ISSN 1814-9545 (PRINT)
ISSN 2412-4354 (ONLINE)

ВОПРОСЫ ОБРАЗОВАНИЯ

Educational Studies Moscow

3 2023

Учредитель: Национальный исследовательский университет «Высшая школа экономики»

Вопросы образования / Educational Studies Moscow № 3, 2023
Ежеквартальный научно-образовательный журнал. Издается с 2004 г.
ISSN 1814-9545 (Print) ISSN 2412-4354 (Online)

Свидетельство о регистрации средства массовой информации ПИ № ФС77-68125
от 27 декабря 2016 г. выдано Федеральной службой по надзору в сфере связи,
информационных технологий и массовых коммуникаций

Содержание № 3, 2023

National Research University Higher School of Economics

Voprosy obrazovaniya / Educational Studies Moscow
No 3, 2023

established in 2004, is an academic journal published quarterly by the Higher School of Economics (HSE)

ISSN 1814-9545 (Print)
ISSN 2412-4354 (Online)

The mission of the journal is to provide a medium for professional discussion on a wide range of educational issues. The journal publishes original research and perceptive essays from Russian and foreign experts on education, development and policy. "Voprosy obrazovaniya / Educational Studies Moscow" strives for a multidisciplinary approach, covering traditional pedagogy as well as the sociology, economics and philosophy of education.

Conceptually, the journal consists of several parts:

- Theoretical materials and empirical research aimed at developing new approaches to understanding the functioning and development of education in modern society
- Papers on current projects, practical developments and policy debates in the field of education, written for professionals and the wider public
- Statistical data and case studies published as "information for reflection" with minimal accompanying text
- Information about and analysis of the latest pedagogical projects
- Reviews of articles published in international journals

Target audience: Leading Russian universities, government bodies responsible for education, councils from federal and regional legislatures, institutions engaged in education research, public organizations and foundations with an interest in education.

All papers submitted for publication in the "Voprosy obrazovaniya / Educational Studies Moscow" journal undergo peer review.
Distributed by subscription and direct order
Address
National Research University Higher School of Economics
20 Myasnitskaya St, Moscow, Russia 101000
Tel: +7 (495) 772 95 90 *15511 *15512
E-mail: edu.journal@hse.ru
Homepage: http://vo.hse.ru/en/

National Research University Higher School of Economics

Voprosy obrazovaniya / Educational Studies Moscow

Table of contents
No 3, 2023

Психометрические исследования: современные методы и новые возможности для образования

Елена Карданова, Алина Иванова

Статья поступила в редакцию в сентябре 2023 г.

Карданова Елена Юрьевна — кандидат физико-математических наук, научный руководитель центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики». E-mail: ekardanova@hse.ru. ORCID: https://orcid.org/0000-0003-2280-1258

Иванова Алина Евгеньевна — кандидат наук об образовании, старший научный сотрудник центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики». Адрес: 101000, Москва, Потаповский пер., 16, стр. 10. E-mail: aeivanova@hse.ru. ORCID: https://orcid.org/0000-0003-3340-7651 (контактное лицо для переписки)

Аннотация

Качественные измерения — фундаментальное требование к исследовательской практике в сфере социальных наук. Качество измерений определяет валидность интерпретаций и выводов, которые мы можем сделать, решений, которые мы можем принять на основе полученных в результате измерений данных. Для качественных измерений в социальных науках необходимы инструменты оценки, а также методы анализа данных, позволяющие связать наблюдаемые результаты измерений с теоретическими атрибутами. Научную основу для их разработки дает психометрика.

Предваряя специальный выпуск журнала «Вопросы образования / Educational Studies Moscow», посвященный психометрике, приглашенные редакторы этого выпуска освещают основные вехи истории психометрики, выделяют несколько исключительно значимых публикаций, отмечают профессиональные институции и авторов, которые внесли весомый вклад в развитие данной отрасли науки. Особое внимание авторы уделяют истории психометрики в России. Оценивая возможности, перспективы и ограничения психометрики, авторы высказывают свою точку зрения на дискуссионные вопросы, и она не всегда совпадает с мнением авторов специального выпуска.

В этом выпуске представлены примеры использования современных психометрических методов для решения актуальных проблем в исследованиях образования, а также в исследованиях на стыке образования и психологии, образования и разных сфер бизнеса. Всех авторов представленных статей объединяет стремление совершенствовать исследовательскую практику в социальных науках за счет по-настоящему качественных измерений.

Ключевые слова

психометрика, измерения в социальных науках, история психометрики

Для цитирования

Карданова Е.Ю., Иванова А.Е. (2023) Психометрические исследования: современные методы и новые возможности для образования. *Вопросы образования / Educational Studies Moscow*, № 3, сс. 8–19. https://doi.org/10.17323/vo-2023-17951

Psychometric Research: Modern Methods and New Opportunities for Education

Elena Kardanova, Alina Ivanova

Elena Yu. Kardanova — Candidate of Sciences (PhD) in Differential Equations, Dynamic Systems and Optimal Control; Scientific Supervisor, Centre for Psychometrics and Measurement in Education, Institute of Education, National Research University, Higher School of Economics. E-mail: ekardanova@hse.ru. ORCID: https://orcid.org/0000-0003-2280-1258

Alina E. Ivanova — Senior Researcher at the Center for Psychometrics and Measurements in Education, Institute of Education, National Research University Higher School of Economics. Address: Bld. 10, 16 Potapovsky Ln, 101000 Moscow, Russian Federation. E-mail: aeivanova@hse.ru. ORCID: https://orcid.org/0000-0003-3340-7651 (corresponding author)

Abstract

Qualitative measurement is a fundamental requirement for research practice in the social sciences. The quality of measurements determines the validity of the interpretations, conclusions and decisions we can make based on the data obtained from the measurements. Qualitative measurement in the social sciences requires assessment tools as well as data analysis methods to link observed measurement results to theoretical attributes. The scientific basis for their development is provided by psychometrics.

Preceding the special issue of the journal "Voprosy obrazovaniya / Educational Studies Moscow" devoted to psychometrics, the guest editors of this issue cover the main milestones of the history of psychometrics, highlight some significant publications, note the professional institutions and authors who have made their valuable contribution to the development of this branch of science. The authors pay special attention to the history of psychometrics in Russia. Assessing the possibilities, prospects and limitations of psychometrics, the authors express their point of view on the debatable issues of psychometrics, and it does not always coincide with the opinion of the authors of the special issue. This issue presents examples of using modern psychometric methods to solve actual problems in education research, as well as in research at the intersection of education and psychology, education and different spheres of business. All the authors of the presented articles are united by the desire to improve research practice in the social sciences through truly qualitative measurements.

Keywords

psychometrics, measurement in social sciences, history of psychometrics

For citing

Kardanova E.Yu., Ivanova A.E. (2023) Psikhometricheskie issledovaniya: sovremennye metody i novye vozmozhnosti dlya obrazovaniya [Psychometric Research: Modern Methods and New Opportunities for Education]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 3, pp. 8–19. https://doi.org/10.17323/vo-2023-17951

Психометрика создает теорию и методологию измерений в социальных науках, и поэтому область ее интересов обширна, а возможности применения очень широки. Психометрики разрабатывают инструменты оценки, методы измерения конструктов, предлагают и применяют формализованные модели, которые могут служить для связи наблюдаемых явлений

с теоретическими атрибутами. Иначе говоря, психометрика определяет принципы разработки инструментов измерений в социальных науках, а также принципы работы с данными измерений. Она позволяет получить надежные и валидные данные измерений, которые можно использовать для принятия управленческих решений или проверки исследовательских гипотез.

