там [в задании № 24] допускаются другие формулировки, значит, в силу своего кругозора и знаний уже тот или иной эксперт додумывает, можно ли это взять или нет». Зачастую эксперту довольно сложно припомнить в ситуации проверки все тонкости и детали исторического материала, и тогда оказывается, что ученик владеет информацией лучше, чем эксперт.

Большинство экспертов упоминали, что они «должны оценивать то, что он [испытуемый] написал, но оценивать это нужно объективно», не допускается «додумывание» за испытуемого. При этом некоторые открыто признаются, что сохранить такую установку при работе с экзаменационным материалом им сложно. Из 36 экспертов 30 неоднократно говорили в интервью, что тестирование должно проводиться в пользу экзаменуемых. Они открыто высказывали заботу об учениках, признавали необходимость их поддерживать, несмотря на требование объективной, непредвзятой оценки. Экспертов, участвующих в процедурах апелляции, беспокоит негласный запрет на изменение баллов в пользу ученика.

Количество третьих проверок, которые проводятся в случае существенных расхождений в оценках двух проверяющих то или иное задание, расценивается как показатель качества работы комиссии. Если таких проверок назначается много, требования к работе комиссий ужесточаются, в то время как их замечания по улучшению критериев оценки не учитываются. Две трети экспертов отмечают стремление избежать третьей проверки как препятствие на пути к объективному оцениванию: «Такое внимание к третьей проверке делает эксперта предвзятым. Стремясь к более согласованной проверке, мы можем либо занижать, либо завышать баллы».

4. Выводы и рекомендации

Тематический анализ интервью с экспертами, принимающими участие в проверке заданий ЕГЭ, позволил выявить важные угрозы валидности интерпретации тестовых баллов ЕГЭ по истории.

В первую очередь такие угрозы создают плохо сформулированные критерии для оценки заданий с открытым ответом. К примеру, критерии «роль исторической личности в событиях периода» и «причинно-следственные связи между событиями периода» к заданию № 25 получили отрицательную оценку почти всех экспертов. Эти критерии часто уточняются, чтобы минимизировать вероятность третьей проверки, но такие изменения приводят только к излишней строгости критериев или их механистичности. «Попытка увеличить согласие между экспертами с помощью более объективных критериев оценки часто приводит к сужению факторов, которые включены в оценку, и увеличивает угрозу валидности» [Crooks, Kane, Cohen, 1996]. Именно это, на наш взгляд, и происходит с критериями «роль исторической личности в событиях периода» и «причинно-следственные связи между событиями периода». Чрезмерная детализация критериев в задании № 25 и специфическая формулировка критериев в задании № 24 приводит к тому, что не все экзаменуемые получают объективную оценку.

Эксперты высказывают также существенные претензии к контентной валидности теста. Вопреки установкам ФГОС общего образования задания в ЕГЭ по истории все еще сильно ориентированы на проверку знания исторических фактов, а не компетенций испытуемых. В содержании и планируемых результатах школьной программы за последние пять лет произошли значительные изменения: обновлен ФГОС, принят историко-культурный стандарт. ЕГЭ по истории в спецификации 2020 г. пока не адаптирован таким образом, чтобы полностью удовлетворять новым требованиям. Новая спецификация, как заявляют разработчики, позволит решить этот вопрос, в ней больше внимания будет уделено оценке компетенций школьников. Однако даже краткий анализ документов 2021 г. свидетельствует о том, что угрозы валидности сохраняются.

Интервью позволили также выделить несколько основных угроз, связанных с субъективностью экспертов: чрезмерная или недостаточная строгость при оценивании, недостаток предметных знаний у экспертов, толкование ответов в пользу экзаменуемых и предвзятость оценок в стремлении избежать третьей проверки. Такого рода искажения в оценивании, проводимом с помощью экспертов, хорошо известны и получили название «эффекты оценщиков» (подробный обзор посвященных им исследований см., например, в [Myford, Wolfe, 2003]).

Таким образом, проведенный анализ выявил несколько существенных угроз валидности интерпретации тестовых баллов ЕГЭ по истории. Для преодоления этих угроз нужно скорректировать критерии оценки некоторых заданий и обратить внимание на соответствие содержания теста содержанию школьной программы, овладение которой он должен проверять. В какой степени эти угрозы удастся преодолеть разработкой новой спецификации ЕГЭ, остается вопросом будущего.

