Классический и современный подходы к измерению валидности заданий на взаимное оценивание в МООК

Д. А. Кравченко

Кравченко Дарья Андреевна
аналитик Центра психометрических исследований в онлайн-образовании Национального исследовательского университета «Высшая школа экономики». Адрес: 119607, Москва, ул. Старая Басманная, 21/4, стр. 1. E-mail: dakravchenko@hse.ru

Аннотация. Представлены результаты исследования валидности заданий на взаимное оценивание, применяемых в массовых открытых онлайн-курсах, в рамках двух психометрических теорий: классической и современной. С применением классической теории были получены данные о конвергентной валидности задания на взаимное оценивание, о низком уровне его критериальной валидности и рассогласованности в оценках экспертов. С помощью современной теории тестирования удалось выявить эффекты строгости и снисходительности экспертов, установлено, что эксперты в целом являются снисходительными к студентам и склонны завышать баллы. На основе полученных данных обсуждаются достоинства и недостатки использованных психометрических теорий, а также возможности их комбинирования.

Ключевые слова: массовые открытые онлайн-курсы, взаимное оценивание, классическая теория тестирования, Item Response Theory, пиринговые задания.

DOI: 10.17323/1814-9545-2018-4-99-115

Статья поступила в редакцию в июле 2018 г.

Массовые открытые онлайн-курсы (МООК) как одна из форм дистанционного обучения набирают популярность как среди студентов, так и среди университетов. В 2016 г. более чем в 700 университетах по всему миру действовало 6850 курсов. Крупнейшей платформой МООК с более чем 23 млн учащихся является *Coursera* [Shah, 2016]. В 2017 г. уже более 800 университетов предложили учащимся более 9400 МООК, а *Coursera* достигла отметки 30 млн учащихся и 2700 курсов [Shah, 2017].

МООК предполагают открытый доступ к учебным материалам через интернет, так что число студентов, которых они могут привлечь, не ограничено. Онлайн-курс состоит из видеолекций, материалов для чтения, практических упражнений, экзамена-

Автор признателен своему научному руководителю Д. Ф. Аббакумову за помощь в написании этой статьи.

ционных заданий и форума для общения преподавателя со студентами и студентов между собой. Обычно MOOK разрабатывают университеты и размещают на платформе-провайдере, например *Coursera*, *EdX*, *XuetangX*, *FutureLearn*, *Udacity*, Национальная платформа открытого образования, *Stepik*, Универсариум. Крупнейшими платформами являются *Coursera* и *EdX*, чьи аудитории достигли 30 млн и 14 млн студентов соответственно [Shah, 2017].

Когда результаты онлайн-курсов стали засчитываться студентам наравне с традиционными университетскими курсами, существенно выросли требования к качеству инструментов оценивания — к их валидности и надежности. Проверка знаний и навыков в MOOK чаще всего проводится с помощью заданий с автоматизированной проверкой (тестов) и заданий на взаимное оценивание. Взаимное оценивание предполагает самостоятельное конструирование ответа студентом и последующую его проверку другими студентами. Обычно в оценивании участвуют не менее трех студентов. Рассылка работ на проверку производится платформой автоматически.

Взаимное оценивание дает возможность использовать в проверке открытые задания (например, эссе и дизайн-проекты) и имеет высокий обучающий потенциал, так как студенты получают аналитический опыт, проверяя и комментируя работы друг друга. Однако такое оценивание характеризуется высоким уровнем субъективности, следствием чего является отсутствие уверенности в его валидности и надежности.

Результаты исследования валидности взаимных оценок неоднозначны. В ряде работ выявлена высокая положительная корреляция между результатами взаимного оценивания, оценкой преподавателя и тестами [Kaplan, Bornet, 2014; Dancey, Reidy, 2017]. В других исследованиях обнаружена низкая валидность взаимных оценок, обусловленная тем, что лица, которые оценивают работы, не обучены принципам объективного оценивания [Admiraal, Huisman, van de Ven, 2014], не ориентируются в предмете на высоком уровне [Falchikov, Goldfinch, 2000], а также тем, что не все авторы курсов разрабатывают объективные критерии для оценивания [Falchikov, Goldfinch, 2000].

В данной статье мы рассматриваем классический и современный подходы к исследованию валидности взаимного оценивания в MOOK, иллюстрируем их применение на материале двух онлайн-курсов, обсуждаем их достоинства и недостатки, а также возможности комбинирования.

1. Исследование валидности заданий на взаимное оценивание

В психометрике существуют два подхода к исследованию валидности измерительных инструментов: классический и совре-

менный[1]. Эти подходы не являются взаимоисключающими. Поэтому мы предлагаем их комбинировать.