Психометрика возникла в конце XIX — начале XX в. из социального запроса на объективные и надежные методы измерения психологических характеристик людей. Начало развития психометрики связывают с именами Ф. Гальтона, Ч. Спирмена, А. Бине, Р. Кэттелла, которые создали первые тесты и заложили основы статистического анализа данных тестирования. В дальнейшем большой вклад в развитие психометрики внесли исследователи компании *Educational Testing Service*[1] в Принстоне, среди которых можно выделить Ф. Лорда, Б. Грина, А. Бирнбаума. Своими работами они заложили основы современной теории тестирования и дали ее статистическое обоснование.

В 1968 г. вышла книга *Statistical Theories of Mental Test Scores* [Lord, Novick, 1968]. Она, во-первых, содержала тщательное обоснование классической теории тестирования. Во-вторых, четыре главы этой книги, написанные А. Бирнбаумом, посвящены основам современной теории тестирования — *item response theory* (IRT), активно развивавшейся в предыдущее десятилетие. В-третьих, в книге рассмотрены практические вопросы тестирования, в частности массового. Позднее Ф. Лорд опубликовал книгу *Applications of Item Response Theory to Practical Testing Problems* [Lord, 1980], в которой рассмотрел опыт применения IRT в практике тестирования: как оцениваются параметры заданий и испытуемых, чем полезны на практике характеристическая и информационная функции задания и теста, как можно выровнять разные формы теста. В этой книге Ф. Лорд также описывал базовые идеи компьютерного адаптивного тестирования. В дальнейшем идеи IRT получили развитие в работе Р. Хамблтона с соавторами *Fundamentals of Item Response Theory* [Hambleton, Swaminathan, Rogers, 1991], в книге С. Эмбретсон и С. Рейса *Item Response Theory for Psychologists* [Embretson, Reise, 2000] и многих других изданиях.

Отдельная линия развития IRT связана с именем Георга Раша, датского математика, который изучал свойства моделей измерения и разработал семейство моделей, обладающих так называемой специфической объективностью: в таких моделях параметры испытуемого и задания полностью отделены. Свойства разработанных им моделей измерения Раш описал в книге *Probabilistic Models for Some Intelligence and Attainment*

[1] https://www.ets.org/

Tests [Rasch, 1960]. В дальнейшем они получили название моделей семейства Раша и вдохновили многих ученых. Последователи Раша расширили это семейство моделей и придали развитие его теории. Самым известным из последователей Раша является Б. Райт, который создал в Университете Чикаго лабораторию объективных измерений, где вместе с учениками проводил исследования в области Раш-моделирования. Б. Райт с коллегами написал две книги с изложением основ измерений в рамках моделей Раша: *Best Test Design* [Wright, Stone, 1979] и *Rating Scale Analysis* [Wright, Masters, 1982]. Многие из его учеников внесли большой вклад в развитие психометрики — Д. Эндрич, М. Вилсон, Дж. Мастерс, М. Линакр, К. Майфорд [Myford, Wolfe, 2003; 2004].

В 1935 г. Л.Л. Терстоун основал в Анн Арборе Психометрическое общество, которое сегодня признано наиболее авторитетным институтом, занимающимся психометрической наукой. В 2019 г. группа исследователей попыталась представить академическую генеалогию Психометрического общества [Wijsen et al., 2019] по аналогии с традиционным семейным генеалогическим древом. Отношения между научными руководителями и их учениками образовали пять отдельных ветвей, берущих начало от Джеймса Р. Энджелла, Вильгельма Вундта, Уильяма Джеймса, Карла Фридриха Гаусса или Альберта Мишотта.

В России начало развития психометрики датируют концом XIX в. и связывают с открытием первой Лаборатории психологических экспериментов под руководством В. Бехтерева в 1885 г. и первой Лаборатории экспериментальной педагогической психологии под руководством А. Нечаева в 1901 г. Оба ученых могут считаться учениками В. Вундта: Бехтерев в 1884–1885 гг. слушал в Лейпциге курс лекций Вундта по экспериментальной психологии и занимался в его лаборатории [Будилова, Кольцова, 1985], а Нечаев в 1898 г. во время стажировки в Европе работал в лабораториях Вундта в Лейпциге, а также у А. Бине в Париже [Сироткина, Смит, 2016]. В 1904 г. он открыл первые в России педологические курсы, на которых проводились занятия по технике психологического эксперимента и основам статистических методов. Таким образом, первые российские исследователи, применявшие психометрические методы в своей работе, тесно связаны с академическим древом мировой психометрики.

К сожалению, печально известное постановление 1936 г. «О педологических извращениях в системе наркомпросов»[2] на полвека затормозило развитие психометрики в нашей стране. Но в 1990-х годах в ходе адаптации системы образования в

[2] Постановление ЦК ВКП(б) «О педологических извращениях в системе наркомпросов». 4 июля 1936 г. См.: КПСС в резолюциях... Т. 6. 1933–1937 гг. М., 1985. С. 364–367.

России к новым реалиям постсоветской эпохи и возвращения страны в мировое научное поле интерес к количественным методам в социальных науках, и к психометрике в частности, начал быстро расти. Стали возникать локальные сообщества разработчиков и пользователей тестов, появились первые курсы повышения квалификации в области разработки тестов и первые методологические разработки, начали укрепляться международные связи. В 1995 г. вышла работа М.Б. Челышковой «Разработка педагогических тестов на основе современных математических моделей», в которой впервые на русском языке рассматриваются основные дихотомические модели современной теории тестирования. В 2002 г. появился Приказ Министерства образования РФ[3], в котором впервые нашей стране заданы стандарты разработки и оценки качества педагогических тестов. В этот же период появляются первые русскоязычные руководства по разработке тестов и первичному анализу их результатов — это книги В.С. Аванесова [1996] и М.Б. Челышковой [2002], а также переводы на русский язык известных за рубежом книг, таких как «Справочное руководство по конструированию тестов» [Клайн, 1994] и «Индивидуальные различия» [Купер, 2000].

Дальнейшее развитие психометрики в образовании в России связано с созданием Федерального центра тестирования (ФЦТ) и проведением им, начиная с 1995 г., централизованного тестирования — первого в нашей стране массового обследования на основе стандартизированных тестов и процедур тестирования, обработки, анализа и представления результатов. Осуществление такого тестирования потребовало разработки методологической базы, проведения научных исследований, объединения специалистов в области психометрики. ФЦТ проводил ежегодные конференции по тестированию в образовании, привлекавшие большое число участников со всей страны, выпускал журнал «Вопросы тестирования в образовании», в котором представлялись и методологические проблемы разработки тестов в образовании, и результаты первых в стране научных исследований в области психометрики. В 2000-х сотрудники ФЦТ издали две книги, посвященные теории и практике современной теории тестирования, — «Введение в теорию моделирования и параметризации педагогических тестов» [Нейман, Хлебников, 2000] и «Моделирование и параметризация тестов: основы теории и приложения» [Карданова, 2008].

Наконец, в 2002 г. в связи с подготовкой и введением Единого государственного экзамена в стране появился специализированный институт — Федеральный институт педагогических

[3] Приказ Минобразования РФ от 17 апреля 2000 г. № 1122 «О сертификации качества педагогических тестовых материалов».

измерений (ФИПИ), основной задачей которого стало создание контрольно-измерительных материалов для государственной итоговой аттестации.

Основное препятствие на пути развития психометрики в России — это огромный дефицит профессиональных кадров. Как ответ на этот вызов, в 2010 г. в Институте образования НИУ ВШЭ создана магистерская программа «Измерения в психологии и образовании», которая в 2020 г. преобразована в программу «Обучение и оценивание как наука». До сих пор она является единственной на постсоветском пространстве программой высшего образования, готовящей специалистов в области разработки инструментов измерения в образовании и других социальных науках, психометрики, анализа данных с применением современных методов статистики, машинного обучения. За время существования программа подготовила большое число специалистов в области психометрики, и мы рады, что большинство авторов статей нашего спецвыпуска, прошедших строгий отбор и рецензирование, — выпускники нашей магистерской программы.