5. Ограничения исследования

В настоящем исследовании предпринята попытка проанализировать угрозы валидности интерпретации тестовых результатов через тематический анализ интервью с экспертами. Безусловно, этот метод имеет ряд ограничений.

Во-первых, все выводы основаны только на высказанных экспертами мнениях, а разработка критериев для тематического анализа этих интервью, хотя и основывалась на аналогичных исследованиях и научной литературе, тоже является субъективным действием исследователя. Ограниченный масштаб выборки не позволяет делать масштабные выводы. Вместе с тем квалификация и опыт респондентов свидетельствуют в пользу справедливости и надежности выводов.

Во-вторых, коротких интервью с небольшим числом экспертов недостаточно, чтобы делать серьезные выводы о валидности всего теста. Полное валидизационное исследование ЕГЭ потребовало бы привлечения гораздо больших ресурсов и использования более широкого спектра методов, включая психометрический анализ заданий, факторный анализ результатов ЕГЭ по истории, возможно, исследование объективности экспертов, принимающих участие в оценке. В будущем разработчикам ЕГЭ целесообразно было бы взять на себя эту непростую обязанность.

Литература

1. Артасов И. А., Мельникова О. Н. (2018) Эволюция экзаменационных моделей КИМ ЕГЭ по истории // Педагогические измерения. № 2. С. 48–56.
2. 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. P. 609–625. doi:10.1080/00313831.2015.1066429.
3. Crooks T. J., Kane M. T., Cohen A. S. (1996) Threats to Valid Use of Assessment // Assessment in Education. Vol. 9. No 3. P. 265–285. doi:10.1080/0969594960030302.
4. Dennis I., Newstead S. E. (1994) The Strange Case of the Disappearing Sex Bias // Assessment & Evaluation in Higher Education. Vol. 19. No 1. P. 49–56. doi:10.1080/0260293940190105.
5. Haladyna T. M. (2006) Perils of Standardized Achievement Testing // Educational Horizons. https://files.eric.ed.gov/fulltext/EJ750641.pdf

6. Haladyna T. M., Downing S. M. (2005) Construct-Irrelevant Variance in High-Stakes Testing // Educational Measurement: Issues and Practice. Vol. 23. No 1. P. 17–27. doi:10.1111/j.1745-3992.2004.tb00149.x.
7. Haladyna T. M., Rodriguez M. (2013) Developing and Validating Test Items. London: Routledge. doi:10.4324/9780203850381.
8. Kane M., Crooks T., Cohen A. (2005) Validating Measures of Performance // Educational Measurement: Issues and Practice. Vol. 18. No 2. P. 5–17. doi:10.1111/j.1745-3992.1999.tb00010.x.
9. 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. P. 1–52.
10. Maris G., Bechger T. (2006) 20 Scoring Open Ended Questions // Handbook of Statistics. Vol. 26. P. 663–681. doi:10.1016/S0169-7161(06)26020-6.
11. 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. P. 741–749.
12. 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.
13. Miller M. D., Linn R. L., Gronlund N. E. (2009) Measurement and Assessment in Teaching. Upper Saddle River, NJ: Merrill/Pearson.
14. 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. P. 386–422.
15. Popham W. J. (1990) Modern Educational Measurement: A Practitioner's Perspective. Englewood Cliffs, NJ: Prentice Hall.
16. 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.
17. Traub R. E., Fisher C. W. (1977) On the Equivalence of Constructed-Response and Multiple-Choice Tests // Applied Psychological Measurement. Vol. 1. No 3. P. 355–369.
18. Traub S. L. (1993) Evaluating Medical Tests: Objective and Quantitative Guidelines // American Journal of Health-System Pharmacy. Vol. 50. No 11. P. 2440–2443.
19. 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. P. 11–29.

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. doi:10.1080/0260293940190105.

Haladyna T. M. (2006) Perils of Standardized Achievement Testing // Educational Horizons. Available at: https://files.eric.ed.gov/fulltext/EJ750641.pdf (accessed 20 July 2021).

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.

Kane M., Crooks T., Cohen A. (2005) Validating Measures of Performance. *Educational Measurement: Issues and Practice*, vol. 18, no 2, pp. 5–17. doi:10.1111/j.1745-3992.1999.tb00010.x.

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.

Maris G., Bechger T. (2006) 20 Scoring Open Ended Questions. *Handbook of Statistics*, vol. 26, pp. 663–681. doi:10.1016/S0169-7161(06)26020-6.

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.

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.

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, pp. 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.