1.1. Классический подход

Валидность теста, согласно А. Анастази, означает, что тест достоверно измеряет именно то качество, для измерения которого он создан. В данной работе валидность экспертного оценивания мы рассматриваем как точность оценок, которые студенты ставят друг другу. В рамках классического подхода обычно измеряют конструктную и критериальную валидность, а также классическую надежность [Анастази, Урбина, 2007].

Конструктная валидность — это один из основных теоретических типов валидности, отражающий степень репрезентации заявленного свойства в результатах теста [Шмелев, 2013]. Мы измеряли конвергентную валидность. Под конвергентной валидностью понимается положительная корреляция между оценками, полученными с помощью различных инструментов, измеряющих один и тот же конструкт. Например, для измерения внутренней мотивации человека существует несколько тестов. Для подтверждения конвергентной валидности целесообразно собрать данные по всем тестам и сравнить результаты. Если данные, полученные в разных тестах, будут иметь высокую корреляцию, можно говорить об их конвергентной валидности.

В данном исследовании для оценки конвергентной валидности мы вычисляли коэффициенты корреляции Пирсона между средней оценкой по экспертам (студентам) во взаимном оценивании и оценками по тестам, а также между оценками каждого эксперта во взаимном оценивании и оценками по тестам (так как в курсе были и взаимные оценки, и тесты).

$$r_{xy} = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2(Y-\overline{Y})^2}}, \quad (1)$$

где X, Y — наблюдения, элементы выборки; $\overline{X}$, $\overline{Y}$ — выборочные средние.

Под критериальной валидностью понимается положительная корреляция между оценками и эмпирическим критерием. В качестве критерия могут выступать, например, итоговые баллы студентов по предмету, с которым связаны тесты, измеряющие их знания и способности. В нашем исследовании таким критерием является итоговый балл по курсу. Для измерения критериальной валидности мы рассчитывали коэффициенты корреляции Пирсона (см. уравнение 1) между оценками каждого

[1] The National Council on Measurement in Education: A Professional Organization for Individuals Involved in Assessment, Evaluation, Testing. Philadelphia, PA. http://www.ncme.org/home

эксперта во взаимном оценивании и итоговой оценкой за курс, а также между средней оценкой по экспертам во взаимном оценивании и итоговой оценкой за курс.

Надежность, как правило, рассчитывается как коэффициент корреляции между оценками, полученными от студентов, и оценками, которые поставил профессор. При этом предполагается, что профессор может обеспечить точную и объективную оценку работ студентов. В нашем исследовании классическая надежность понимается как согласованность оценок экспертов, которая оценивается на основании сравнения баллов, выставляемых экспертами. Если все три эксперта поставили максимальный балл студенту, можно говорить о согласованности оценок. Если же оценки трех экспертов противоречат друг другу, имеет место рассогласование.

Мы оценивали согласованность с помощью расчета коэффициента конкордации:

$$W = \frac{12S}{n^2(m^3 - m)}, \quad (2)$$

где S — сумма квадратов отклонений всех оценок рангов каждого объекта экспертизы от среднего значения; n — число экспертов; m — число объектов экспертизы.

1.2. Современный подход

Обоснованность экспертных оценок и сама процедура выставления оценки не раз ставились под сомнение [Charney, 1984; Gere, 1980; Huot, 1990]. Даже если эксперт специализируется в оцениваемой области и способен ставить равноценные объективные оценки, остаются вопросы к интерпретации шкалы оценивания: такая шкала не может быть линейной, и балл 2 в одной задаче не может быть равноценным баллу 2 в другой задаче. Эта и другие особенности шкалы измерения в классической теории тестирования (КТТ) обусловливают проблемы в обеспечении валидности и надежности заданий на взаимное оценивание. В современной теории тестирования (*Item Response Theory, IRT*) шкала является метрической, у нее нет фиксированного начала, а сумма трудностей всех заданий равна нулю. Такой подход позволяет более точно измерить валидность оценок и выявить искажения экспертного оценивания.

Исследования экспертной оценки в основном сфокусированы на анализе ее надежности. Д. Линакр [Linacre, 1989] отмечает, что стремление получить истинный балл в результате оценивания экспертами является предпосылкой превращения вариаций оценок экспертов и вариации нежелательной дисперсии ошибок в проблему измерения, так что эти вариации должны быть уменьшены, насколько это возможно. Другой подход к экспертному оцениванию реализован в многофасетной модели, разра-

ботанной Д. Линакром, взявшим за основу модель Раша. В ней вариации экспертных оценок рассматриваются как неизбежная часть процесса оценивания, более того, они считаются не препятствием для измерения, а преимуществом, поскольку обеспечивают изменчивость, достаточную для вероятностной оценки строгости экспертов, трудности заданий и уровня способностей студентов на линейной шкале.