Цель предлагаемого вниманию читателей специального выпуска журнала «Вопросы образования / Educational Studies Moscow» — предоставить широкому кругу авторов из разных организаций, разных сфер и уровней образования (и не только) возможность ознакомить профессиональное сообщество с примерами использования современных психометрических методов для решения актуальных проблем в исследованиях, проводимых в образовании, а также на стыке образования и психологии, образования и разных сфер бизнеса. При этом, как нам кажется, всех авторов этого выпуска объединяет стремление совершенствовать исследовательскую практику в сфере социальных наук за счет по-настоящему качественных измерений. Именно качественные измерения определяют валидность интерпретаций и выводов, которые мы можем сделать, решений, которые мы можем принять, применяя тот или иной инструмент измерения. Сегодня валидными мы считаем те измерения, для которых собранные доказательства и теория вместе поддерживают интерпретацию результатов тестирования для заявленных целей использования этих тестов [American Educational Research Association et al., 2014]. Не будет преувеличением сказать, что именно соображения валидности составляют сердцевину разработки тестов и оценки их качества.

Методы психометрики, в частности современная теория тестирования и конфирматорный факторный анализ, имеют долгую историю применения в исследовательской практике, они многократно усовершенствованы и доказали свою полезность и высокое качество производимых ими данных. Оба метода

предоставляют уникальную информацию о функционировании отдельных заданий и о качестве тестов или опросников в целом и вносят важный вклад в состав свидетельств валидности результатов измерения, подтверждая возможность использования тех или иных инструментов измерения в научных и практических целях.

Авторы специального выпуска напрямую или косвенно затрагивают проблему валидности, дискутируя об истории психологического знания, полученного с помощью и без помощи психометрики, рассуждая о природе и тонкостях психометрического моделирования, применяя возможности психометрического моделирования для поиска ответов на широкий круг исследовательских вопросов.

Авторы статьи «Апробация шкалы установок к межкультурному обучению в вузе» М. Бульцева и С. Берриос Кальехас считают такие установки критически важным условием развития межкультурной компетентности студентов. Ввиду недостатка надежных средств измерения студенческих установок к межкультурному обучению авторы предлагают свой инструмент. Приведя подробный анализ концептов межкультурного обучения и межкультурной компетентности, авторы описывают процесс апробации Шкалы установок к межкультурному обучению в вузе и используют конфирматорный факторный анализ для ее валидизации. Стоит отметить, что редакторы данного спецвыпуска придерживаются иного подхода в понимании валидности (мы уже привели его выше). Этот подход адресует скорее к работам С. Мессика [Messick, 1995], М. Кейна [Kane, 2001], Р. Мислеви [Mislevy, 2007], которые показали, что о валидности инструмента говорить некорректно. Вслед за ними мы полагаем, что доказательство валидности базируется на последовательном сборе ее свидетельств, на поиске аргументов, поддерживающих или опровергающих интерпретацию и использование результатов тестирования в заданных целях. Тем не менее многие исследователи, в том числе в смежных сферах социальных наук, например в менеджменте, придерживаются иной концепции валидности, применяют наборы статистических показателей валидности и оценивают дискриминантную, конвергентную и другие виды валидности инструментов. Надеемся, что читатели этого спецвыпуска выберут близкую для них оптику, в которой им удобнее будет работать самим и оценивать результаты исследований коллег.

В статье «Опыт использования бифакторных моделей для снижения эффектов социальной желательности на материале нормативного опросника универсальных компетенций» Е. Сагитов, И. Брун и С. Павлов используют метод конфирматорного факторного анализа для решения давней проблемы психо-

логического тестирования — искажения итоговых баллов по измеряемым конструктам, обусловленного социальной желательностью ответов респондентов. Авторы описывают метод внесения корректировок в итоговые баллы респондентов на примере своего нормативного опросника универсальных компетенций. Представленный ими подход к обработке и анализу данных позволяет минимизировать эффект социальной желательности при измерении психологических конструктов.

И. Антипкина и А. Иванова в статье «Декомпозиция трудности заданий в тесте читательской грамотности» выясняют, что именно составляет трудность заданий в тесте, проверяющем навык чтения у учеников начальной школы. Авторы показывают возможность с помощью современной теории тестирования не просто описать тестовое поведение испытуемых и их результаты по тесту, но и объяснить, какие характеристики заданий могут обусловливать степень их трудности. Применяемая в работе линейная логистическая тестовая модель с ошибкой (LLTM+e) использует характеристики задания как предикторы вероятности верного решения этого задания. С помощью предложенного методологического подхода авторы проверяют предположение о том, что при контроле «внешних» характеристик заданий (например, формата), параметры трудности, связанные с заложенными в задании группами читательских умений, будут образовывать иерархию — от поиска информации, данной в явном виде (наиболее простые задания) до оценивания текста в целом (наиболее трудные задания). Статья наглядно показывает, что гипотезы исследователей не всегда оказываются верны, но полученные результаты не становятся при этом менее интересными или полезными с точки зрения разработки тестовых материалов.

С. Тарасов, И. Зуева и Д. Федерякин в статье «Измерение образовательного прогресса на основе когнитивных операций» также используют модель LLTM, но значительно обогащают ее благодаря синтезу с одной из моделей для измерения образовательного прогресса — моделью Андерсена. Авторы отмечают, что, несмотря на все технологические и методологические усовершенствования в области исследований образования в последние годы, измерение динамики достижений учащихся, их образовательного прогресса остается нетривиальной методологической задачей. И чтобы эту задачу решить, авторы предлагают измерять образовательный прогресс с помощью когнитивных операций, освоение которых заранее закладывается и затем проверяется в тесте. Таким образом можно не просто измерить образовательный прогресс, но и существенно расширить возможности интерпретации тестовых баллов учеников как раз за счет когнитивных операций. Для иллюстрации

предлагаемого подхода использована линейка тестов, применявшихся для мониторинга образовательного прогресса в математике у учащихся 8–9-х классов средней школы.

Статья Д. Грачевой «Роль контекста в заданиях сценарного типа при измерении универсальных навыков: применение теории генерализации» также иллюстрирует методологические достижения психометрики в решении практических задач оценивания в образовании. В современных быстро меняющихся и все более «цифровизирующихся» условиях большое внимание в образовании уделяется развитию и оцениванию универсальных навыков у школьников. Для такого оценивания необходимы новые тестовые форматы, основанные на наблюдаемых действиях учащегося в цифровой среде. Однако эти новые и, как правило, технологически сложные контекстуализированные форматы тестов несут с собой новые методологические вызовы. Каков вклад контекста в результаты оценивания универсальных компетенций? Какое количество контекстов сценарных заданий необходимо для надежного измерения универсальных навыков? Автор ищет ответы на эти вопросы, опираясь на методы теории генерализации. В статье содержится детальное описание теории генерализации и дизайна проведения анализа, она может стать методическим пособием для всех заинтересованных в изучении и применении психометрики в образовательных исследованиях.

Авторы еще двух статьей этого специального выпуска ставят важные теоретические вопросы, открывая дискуссию о перспективах и ограничениях психометрических исследований. В статье «Психометрика и когнитивные исследования: противоречия и возможности кооперации» Ю. Кузьмина рассматривает историю взаимоотношений между экспериментальными исследованиями и психометрикой с конца XIX в. до настоящего времени. Когнитивная психология в основном развивалась в рамках экспериментальной парадигмы, в отличие от психометрики, занимающейся оценкой индивидуальных различий и корреляционными исследованиями. Автор статьи показывает, как возник и ширился разрыв между когнитивными исследованиями и психометрикой и как он связан с разной исследовательской логикой в этих двух подходах. Важно, что этот разрыв не предопределен — а следовательно, может быть преодолен. Опираясь на огромный пласт научных работ, автор показывает, что на всем протяжении развития психологии многие исследователи подчеркивали возможность сближения рассматриваемых подходов, которое способно обогатить психологию в целом.

В статье «Так ли полезна психометрика для академической психологии?» Ю. Тюменева анализирует логику психометрического моделирования как способа репрезентировать латент-

ный конструкт или обнаружить его структуру. Психологические теории относительно способностей и личностных черт часто полагаются на результаты психометрического моделирования. Ю. Тюменева задается вопросом: а действительно ли психометрическое моделирование является моделированием в общенаучном значении этого термина? Цепь логических рассуждений приводит автора к отрицательному ответу на этот вопрос. Поэтому, утверждает Ю. Тюменева, на основании психометрического моделирования можно делать выводы только о структуре тестовых данных, но не о структуре латентного конструкта.

Начиная обзор статей специального выпуска, мы отмечали, что наше видение как редакторов этого выпуска предложенных в статье идей и перспектив совсем не обязательно совпадает с видением авторов. Но мы верим, что именно различия во взглядах исследователей и возможность открыто обсуждать эти различия формируют академическую свободу.

Наконец, завершается специальный выпуск рецензиями на две выдающиеся, с нашей точки зрения, книги в области психометрики, изданные в течение последних десяти лет. И. Угланова знакомит читателей с книгой Р. Леви и Р. Мислеви *Bayesian Psychometric Modeling* (2016), посвященной байесовскому подходу к психометрике. К. Тарасова и Д. Грачева делятся своим видением психометрических перспектив, представленных А. фон Давьер, Р. Мислеви и Дж. Хао в книге *Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment. With Examples in R and Python* (2022).

Мы надеемся, что этот специальный выпуск, в котором обсуждаются возможности, перспективы и ограничения применения психометрики в сфере образования, подарит читателям интеллектуальное удовольствие — такое же, какое он подарил нам во время его планирования, подготовки и публикации. Мы благодарим команду редакции журнала «Вопросы образования / Educational Studies Moscow» за предоставленную возможность провести интересную научную дискуссию на страницах журнала. И конечно же, мы благодарим авторов и рецензентов за проделанную работу и неоценимый вклад в эту дискуссию.

Литература

1. Аванесов В.С. (1996) *Композиция тестовых заданий*. М.: Ассоциация инженеров-педагогов.
2. Будилова Е.А., Кольцова В.А. (1985) 100-летие первой русской экспериментальной психологической лаборатории. *Вопросы психологии*, № 6, сс. 96–102.
3. Карданова Е.Ю. (2008) *Моделирование и параметризация тестов: основы теории и приложения*. М.: Федеральный центр тестирования.
4. Клайн П. (1994) *Справочное руководство по конструированию тестов. Введение в психометрическое проектирование*. Киев: ПАН.

5. Купер К. (2000) *Индивидуальные различия*. М.: Аспект.
6. Нейман Ю.М., Хлебников В.А. (2000) *Введение в теорию моделирования и параметризации педагогических тестов.* М.: Прометей.
7. Сироткина И., Смит Р. (2016) *История психологии в России: краткий очерк с авторскими акцентами. Препринт WP6/2016/01*. М.: НИУ ВШЭ.
8. Челышкова М.Б. (2002) *Теория и практика конструирования педагогических тестов*. М.: Логос.
9. Челышкова М.Б. (1995) *Разработка педагогических тестов на основе современных математических моделей*. М.: МИСИС.
10. American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.) (2014) *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
11. Embretson S.E., Reise S.P. (2000) *Item Response Theory for Psychologists*. Mahwah, NJ: Erlbaum.
12. Hambleton R.K., Swaminathan H., Rogers H.J. (1991) *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.
13. Kane M.T. (2001) Current Concerns in Validity Theory. *Journal of Educational Measurement*, vol. 38, no 4, pp. 319–342. http://dx.doi.org/10.1111/j.1745-3984.2001.tb01130.x
14. Lord F.M. (1980) *Applications of Item Response Theory to Practical Testing Problems*. New York, NY: Routledge. doi.org/10.4324/9780203056615
15. Lord F.M., Novick M.R. (1968) *Statistical Theories of Mental Test Scores*. Menlo Park, CA: Addison-Wesley.
16. Messick S. (1995) Standards of Validity and the Validity of Standards in Performance Assessment. *Educational Measurement: Issues and Practice*, vol. 14, no 4, pp. 5–8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x
17. Mislevy R.J. (2007) Validity by Design. *Educational Researcher*, vol. 36, no 8, pp. 463–469.
18. Myford C.M., Wolfe E.W. (2004) Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part 2. *Journal of Applied Measurement,* vol. 5, no 2, pp. 189–227.
19. Myford C.M., Wolfe E.W. (2003) Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part 1. *Journal of Applied Measurement,* vol. 4, no 4, pp. 386–422.
20. Rasch G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danmarks Paedagogiske Institut.
21. Wijsen L.D., Borsboom D., Cabaço T., Heiser W.J. (2019) An Academic Genealogy of Psychometric Society Presidents. *Psychometrika*, vol. 84, no 2, pp. 562–588. http://dx.doi.org/10.1007/s11336-018-09651-4
22. Wright B.D., Masters G.N. (1982) *Rating Scale Analysis. Rasch Measurement*. Chicago, IL: Mesa.
23. Wright B.D., Stone M.N. (1979) *Best Test Design. Rasch Measurement*. Chicago, IL: Mesa.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.) (2014) *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Avanesov V.S. (1996) *Kompozitsiya testovykh zadaniy* [Composition of Test Tasks]. Moscow: Association of Engineers-Teachers.

Budilova E.A., Koltsova V.A. (1985) 100-letie pervoy russkoy eksperimental´noy psikhologicheskoy laboratorii [100th Anniversary of the First Russian Experimental Psychological Laboratory]. *Voprosy Psychologii*, no 6, pp. 96–102.

Chelyshkova M.B. (2002) *Teoriya i praktika konstruirovaniya pedagogicheskikh testov* [Theory and Practice of Designing Pedagogical Tests]. Moscow: Logos.

Chelyshkova M.B. (1995) *Razrabotka pedagogicheskikh testov na osnove sovremennykh matematicheskikh modeley* [Development of Pedagogical Tests Based on Modern Mathematical Models]. Moscow: MISIS.

Cooper C. (2000) *Individual´nye razlichiya* [Individual Differences]. Moscow: Aspekt.

Embretson S.E., Reise S.P. (2000) *Item Response Theory for Psychologists*. Mahwah, NJ: Erlbaum.

Hambleton R.K., Swaminathan H., Rogers H.J. (1991) *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Kane M.T. (2001) Current Concerns in Validity Theory. *Journal of Educational Measurement*, vol. 38, no 4, pp. 319–342. http://dx.doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kardanova E.Yu. (2008) *Modelirovanie i parametrizatsiya testov: osnovy teorii i prilozheniya* [Modeling and Parameterization of Tests: Fundamentals of Theory and Applications]. Moscow: Federal Testing Center.

Klein P. (1994) *Spravochnoe rukovodstvo po konstruirovaniyu testov. Vvedenie v psikhometricheskoe proektirovanie* [Reference Guide for Designing Tests. Introduction to Psychometric Design]. Kiev: PAN.

Lord F.M. (1980) *Applications of Item Response Theory to Practical Testing Problems*. New York, NY: Routledge. doi.org/10.4324/9780203056615

Lord F.M., Novick M.R. (1968) *Statistical Theories of Mental Test Scores*. Menlo Park, CA: Addison-Wesley.

Messick S. (1995) Standards of Validity and the Validity of Standards in Performance Assessment. *Educational Measurement: Issues and Practice*, vol. 14, no 4, pp. 5–8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x

Mislevy R.J. (2007) Validity by Design. *Educational Researcher*, vol. 36, no 8, pp. 463–469.