Сторонники применения модели Раша утверждают, что важно дать экспертам понимание рейтинговой шкалы, с помощью которой они будут оценивать студентов [Lunz, Wright, Linacre, 1990]. На самом деле использование модели Раша устраняет необходимость приведения оценок экспертов к согласованности, поскольку оценки способностей испытуемых не зависят от строгости конкретного эксперта.

В рамках *IRT* оценка, полученная студентом при взаимном оценивании, рассматривается как функция трех переменных: параметра студента (уровня его знаний), параметра задания (его трудности) и параметра эксперта (его строгости или снисходительности) [Lunz, Wright, Linacre, 1990], а оценка, полученная студентом в рамках тестирования, рассматривается как функция двух переменных: параметра студента (уровня его знаний) и параметра задания (его трудности).

Мы использовали многофасетную модель Раша [Lunz, Wright, Linacre, 1990]:

$$\log\left(\frac{P_{nijk}}{P_{nij}(k-1)}\right) = B_n - D_i - C_j - F_{jk}, \quad (3)$$

где P_{ni} — вероятность того, что испытуемый n успешно выполнит задание i, испытуемый n имеет показатель способности B_n и показатель трудности задания D_i. C_j является показателем строгости или снисходительности в оценивании экспертами j, которые присуждают рейтинги k испытуемому n по заданию i.

В данной модели о низкой валидности говорит высокий уровень неожиданных оценок и отличных от критических статистик показателей. Неожиданная оценка — это рассогласование между баллами, которые были выставлены экспертами, и ожидаемыми баллами, т. е. теми, которые предсказаны моделью.

2. Применение классического и современного подходов к исследованию валидности на материале онлайн-курсов

2.1. Классический подход

Мы проанализировали данные 1308 слушателей курса «Философия культуры»[2] — это все записавшиеся на курс. Женщины составили 66% слушателей, мужчины — 34%. Возраст испытуемых — от 15 до 50 лет (M = 30 лет). 46% из них имеют высшее

[2] National Research University Higher School of Economics. Философия культуры (Philosophy of Culture). https://www.coursera.org/learn/filosofiya-kultury

профессиональное образование (бакалавр/специалист). Большинство (67%) родились и проживают в России.

Нас интересовали студенты, завершившие курс полностью и при этом участвовавшие во взаимном оценивании и получившие баллы минимум от трех экспертов. Таким образом, итоговая выборка составила 188 человек.

Данные о взаимных оценках, оценках за тесты и итоговых баллах по курсу «Философия культуры» на платформе *Coursera* были получены из итогового отчета по опросу студентов, который проводит Центр внутреннего мониторинга.

Курс «Философия культуры» включает пять тестов с множественным выбором и два задания на взаимную оценку. С помощью КТТ мы проанализировали одно задание на взаимное оценивание с критериями для оценки. Студенты писали краткий текст на заданную тему (эссе). Для проведения анализа мы использовали данные только тех студентов, чьи тексты были оценены как минимум тремя экспертами. Задание оценивалось по четырем критериям, по каждому из них могло быть начислено от 0 до 3 баллов. Максимальный балл, который мог выставить каждый эксперт, — 12.

Задание на взаимное оценивание звучало следующим образом: «Выберите конкретный эпизод из истории (можно тот, который разбирал лектор) и сформулируйте характерные примеры конфликтов „природа против культуры“, „природа против духа“, „культура против духа“. По желанию можно изобразить их на общей схеме (круги Эйлера)». Студентам были предоставлены примеры схем для выполнения задания. Ниже представлен пример критерия.

Критерий № 1. Какие элементы присутствуют на схеме? Список элементов, наличие которых оценивается: заголовок схемы, два примера категорий, их конфликт.

3 балла — есть заголовок, два примера из разных категорий, их конфликт;
2 балла — есть три элемента из четырех;
1 балл — есть два элемента из четырех;
0 баллов — есть только один элемент.

В задании были даны примеры для облегчения оценивания, на которые можно было ссылаться также и при выполнении задания.

Расчет итогового балла производился по формуле:

Итоговая оценка = Средний балл за тесты и взаимно оцениваемые задания (за 7 недель) × 0,5 + Балл за итоговый тест × 0,4 + + Активное участие на форуме × 0,1.

Коэффициент вклада каждого вида активности устанавливал автор — преподаватель онлайн-курса. В данном курсе задания на взаимное оценивание вносят большой вклад в итоговую оценку: они составляют половину итогового балла, и поэтому объективность данного оценивания очень важна.

Из полученных оценок по четырем критериям рассчитывался общий балл по заданию от каждого эксперта (медиана). Далее каждый студент получал по три оценки от экспертов. Эти баллы мы использовали при расчете коэффициента конкордации. Общий балл за задание на взаимное оценивание, который использовался при определении итогового балла, рассчитывался как среднее арифметическое по трем оценкам экспертов. Данные баллы мы использовали при оценке корреляций.