Myford C.M., Wolfe E.W. (2004) Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part 2. *Journal of Applied Measurement,* vol. 5, no 2, pp. 189–227.

Myford C.M., Wolfe E.W. (2003) Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part 1. *Journal of Applied Measurement,* vol. 4, no 4, pp. 386–422.

Neiman Yu.M., Khlebnikov V.A. (2000) *Vvedenie v teoriyu modelirovaniya i parametrizatsii pedagogicheskikh testov* [Introduction to the Theory of Modeling and Parametrization of Pedagogical Tests]. Moscow: Prometey.

Rasch G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danmarks Paedagogiske Institut.

Sirotkina I., Smith R. (2016) *Istoriya psikhologii v Rossii: kratkiy ocherk s avtorskimi aktsentami. Preprint WP6/2016/01* [History of Psychology in Russia: Short Review with the Authors' Emphasis. Working paper no WP6/2016/01]. Moscow: HSE.

Wijsen L.D., Borsboom D., Cabaço T., Heiser W.J. (2019) An Academic Genealogy of Psychometric Society Presidents. *Psychometrika*, vol. 84, no 2, pp. 562–588. http://dx.doi.org/10.1007/s11336-018-09651-4

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

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

Апробация Шкалы установок к межкультурному обучению в вузе

Мария Бульцева, Соня Алехандра Берриос Кальехас

Статья поступила в редакцию в феврале 2023 г.

Бульцева Мария Александровна — кандидат психологических наук, научный сотрудник Центра социокультурных исследований, Национальный исследовательский университет «Высшая школа экономики». 101000, Москва, ул. Мясницкая, 20. E-mail: mbultseva@hse.ru. ORCID: https://orcid. org/0000-0002-5899-9916 (контактное лицо для переписки)

Берриос Кальехас Соня Алехандра — аспирант, стажер-исследователь Центра социокультурных исследований, Национальный исследовательский университет «Высшая школа экономики». ORCID: https://orcid. org/0000-0002-9572-2289

Аннотация

Успешность современного человека во многом зависит от способности справляться с вызовами культурного разнообразия. Перед вузами стоит задача развивать у студентов межкультурную компетентность как одну из ключевых. Для этого необходимо организовать межкультурное обучение, эффективность которого зависит от установок студентов к нему. Проведено исследование с целью разработки и апробации инструментария по оценке установок к межкультурному обучению в вузе. В основу создания опросника положены представления о доступных студентам источниках такого обучения, о процессах, входящих в состав межкультурного обучения, а также о структуре конструкта «установка». Опросник состоит из четырех субшкал, каждая из которых оценивает три компонента установок: когнитивный, аффективный и поведенческий. Надежность и валидность инструментария подтверждены на выборке из 399 студентов российских вузов при помощи таких статистических процедур, как эксплораторный и конфирматорный факторный анализ, анализ согласованности шкал по критериям альфа Кронбаха и омега МакДональда, анализ показателей извлеченной средней дисперсии (AVE) и соотношения гетеропризнаков и монопризнаков. Обсуждаются перспективы дальнейших исследований межкультурного обучения, адаптация и использование разработанных шкал не только в системе высшего образования, но и в других сферах деятельности и разных возрастных группах.

Ключевые слова

межкультурное обучение, социальные установки, валидизация шкалы, вуз, студенты

Для цитирования

Бульцева М.А., Берриос Кальехас С.А. (2023) Апробация Шкалы установок к межкультурному обучению в вузе. *Вопросы образования / Educational Studies Moscow*, № 3, сс. 20–61. https://doi.org/10.17323/vo-2023-16819

Approbation of the Scale of Attitudes towards Intercultural Learning in Higher Education

Maria Bultseva, Sonia Alejandra Berrios Callejas

Maria A. Bultseva — PhD in Psychology, Research Fellow at the Centre for Sociocultural Research, National Research University Higher School of Economics. Address: 20 Myasnitskaya St, 101000 Moscow, Russian Federation. E-mail: mbultseva@hse.ru. ORCID: https://orcid. org/0000-0002-5899-9916 (corresponding author)

Sonia Alejandra Berrios Callejas — PhD student, Research Intern at the Centre for Sociocultural Research, National Research University Higher School of Economics. ORCID: https://orcid. org/0000-0002-9572-2289

Abstract

The ability to cope with the challenges of cultural diversity is one of the key competencies that students need to develop in the contemporary world. This actualizes the question of intercultural learning and students' attitudes towards it. The study aimed at development and validation of the instrument to measure students' attitudes towards intercultural learning. Taking into account the "sources" and processes of intercultural learning available to students, as well as the structure of the attitude, an instrument consisting of four scales were created — each of them includes three aspects of the attitudes: cognitive, affective and behavioral ones. The reliability and validity of each of the instrument were confirmed on a sample of 399 students of Russian universities using statistical procedures such as exploratory and confirmatory factor analysis, analysis of the consistency of the scales according to Cronbach's alpha and McDonald's omega, analysis of extracted mean variance and heterotrait-monotrait ratio. It seems promising to consider various aspects of intercultural learning in further research, adapting and using the scales not only in the higher education system, but also in other contexts and age groups.

Keywords

intercultural learning, social attitudes, scale validation, university, students

For citing

Bultseva M.A., Berrios Callejas S.A. (2023) Aprobatsiya shkaly ustanovok k mezhkul´turnomu obucheniyu v vuze [Approbation of the Scale of Attitudes towards Intercultural Learning in Higher Education]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 3, pp. 20–61. https://doi.org/10.17323/vo-2023-16819

Что стимулирует студентов включаться в обучение, что помогает им сохранять интерес к учебе в течение долгого времени, что заставляет проявлять самостоятельность и инициативу в учебном процессе? Эти вопросы активно изучает современная психология образования. В качестве одного из важных мотивационных факторов исследователи рассматривают установки к обучению [Kara, 2010]. Установка — это психологическая предрасположенность к тому, чтобы оценивать объект определенным образом [Haddock, Maio, 2008], чаще всего установки различают на основании одобрения или неодобрения предмета установки. Установки к обучению, соответственно, отражают

отношение учащихся к предмету обучения и их поведение в процессе обучения [Mašić, Bećirović, 2021].

Установки к обучению могут влиять на учебные достижения [Buschenhofen, 1998], а их формирование необходимо для успешной реализации учебных программ. Исходя из целей педагогической практики выделяют позитивные и негативные установки к обучению. Они представляют собой предпосылку включенности студентов в обучение [Şen, 2013]. Так, негативные установки и ожидания по отношению к обучению снижают мотивацию и препятствуют образовательному успеху. Позитивные установки способствуют включенности в изучение темы [Kara, 2010], усиливают открытость студентов к новым знаниям, снижают их уровень тревожности — а значит, помогают обрабатывать и структурировать новую информацию наиболее эффективным образом.

Так как студенты обучаются разным дисциплинам и навыкам, в современной образовательной науке установки к обучению часто изучаются в отношении определенной дисциплины или образовательного направления. Например, исследуются установки к онлайн-обучению и обучению на дому [Abirin, 2023; Uyar, 2023], установки к науке и изучению естественнонаучных дисциплин [Kurt, Benzer, 2020], много работ посвящено изучению установок к изучению иностранных языков [Rahman et al., 2021].

Однако процесс обучения не ограничивается дисциплинами, включенными в учебные планы. Глобализация и возрастание культурного разнообразия в обществе [Czaika, de Haas, 2014] выдвигают перед высшими учебными заведениями задачу культурно сенситивного развития студентов. С целью повышения межкультурной компетентности студентов вузы активно вовлекаются в процесс интернационализации образовательной среды: способствуют студенческой мобильности, вносят изменения в учебные программы и принципы функционирования кампусов [Gregersen-Hermans, 2017]. Межкультурные контакты могут мотивировать человека к адаптации, совершенствованию своих межкультурных знаний и навыков, когнитивному развитию, только если человек способен преодолеть вызовы нового опыта и принять новую культурно специфическую информацию [Crisp, Turner, 2011]. То есть ключевым механизмом, определяющим влияние межкультурного опыта на жизнь конкретного человека, является межкультурное обучение [Leung, Chiu, 2010]. Межкультурное обучение — это динамический коммуникативный процесс усвоения знаний о нормах, традициях, ценностях другой культуры [Xu, Chen, 2017], направленный на развитие готовности и способности представителей разных культур понимать друг друга и жить вместе [O'Brien et al., 2019].

1. Межкультурное обучение: источники, установки, результаты

Межкультурное обучение имеет много общего с изучением иностранных языков: в процессе освоения нового языка студенты тоже не только получают лингвистические знания и навыки, но и формируют свое отношение к людям, говорящим на других языках [Stern, 1983]. Установки к изучению иностранных языков включают убеждения учащегося относительно изучаемой языковой группы и относительно собственной культуры [Tahaineh, Hana, 2013]. С установками к обучению и к изучению иностранных языков тесно связано понятие коммуникативной межкультурной компетентности [Tran, Seepho, 2022]. С точки зрения углубления понимания детерминантов включенности студентов в изучение языков и культур их носителей исследователи считают целесообразным рассматривать установки как многосоставный конструкт [Mašić, Bećirović, 2021].

Таким образом, установки являются устойчивой системой убеждений, в которой выделяются когнитивный, аффективный и поведенческий компоненты [Gawronski, 2007]. Когнитивный компонент отражает представления об объекте установки; в контексте межкультурного обучения в составе когнитивного компонента можно рассматривать, например, мнение студента о том, насколько полезно изучать другие культуры на курсах, посвященных межкультурной коммуникации. Аффективный компонент установки — это чувства и эмоции, ассоциируемые с объектом установки; в контексте межкультурного обучения это, например, удовольствие или страх, интерес или скука, которые переживает студент, общаясь с представителями других культур. Поведенческий компонент установки описывает намерения и поступки, связанные с объектом установки: готов ли, например, студент посещать внеучебные мероприятия, чтобы узнать что-то новое о других культурах.

При безусловном сходстве с изучением иностранных языков межкультурное обучение охватывает более широкий набор знаний и навыков, поэтому данный феномен и подходы к его исследованию необходимо рассматривать отдельно. Чаще всего результаты межкультурного обучения оцениваются косвенно, на основании тех социально-психологических преобразований, к которым оно приводит. Межкультурное обучение в таком понимании воплощается в первую очередь в росте межкультурной компетентности. В самом широком смысле под межкультурной компетентностью имеется в виду сформированная способность жить, работать и отдыхать в условиях культурных различий, существующих в повседневной жизни [Мацумото, 2003]. В современной психологии и коммуникационных исследованиях сложились разные подходы к определению межкультурной компетентности [Spitzberg, Changnon, 2009]. Многообразие подходов реализуется и в исследованиях меж-

культурного обучения. Его операционализируют через его результаты, такие как рост межкультурной компетентности [Fenech, Baguant, Abdelwahed, 2020; Gondra, Czerwionka, 2018; King, Perez, Shim, 2013; Lane, 2012; Shao, Crook, 2015], выбор инклюзивных стратегий аккультурации и успешная адаптация к новой культуре или культурному разнообразию [Ward et al., 1998; Wilson, Ward, Fischer, 2013], формирование культурной сенситивности [Ramirez, 2021] и позитивных установок к мультикультурализму [Ari, Laron, 2013]. Мультикультурализм — одна из межгрупповых идеологий, т.е. система убеждений относительно природы культурных и этнических групп. Эти убеждения определяют, как люди воспринимают эти группы и ведут себя с их представителями [Rosenthal, Levy, 2010]. Поддержку определенных межгрупповых идеологий можно рассматривать как результат межкультурного обучения.

Однако рассмотрение межкультурного обучения только с точки зрения его результатов существенно ограничивает возможности исследователей и практиков в выявлении источников и психологических механизмов его формирования. Межкультурное обучение — сложный системный процесс [Chen, Ju, 2005], глубоко интегрированный в контекст, так что далее мы будем рассматривать его конкретно в условиях образовательной среды вуза.

Теория культурного обучения в основном описывает процесс инкультурации у детей [Tomasello, Kruger, Ratner, 1993], т.е. освоение ими собственной культуры. Однако ее основные принципы применимы и к межкультурному обучению. Культурное обучение рассматривается как социально-когнитивный навык, который приобретается посредством разных видов коммуникации. Выделяются три способа культурного обучения: имитация, инструктивное обучение, при котором ученика целенаправленно обучает кто-то другой, например учитель, преподаватель, ментор, и совместное обучение. Эти способы обучения можно реализовать и в вузе, используя возможности образовательных планов (*curriculum*).

Образовательные планы могут быть формальными и неформальными [Lane, 2012], они предусматривают комплекс академической и социокультурной деятельности [McKay, O'Neill, Petrakieva, 2016], вплетенной в общую стратегию интернационализации образования [Ippolito, 2007]. В результате многочисленных исследований межкультурного обучения в образовательной среде установлено, какие возможности взаимодействия с другими культурами предлагают вузы своим студентам — их можно условно назвать источниками межкультурного обучения:

- официальные учебные курсы и разнообразные виды деятельности, включенные в них;

- внеучебные межкультурные мероприятия;
- обучение через общение;
- глубокое погружение в другую культуру.

Официальные учебные курсы и разнообразные виды деятельности на них составляют формальное межкультурное обучение. В теории культурного обучения им соответствует инструктивное обучение, эти курсы входят в состав формального учебного плана. Исследования, в частности, показали эффективность с точки зрения роста межкультурной компетентности таких видов деятельности на лекциях и семинарах, посвященных культурам и межкультурной коммуникации, как дискуссии, ролевые игры, чтение статей, решение кейсов (критические инциденты), подготовка презентаций о разных культурах [Mahoney, Schamber, 2004; Sizoo, Serrie, 2004; Zhu, Bargiela-Chiappini, 2013]. Развитию межкультурной компетентности способствует также непосредственное инструктивное обучение по межкультурной тематике в образовательной организации [Klak, Martin, 2003; Constantin, Cohen-Vida, Popescu, 2015].

Внеучебные межкультурные мероприятия предоставляют студентам разнообразные возможности получить новые знания о других культурах и навыки межкультурной коммуникации в университете, но вне предметов, включенных в учебную программу. Они представляют собой неформальное межкультурное обучение. Установлено, что росту межкультурной компетентности способствуют гостевые лекции, культурные воркшопы, мероприятия, посвященные той или иной культуре, например праздники, тематические ланчи [Leask, 2009; McKay et al., 2016]. В организации внеучебных межкультурных мероприятий могут использоваться как методы инструктивного обучения, так и элементы совместного обучения или имитации.

Обучение через общение дополняет первый и второй источники межкультурного обучения. В теории культурного обучения ему соответствуют процессы имитации и совместного обучения. Развитию межкультурной компетентности способствуют учеба совместно с представителями других культур, наблюдение за их поведением, сравнение их паттернов поведения с привычными, менторство и помощь инокультурным студентам, выполнение групповых заданий в классе и дома, в том числе в онлайн-формате, интенсивная коммуникация с инокультурными студентами, в частности на внеучебные темы [Lin, Shen, 2020; Lu et al., 2017]. Межкультурное общение, результатом которого становится понимание убеждений, мотивации, системы ценностей представителей другой культуры, стимулирует и облегчает межкультурную коммуникацию [Gudykunst, Ting-Toomey, 1988]. Интенсивные дружеские контакты с пред-

ставителями культур, отличных от собственной, — значимая составляющая процесса обучения, так как они повышают культурную осведомленность и желание учиться у других [Bazron, Osher, Fleischman, 2005].