2.2. Современный подход

Выборка исследования составила 1483 работы студентов (868 работ по курсу «Философия культуры» и 615 работ по курсу на английском языке *Understanding Russians: Contexts of Intercultural Communication*[3]). Всего было получено 4449 взаимных оценок, так как каждая работа была оценена тремя экспертами.

Задание на взаимное оценивание в курсе *Understanding Russians: Contexts of Intercultural Communication* также состояло в написании эссе. Студентам предлагались две темы на выбор. Инструкция по написанию эссе включала описание структуры, ключевые слова, которые должны быть использованы при написании, и объем текста в словах.

Эксперты получали инструкции для проверки. Эссе необходимо было оценить на основании шести критериев. Максимальный балл по одному из критериев текст получал в том случае, если эссе отвечает на вопрос, как преодолеть культурные пробелы в межкультурной коммуникации, если в тексте есть детализация культурных барьеров и обсуждение их с точки зрения культурных измерений. Есть и другие требования к эссе: оно должно состоять из 500–1000 слов, отличаться новизной и содержать ссылки на внешние материалы или материалы курса. В зависимости от наличия необходимых содержательных элементов в тексте эссе получает то или иное количество баллов.

У каждого студента есть свой *id*-номер, для которого прописываются все действия на онлайн-платформе. Для анализа были использованы *id* студентов и *id* экспертов, которые выставляли баллы. Данные были помещены в контрольный файл для анализа в программе *FACETS*. Файл содержал *id* студента и соответствующие три *id* экспертов с баллами по шести критери-

[3] National Research University Higher School of Economics. Understanding Russians: Contexts of Intercultural Communications. https://www.coursera.org/learn/intercultural-communication-russians

Таблица 1. **Корреляции между заданиями на взаимное оценивание и тестами автоматизированной системы оценивания в онлайн-курсах**

Тест	Задание на взаимное оценивание
1	0,57**
2	0,04
3	0,26
4	0,18
5	0,02
6	0,01

** $p \leqslant 0{,}01$.

ям. В файл были помещены все данные о студентах и об оценках, которые они получили от экспертов.

С помощью этого анализа мы получили информацию об эффекте экспертов, о завышении или занижении баллов при оценивании работ студентов.

3. Результаты определения валидности взаимного оценивания

3.1. Классический подход

Результаты исследования конвергентной валидности приведены в табл. 1.

Корреляции между тестами 2, 3, 4, 5, 6 и заданием на взаимное оценивание являются незначимыми и низкими. Тесты и задания на взаимное оценивание содержательно различны. Коэффициенты не должны быть значимыми, потому что задания направлены на измерение знаний по разным темам философии культуры. При этом, однако, получен значимый коэффициент корреляции между первым тестом и заданием на взаимное оценивание — 0,57. На этом основании мы можем заключить, что взаимное оценивание характеризуется конвергентной валидностью, поскольку содержательно первое задание на взаимное оценивание направлено на измерение знаний о тех же конструктах, что и первый тест.

Коэффициент корреляции итогового балла с заданием на взаимное оценивание составляет 0,73 ($p \leqslant 0{,}01$) и является значимым и высоким. Он свидетельствует о том, что взаимное оценивание вносит большой вклад в итоговый балл и имеет большую прогностическую силу. Также в данном случае можно говорить о критериальной валидности взаимного оценивания

в курсе «Философия культуры». В качестве критерия выступает итоговый балл по курсу.

О надежности заданий на взаимное оценивание позволяет судить коэффициент конкордации. Он составляет 0,53 (p = 0,000). Такой уровень согласованности является средним. Это означает, что эксперты могут расходиться во мнениях относительно оценок по критериям. Несогласованность их оценок может быть связана с тем, что эксперты неодинаково понимают критерии оценки или рубрики составлены некорректно. Коэффициент конкордации является простым и понятным параметром для измерения согласованности оценок экспертов, поэтому мы ограничились рассмотрением одного примера задания на взаимное оценивание.

Анализируя оценки, которые эксперты ставили по каждому критерию, мы установили, что они склонны выставлять исключительно высокие или низкие баллы, минуя средние категории. В научной литературе также есть описания эффектов строгости или снисходительности экспертов. Эти данные получены в исследованиях, осуществленных в рамках современной теории тестирования [Falchikov, 1986; Orpen, 1982; Ueno, Okamoto, 2016; Lunz, Wright, Linacre, 1990].

Важнейшие результаты исследования валидности задания взаимного оценивания, применяемого в рассмотренном курсе, в рамках классической теории тестирования состоят в следующем.

1. Уровень конвергентной валидности задания — средний.
2. Вклад балла за задание на взаимное оценивание в итоговый балл следует считать значимым. Показатель критериальной валидности — средний с тенденцией к низкому.
3. Уровень надежности критериев — средний, т. е. эксперты могут расходиться во мнениях при вынесении оценок по критериям. Недостижение высокого уровня надежности, на наш взгляд, связано с неточной формулировкой критериев. Четыре критерия, предложенные для оценивания задания, допускали субъективную трактовку их смысла, отсюда и большие расхождения в оценках, полученных от трех экспертов. Критерии необходимо формулировать более точно и просто. Правила начисления того или иного количества баллов также нуждаются в более детальном описании, которое будет способствовать более точному оцениванию студентами работ своих сокурсников.

Используя выводы, полученные по результатам проведенного анализа, следует иметь в виду ряд серьезных ограничений данного исследования. Оно основано на материалах взаим-

ного оценивания одного задания в рамках гуманитарного онлайн-курса. У нас не было возможности сравнить взаимное оценивание выполнения этого задания с взаимным оцениванием в других онлайн-курсах (гуманитарных или технических). Еще одним значимым ограничением являлась выборка испытуемых, которая составила менее тысячи человек. Такие ограничения устранимы: для этого необходимо провести исследования на разных онлайн-курсах (гуманитарных и технических), в которых используются задания на взаимное оценивание.

Проведение анализа данных в рамках классической теории тестирования также накладывает определенные ограничения, а именно: отсутствует возможность оценить ошибку измерения и получить показатели строгости или снисходительности экспертов. Данные ограничения были устранены с помощью проведения анализа в рамках современной теории тестирования.

3.2. Современный подход

Результаты исследования валидности заданий на взаимное оценивание в курсе «Философия культуры» в рамках *IRT* приведены на рис. 1 в виде графических мер студентов, экспертов и задания (с критериями). В левой части фигуры — шкала логитов (логарифмический шанс), которая одинакова для всех трех граней (студенты, эксперты, критерии). Масштаб карты — каждые четыре студента и эксперта обозначены звездочкой.

Испытуемые упорядочиваются от наиболее способных в верхней части до наименее способных в нижней части карты данных. Критерии упорядочены от наиболее сложных элементов вверху карты до наименее сложных в нижней части карты данных. Эксперты расположены от наименее строгих в верхней части карты данных до наиболее строгих — в нижней части.

В крайнем правом столбце приведены наиболее вероятные показатели для каждого уровня способностей. Различия на рисунке представлены как разница между элементами фасетов.

В нашем случае данные распределены от –8 до +10 логитов. Судя по колонке экспертов, 28 из них являются наименее строгими, т. е. их оценки по всем критериям выше, чем у других экспертов. По взаимному расположению экспертов относительно студентов на карте можно заключить, что эксперты склонны завышать баллы: основная масса экспертов расположена в промежутке от 0 до +4 логитов, а студентов — от –2 до +2 логитов, следовательно, эксперты были не строги при оценивании способностей студентов. Распределение способностей студентов смещено вниз, т. е. большинство из них имеют средний уровень способностей, и он ниже, чем те оценки, которые выставляют эксперты. Распределение строгости экспертов смещено вверх, т. е. они не склонны быть строгими. Такое рассогласование между оценками, которые выставляют эксперты, и уровнем под-

Рис. 1. **Карта данных для исследования валидности заданий на взаимное оценивание в курсе «Философия культуры»**

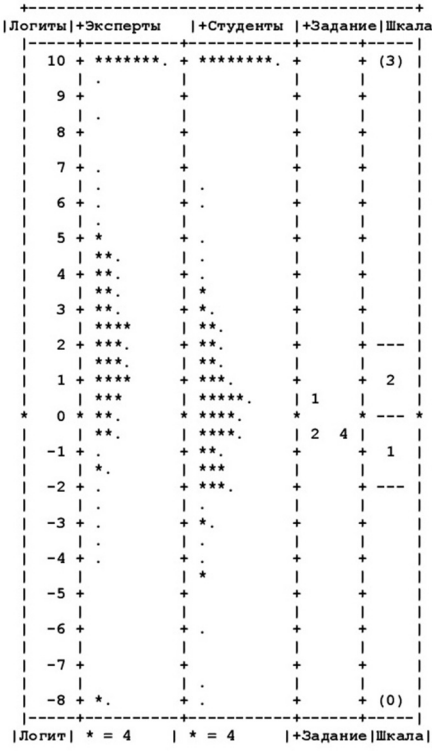
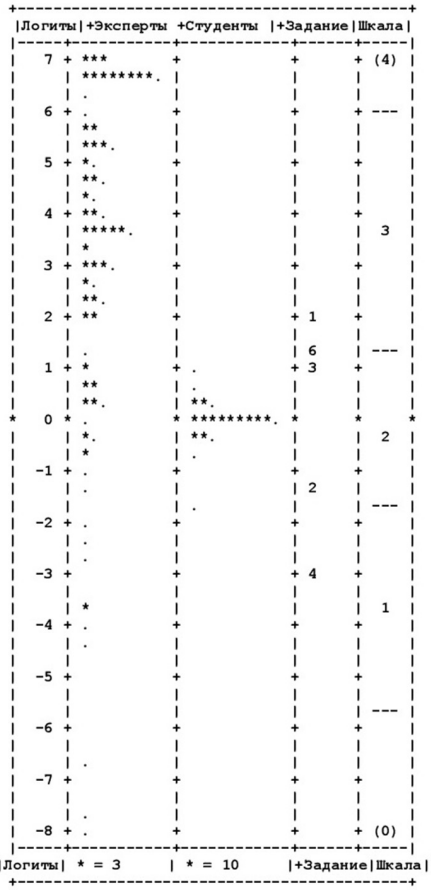