Глубокое погружение в другую культуру достигается, если вуз обеспечивает возможность студенческой мобильности. Экскурсии и экспедиции, краткосрочные и долгосрочные зарубежные стажировки способствуют межкультурному обучению и развитию межкультурных компетенций [Bennett, Salonen, 2007; Gondra, Czerwionka, 2018]. В рамках данного источника обучения в полной мере реализуются имитация и совместное обучение, а также могут присутствовать элементы инструктивного обучения.

Позитивные последствия межкультурных контактов и значимого культурного опыта широко освещены в научной литературе [Chan et al., 2017; Lee at el., 2014]. Однако чаще всего в этих исследованиях оцениваются результаты именно погружения в другую культуру, а не опыта совместного с представителями иных культур обучения в собственной стране [O'Brien et al., 2019]. Возможность участвовать в зарубежных стажировках получает небольшая часть студентов, но всем доступны источники межкультурного обучения в собственной стране: формальное и неформальное обучение, а также обучение через общение способствуют развитию межкультурной компетентности [Brown et al., 2007; Griffiths, Kopanidis, Steel, 2018; Jarosiński, Kozma, Sekliuckiene, 2021; Kulich, Wang, 2015].

Для эффективного межкультурного обучения недостаточно конвенциональных инструктивных методов, обучение должно включать разные элементы и не только обеспечивать получение знаний, но и стимулировать когнитивные и эмоциональные изменения у студентов [Binkley, Minor, 2021]. Поэтому процесс межкультурного обучения часто рассматривается в терминах теорий трансформационного обучения и экспириентального обучения.

Трансформационное обучение предполагает не просто получение знаний, но качественное изменение представлений, свойственных человеку [Cranton, 2023; Mezirow, 2003]. В условиях межкультурного обучения возможен первоначальный дискомфорт, обусловленный столкновением с новой культурой [Santoro, Major, 2012; Pence, Macgillivray, 2008]. Он может быть вызван стрессом аккультурации или когнитивным диссонансом [Chinnappan, McKenzie, Fitzsimmons, 2013; Mitchell, Paras, 2018]. Преодоление такого дискомфорта будет способствовать эффективности межкультурного обучения [Chwialkowska, 2020; Dai, Garcia, 2019; DeRobertis, Bland, 2020].

Теория экспириентального обучения описывает, что происходит с человеком при проживании определенного опыта в

той или иной сфере [Kolb, Kolb, 2022]. В ходе обучения выделяются этапы: включения в конкретный опыт, наблюдения и рефлексии, выстраивания догадок, т.е. абстрактного теоретизирования и создания собственных концепций и интерпретаций, и экспериментирования — попыток на практике применить свои новые знания. Применительно к межкультурному обучению вовлеченность в эти четыре процесса при столкновении с новой культурно специфической информацией приводит к развитию межкультурной компетентности [Yamazaki, Kayes, 2004; Zhu et al., 2017].

Таким образом, оказываясь в ситуации межкультурного взаимодействия, которое может быть представлено разными источниками обучения, студенты становятся более компетентными посредством обучения, т.е. включаясь в новый опыт, рефлексируя, выстраивая и проверяя свои догадки о других культурах и оттачивая новые знания и навыки в поведении. Но всегда ли простое присутствие источника обучения актуализирует процессы межкультурного обучения и приводит к повышению межкультурной компетентности? Наличие контактов с представителями других культур или информации о других культурах не означает, что человек обязательно усвоит знания и навыки или пройдет через трансформацию установок и представлений и, как следствие, реализует потенциал межкультурного обучения [Бульцева, 2020]. Надлежащее межкультурное обучение побуждает учащихся развивать свою культурную осведомленность, но студенты, пережившие негативный опыт взаимодействия с другой культурой, могут не проявлять к нему интереса [Syahreal et al., 2019; Rapanta, Trovão, 2021]. Действительно, если новый опыт воспринимается как угроза, он усиливает настороженность и предубеждения по отношению к другим культурам [Chiu, Cheng, 2007] — а значит, затрудняет процесс межкультурного обучения. Таким образом, критически важным для эффективности межкультурного обучения является отношение к этому обучению, открытость новому опыту и знаниям [Jin, Cortazzi, 2016]. Обусловленность успешности межкультурного обучения внутренним состоянием и отношением человека к происходящему выдвигает на первый план вопрос об установках к межкультурному обучению и определяет необходимость рассматривать его в контексте разных источников обучения и разных процессов, формирующих его содержание.

Итак, межкультурное обучение — это комплексный процесс получения новых знаний и навыков, формирования позитивного отношения к другим культурам и межкультурному взаимодействию, в который можно включаться посредством разных источников обучения. Учитывая результаты проведенного анализа, в теоретической модели опросника будем опираться

на представление о межкультурном обучении как о процессе, который реализуется с опорой на четыре источника (формальное и неформальное обучение, обучение через общение и через погружение в другую культуру) посредством пяти процессов (включение в определенный опыт, наблюдение и рефлексия, выстраивание догадок, экспериментирование, преодоление дискомфорта), а установки к нему объединяют когнитивные представления, эмоциональные оценки и намерения вовлекаться в ассоциированное с обучением поведение. При этом именно позитивные установки к обучению составляют основу обучения, делают студентов более восприимчивыми к новой информации и позволяют реализовать те возможности, которые им предлагает вуз. Оценивание установок студентов к межкультурному обучению и формирование позитивных установок критически важно для развития межкультурной компетентности студентов — а значит, необходим инструментарий для их измерения. Существующие методики оценивают либо включенность студента в те или иные активности (источники) межкультурного обучения, либо стили обучения в целом (например, [Kolb, Kolb, 2005]). В данном исследовании проведена апробация нового опросника установок студентов к межкультурному обучению.

2. Метод

2.1. Выборка

После исключения некачественно заполненных анкет в итоговую выборку исследования вошли ответы 399 респондентов — российских студентов в возрасте от 18 до 58 лет (88,5% выборки составили респонденты в возрасте до 30 лет, средний возраст 23,18 года). Выборка сбалансирована по полу: 58,9% респондентов — женщины. Выборка достаточно разнородна по направлениям обучения и по типам вузов и их локализации. В ней представлены студенты, изучающие социальные науки (25,31%), студенты технического (20,8%), естественнонаучного (13,7%), гуманитарного (11,8%), экономического (13,03%), юридического (12,28%) направлений обучения и студенты творческих вузов (3,01%). Студенты бакалавриата и специалитета составляют 84,2% выборки. В выборке представлены студенты более 60 вузов из 53 регионов Российской Федерации, 34,8% респондентов — студенты вузов Москвы и Санкт-Петербурга. Все респонденты имеют российское гражданство, при этом 88,22% респондентов считают себя русскими, 5,76% респондентов не сообщили, к какой этнокультурной группе себя относят, 6,02% респондентов указали этнокультурные группы, отличные от русской: евреи — 2, казахи — 2, белорусы — 3, татары — 4, башкиры — 2, чеченцы — 3, армяне — 6, бурят — 1, якут — 1, карел — 1, азербайджанец —1. Меньше половины респондентов

имеют межкультурный опыт: 48,37% проходили межкультурные курсы в вузе, 37,34% участвовали во внеучебных межкультурных мероприятиях, 37,84% общались с иностранными студентами в вузе, 21,8% были на зарубежных учебных стажировках.