Рис. 2. **Карта данных для исследования валидности задания на взаимное оценивание в курсе *Understanding Russians: Contexts of Intercultural Communication***



готовленности студентов свидетельствует о низкой валидности задания.

Таким образом, выявлено, что эксперты склонны завышать баллы студентам, а уровень подготовленности студентов ниже, чем полученные ими оценки.

Для того чтобы показать возможности многофасетной модели Раша в выявлении искажений в экспертном оценивании, мы проанализируем еще один курс.

На рис. 2 представлены результаты исследования валидности задания на взаимное оценивание в курсе *Understanding Russians: Contexts of Intercultural Communication*. Масштаб кар-

ты — каждые три студента и десять экспертов обозначены звездочкой.

Здесь данные распределены от –8 до +7 логитов. Судя по колонке экспертов, 9 из них являются наименее строгими.

Основная масса экспертов расположена в промежутке от 0 до +6 логитов, а студентов — от –1 до +1 логита. Очевидно, и в этом задании эксперты были не строги в оценках. По взаимному расположению экспертов относительно студентов на карте можно заключить, что эксперты склонны завышать баллы. Такое рассогласование между оценками экспертов и уровнем подготовленности студентов свидетельствует о низкой валидности задания и подтверждает данные, полученные при анализе задания в другом онлайн-курсе.

Таким образом, анализ данных по второму заданию также показывает, что эксперты склонны завышать баллы студентам. Их оценки не соответствуют уровню подготовленности студентов.

Важнейшими результатами исследования валидности взаимного оценивания в рамках современной теории тестирования (*IRT*) являются следующие.

1. В обоих курсах оценки экспертов не соответствуют способностям студентов — эксперты в целом снисходительны и склонны завышать баллы.
2. В обоих курсах у экспертов имеют место неожиданные оценки. Неожиданной считается оценка, существенно отклоняющаяся от прогнозируемой на основании модели. То есть при общей тенденции к снисходительности в оценках есть эксперты, которые занижают баллы. Занижение баллов студентам с высоким уровнем подготовленности создает неравные условия для выполнения задания и прохождения курса в целом. Мы полагаем, что такие оценки необходимо выбраковывать и не брать в расчет при выставлении общего балла за задание на взаимное оценивание и итогового балла. Так можно сделать измерения максимально объективными.

Исследование в рамках современной теории тестирования также имеет ряд ограничений:

- невозможно проследить, намеренно ли эксперты выставляют завышенные, заниженные или случайные оценки;
- в рамках применяемой модели не проводится анализ данных с учетом пола и возраста студентов, уровня их мотивации и времени, затраченного на выполнение задания;
- исследование было проведено на материале только двух заданий на взаимное оценивание в рамках гуманитарных онлайн-курсов.

Таблица 4. **Оценка валидности и надежности взаимного оценивания в КТТ и *IRT***

	КТТ (классическая теория тестирования)	*IRT* (современная теория тестирования)
1	Уровень надежности задания оценен как средний. Причиной являются ограничения исследования. Показатель надежности ниже критического порога	Надежность задания была оценена отдельно от надежности студентов и экспертов. Показатель надежности высокий
2	Уровень надежности критериев средний. Наименьшие показатели получены по первому и третьему критерию. При их исключении из рассмотрения значительного роста надежности не произошло	Анализ критериев показал, что реже всего выставляются баллы 1 и 0. Существует необходимость в более строгой оценке выполнения задания с помощью существующих критериев. Возможно, их содержание нуждается в доработке
3	Уровень конвергентной валидности средний. Балл за взаимное оценивание вносит значимый вклад в итоговый балл. Показатель критериальной валидности чуть ниже среднего	Данные хорошо согласуются с моделью. Однако нет оснований для вывода о высоком уровне валидности, так как выявлено большое количество неожиданных оценок и отличных от критических статистик показателей
4	Удалось оценить согласованность оценок и точность суждений экспертов	Удалось отдельно оценить трудность задания, уровень способностей студентов и строгость экспертов
5	Выявлена необходимость в доработке критериев оценивания	Выявлено наличие искажений в оценивании — завышение балла по выборке в целом
6	Анализ достаточно прост в осуществлении	Анализ сложнее, чем в КТТ, в реализации и в интерпретации
7	Ошибка измерений не была оценена	Оценена ошибка измерения для уровня способностей студентов и для показателя строгости экспертов

Для устранения этих ограничений требуются дополнительные исследования с анкетированием экспертов и применением других моделей, которые будут учитывать большее количество параметров (пол, возраст, страна и др.).

Обсуждение и выводы

Таким образом, мы измерили валидность и надежность взаимного оценивания в рамках двух заданий из гуманитарных онлайн-курсов двумя способами: с помощью классической теории тестирования и современной теории тестирования. В табл. 4 представлены их преимущества и недостатки.

Результаты анализа, полученные с помощью классической теории тестирования и *IRT*, схожи. Тем не менее каждая из теорий имеет свои достоинства и недостатки.

Достоинством классической теории тестирования в сравнении с современной теорией является простота анализа данных и интерпретации результатов. Этот метод удобно применять в качестве экспресс-диагностики заданий на взаимную оценку. При этом важно учитывать, что, во-первых, согласованность экспертных баллов зависит от подготовленности конкретной вы-

борки экспертов, а во-вторых, анализ концентрируется только на измерении согласованности оценок экспертов, отсутствует возможность оценить ошибку измерения и объективно судить о степени строгости или снисходительности экспертов. Данные ограничения устраняются с помощью применения современной теории тестирования. Этот метод более сложный, но предоставляет возможность получить данные о наличии искажений — завышения или занижения экспертных оценок.

Экспресс-диагностика с помощью классической теории тестирования является неотъемлемой частью анализа данных. Она позволяет выявить основные проблемные места и наметить траекторию более углубленного исследования с помощью современной теории тестирования, именно поэтому мы считаем, что комбинированный подход является оптимальным для корректировки и доработки заданий на взаимную оценку.

Литература

1. Анастази А., Урбина С. (2007) Психологическое тестирование. СПб.: Питер.
2. Шмелев А. Г. (2013) Практическая тестология. Тестирование в образовании, прикладной психологии и управлении персоналом. М.: Маска.
3. Admiraal W., Huisman B., van de Ven M. (2014) Self- and Peer Assessment in Massive Open Online Courses // International Journal of Higher Education. Vol. 3. No 3. P. 110–128. doi: 10.5430/ijhe.v3n3p119.
4. Charney D. (1984) The Validity of Using Holistic Scoring to Evaluate Writing: A Critical Overview // Research in the Teaching of English. Vol. 18. No 1. P. 65–81.
5. Dancey C. P., Reidy J. (2017) Statistics without Maths for Psychology. Upper Saddle River: Pearson.
6. Falchikov N. (1986) Product Comparisons and Process Benefits of Peer Group and Self-Assessments // Assessment and Evaluation in Higher Education. Vol. 11. No 2. P. 146–166. doi: 10.1080/0260293860110206.
7. Falchikov N., Goldfinch J. (2000) Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks // Review of Educational Research. Vol. 70. No 3. P. 287–322.
8. Gere A. R. (1980) Written Composition: Toward a Theory of Evaluation // College English. Vol. 42. No 1. P. 44–48, 53–58.
9. Huot B. (1990) The Literature of Direct Writing Assessment: Major Concerns and Prevailing Trends // Review of Educational Research. Vol. 60. No 2. P. 237–263.
10. Kaplan F., Bornet C. (2014) A Preparatory Analysis of Peer-Grading for a Digital Humanities MOOC // Digital Humanities: Book of Abstracts. Lausanne: University of Lausanne. P. 227–229.
11. Linacre J. M. (1989) Many-Faceted Rasch Measurement. Chicago, IL: MESA.
12. Lunz M. E., Wright B. D., Linacre J. M. (1990) Measuring the Impact of Judge Severity on Examination Scores // Applied Measurement in Education. Vol. 3. No 4. P. 331–345.
13. Orpen C. (1982) Student versus Lecturer Assessment of Learning // Higher Education. Vol. 11. No 5. P. 567–572.
14. Shah D. (2016) Monetization over Massiveness: Breaking down MOOCs by the Numbers in 2016. https://www.edsurge.com/news/2016–12–29-

monetization-over-massiveness-breaking-down-moocs-by-the-numbers-in-2016
15. Shah D. (2017) Coursera's 2017: Year in Review. https://www.class-central.com/report/coursera-2017-year-review/
16. Shah D. (2018) A Product at Every Price: A Review of MOOC Stats and Trends in 2017. https://www.class-central.com/report/moocs-stats-and-trends-2017/
17. Ueno M., Okamoto T. (2016) Item Response Theory for Peer Assessment // IEEE Transactions on Learning Technologies. Vol. 9. No 2. P. 157–170.
18. Wright B. D., Masters G. N. (1982) Rating Scale Analysis: Rasch Measurement. Chicago: MESA.

Classical Test Theory and Item Response Theory in Measuring Validity of Peer-Grading in Massive Open Online Courses

Author **Daria Kravchenko**
Analyst, Centre for Psychometrics in eLearning, National Research University Higher School of Economics. Address: Bld. 1, 21/4 Staraya Basmannaya Str., 119607 Moscow, Russian Federation. E-mail: dakravchenko@hse.ru

Abstract The article presents the results of research on validity of peer-review assignments in massive open online courses within the framework of classical test theory (CTT) and item response theory (IRT). CTT-based analysis yielded data on convergent validity of the peer-review assignment, the low level of its criterion validity, and rater disagreement. IRT-based analysis revealed rater bias and established that experts largely tend to be lenient and overrate their peers. The findings are used to discuss the advantages and disadvantages of the psychometric theories in question and the opportunities for combining the two.

Keywords massive open online courses, peer grading, classical test theory, item response theory, peer-review assignments.

References

Admiraal W., Huisman B., van de Ven M. (2014) Self- and Peer Assessment in Massive Open Online Courses. *International Journal of Higher Education*, vol. 3, no 3, pp. 110–128. doi: 10.5430/ijhe.v3n3p119.

Anastazi A., Urbina S. (2007) *Psihologicheskoe testirovanie* [Psychological Testing]. Saint Petersburg: Piter.

Charney D. (1984) The Validity of Using Holistic Scoring to Evaluate Writing: A Critical Overview. *Research in the Teaching of English*, vol. 18, no 1, pp. 65–81.

Dancey C. P., Reidy J. (2017) *Statistics without Maths for Psychology*. Upper Saddle River: Pearson.

Falchikov N. (1986) Product Comparisons and Process Benefits of Peer Group and Self-Assessments. *Assessment and Evaluation in Higher Education*, vol. 11, no 2, pp. 146–166. doi: 10.1080/0260293860110206.

Falchikov N., Goldfinch J. (2000) Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks. *Review of Educational Research*, vol. 70, no 3, pp. 287–322.

Gere A. R. (1980) Written Composition: Toward a Theory of Evaluation. *College English*, vol. 42, no 1, pp. 44–48, 53–58.

Huot B. (1990) The Literature of Direct Writing Assessment: Major Concerns and Prevailing Trends. *Review of Educational Research*, vol. 60, no 2, pp. 237–263.

Kaplan F., Bornet C. (2014) A Preparatory Analysis of Peer-Grading for a Digital Humanities MOOC. *Digital Humanities: Book of Abstracts.* Lausanne: University of Lausanne, pp. 227–229.

Linacre J. M. (1989) *Many-Faceted Rasch Measurement*. Chicago, IL: MESA.

Lunz M. E., Wright B. D., Linacre J. M. (1990) Measuring the Impact of Judge Severity on Examination Scores. *Applied Measurement in Education*, vol. 3, no 4, pp. 331–345.

Orpen C. (1982) Student versus Lecturer Assessment of Learning. *Higher Education*, vol. 11, no 5, pp. 567–572.

Shah D. (2016) *Monetization over Massiveness: Breaking down MOOCs by the Numbers in 2016*. Available at: https://www.edsurge.com/news/2016–12–

29-monetization-over-massiveness-breaking-down-moocs-by-the-numbers-in-2016 (accessed 10 October 2018).

Shah D. (2017) *Coursera's 2017: Year in Review.* Available at: https://www.class-central.com/report/coursera-2017-year-review/ (accessed 10 October 2018).

Shah D. (2018) *A Product at Every Price: A Review of MOOC Stats and Trends in 2017*. Available at: https://www.class-central.com/report/moocs-stats-and-trends-2017/ (accessed 10 October 2018).

Shmelev A. G. (2013) *Prakticheskaja testologija. Testirovanie v obrazovanii, prikladnoj psihologii i upravlenii personalom* [Practical test. Testing in education, applied psychology and human resource management]. Moscow: Maska.

Ueno M., Okamoto T. (2016) Item Response Theory for Peer Assessment. *IEEE Transactions on Learning Technologies*, vol. 9, no 2, pp. 157–170.

Wright B. D., Masters G. N. (1982) *Rating Scale Analysis: Rasch Measurement*. Chicago: MESA.