2.2. Дизайн и процедура исследования

Исследование имеет кросс-секционный дизайн и нацелено на апробацию разработанного опросника установок к межкультурному обучению. Опросник был размещен на интернет-платформе *anketolog.ru*. Заполнение опросника начиналось с формы информированного согласия и фильтрующих вопросов относительно очного обучения в вузе в настоящее время и российского гражданства. Далее респонденты отвечали на вопросы относительно их межкультурного опыта в вузе, участия в межкультурных курсах, внеучебных активностях, общения с инокультурными студентами, а также опыта обучения и/или проживания за рубежом. Далее респондентам предлагалось ответить на вопросы относительно установок к межкультурному обучению по четырем источникам обучения. После этого они отвечали на вопросы шкал культурного интеллекта, поддержки мультикультурной идеологии и идеологии ассимиляционизма. В конце респонденты сообщали свои социально-демографические характеристики: пол, возраст, вуз, направление обучения, город. На заполнение опросника требовалось в среднем 27 минут. При условии полного и качественного заполнения опросника респонденты получали небольшое денежное вознаграждение.

2.3. Инструментарий

Установки к межкультурному обучению измерялись при помощи опросника, состоящего из четырех частей, каждая со своей инструкцией (см. Приложение). Изначальный пул вопросов разработан на основании сформированного в результате теоретического анализа представления о том, что межкультурное обучение реализуется в рамках четырех источников обучения, при этом обучение включает пять процессов, а установки к межкультурному обучению представляют собой когнитивный, аффективный и поведенческий аспекты отношения студентов к межкультурному обучению. Далее для отбора наиболее подходящих вопросов использовался метод «Дельфи», доказавший свою применимость и полезность при разработке опросников [Mya, Zaw, Mya, 2021]. Вопросы оценивали 10 экспертов: два преподавателя российских вузов, два филолога, два студента-бакалавра и два студента-магистра разных направлений обучения, а также два российских студента, обучающихся за рубежом. В итоговом варианте опросника остались вопросы, которые

большинство экспертов (60% и более) оценили как хорошо отражающие отношение к межкультурному обучению. Вопросы и инструкции к каждой из частей опросника проверены на предмет корректности, понятности и удобочитаемости с помощью когнитивных интервью в формате онлайн с использованием техники *think-aloud*. В интервью участвовали 8 респондентов, свободно владеющих русским языком: два менеджера по работе со студентами, курирующие культурно гетерогенную группу студентов, три преподавателя-психолога, работающие как с русскими, так и с инокультурными студентами, и четыре студента НИУ ВШЭ разных направлений обучения. Перед запуском сбора данных опросник отредактирован в соответствии с предложениями участников по изменению формулировок. В итоговый вариант опросника вошли 30 вопросов для измерения установок к формальному межкультурному обучению на курсах (включая 12 обратных), 30 вопросов для измерения установок к неформальному межкультурному обучению на внеучебных мероприятиях (включая 10 обратных), 30 вопросов для измерения установок к межкультурному обучению в процессе общения с представителями других культур в вузе (включая 8 обратных) и 31 вопрос для измерения установок к межкультурному обучению путем погружения во время поездок в инокультурные регионы или страны (включая 9 обратных). Показатель по каждой из субшкал высчитывался как среднее ответов на включенные в нее пункты.

Для проверки конструктной валидности Шкалы установок к межкультурному обучению оценивалось соотношение ответов респондентов по этой шкале с ответами на два других опросника, связанных с межкультурной компетентностью: на Шкалу культурного интеллекта [Dyne van, Ang, Koh, 2009], адаптированную на русском языке [Беловол, Шкварило, Хворова, 2012], и на методику Л. Розенталя и С. Леви [Rosenthal, Levy, 2012], также адаптированную на русском языке [Григорьев, Батхина, Дубров, 2018]. Культурный интеллект — это способность человека успешно адаптироваться к новой культурной среде. Шкала культурного интеллекта состоит из четырех субшкал: метакогнитивной (4 пункта), когнитивной (6 пунктов), мотивационной (5 пунктов) и поведенческой (5 пунктов) [Ang et al., 2007]. Респондентов просили оценить степень своего согласия с каждым из утверждений опросника по шкале от 1 (полностью не согласен) до 7 (полностью согласен). Показатель по каждому из компонентов культурного интеллекта высчитывался как сумма ответов на включенные в соответствующую субшкалу пункты. Уровень надежности шкалы в целом $\alpha = 0{,}92$, метакогнитивной субшкалы — $\alpha = 0{,}83$, когнитивной субшкалы — $\alpha = 0{,}78$, мотивационной субшкалы — $\alpha = 0{,}87$, поведенческой субшкалы —

α = 0,82. Предполагалось, что установки к межкультурному обучению положительно связаны с культурным интеллектом как позитивным результатом межкультурного обучения.

Методика Л. Розенталя и С. Леви измеряет поддержку идеологии мультикультурализма. В нашем исследовании использовался адаптированный вариант опросника [Григорьев, Батхина, Дубров, 2018]. Респондентам предлагалось по 7-балльной шкале от 1 (абсолютно не согласен) до 7 (совершенно согласен) оценить степень своего согласия с пятью утверждениями относительно мультикультурализма и пятью — относительно ассимиляционизма. Утверждения касались норм проживания разных этнических групп в мультикультурном обществе. Уровень надежности шкалы поддержки идеологии мультикультурализма составил α = 0,85, уровень надежности шкалы поддержки идеологии ассимиляционизма — α = 0,79. Показатель по каждой шкале высчитывался как среднее ответов на включенные в них пункты. Идеология мультикультурализма состоит в признании и позитивной оценке культурных различий и границ между группами для достижения равенства, разнообразия и равного включения всех групп в жизнь общества [Stevens, Plaut, Sanchez-Burks, 2008], и поддержка данной идеологии может рассматриваться как признак межкультурной компетентности [Ari, Laron, 2013]. Идеологию ассимиляционизма можно в некотором смысле расценивать как противоположность идеологии мультикультурализма: она отрицает ценность культурного разнообразия, делает акцент на культуру доминирующей группы и предполагает отказ представителей этнокультурных аутгрупп от собственных культур в пользу культуры большинства [Rosenthal, Levy, 2010]. Ассимиляционизм положительно связан с этноцентризмом [Григорьев, Батхина, Дубров, 2018], и некоторые исследователи считают его формой дискриминации. Следовательно, поддержка этой идеологии будет, скорее, свидетельствовать о невысокой межкультурной компетентности. Поэтому предполагалось, что положительные установки к межкультурному обучению будут позитивно связаны с поддержкой идеологии мультикультурализма и не будут связаны или окажутся связаны негативно с поддержкой идеологии ассимиляционизма.

2.4. Методы обработки данных

Для проверки факторной структуры инструментария и отбора вопросов, которые должны войти в итоговый вариант шкалы, использованы эксплораторный факторный анализ (*exploratory factor analysis*, EFA) и конфирматорный факторный анализ (*confirmatory factor analysis*, CFA). При проведении EFA определялись критерий Кайзера — Мейера — Олкина (KMO) и критерий сфе-

ричности Бартлетта. Для вычисления количества извлекаемых факторов на каждом новом круге EFA использован параллельный анализ [Horn, 1965], т.е. проведены сравнения значений, полученных на данной выборке, со значениями, полученными из смоделированной методом Монте-Карло матрицы (скрипт О'Коннора для SPSS). EFA реализован в программе SPSS (22-я версия) методом главных компонент (*principal component analysis*) с нормализацией Кайзера и вращением Варимакс.

В рамках CFA построены альтернативные SEM-модели в надстройке к SPSS — *Amos* (22-я версия) — с опорой на результаты EFA и модификационные индексы и оценены показатели соответствия модели: относительный хи-квадрат ($\chi^2/df < 3$), сравнительный индекс соответствия (CFI > 0,9), среднеквадратичная ошибка аппроксимации (RMSEA ≤ 0,08). Кроме того, оценены надежность, конвергентная и дискриминантная валидность инструментария. Надежность шкал и субшкал инструментария оценивались в макросе Э. Хайесса для SPSS с использованием критериев альфа Кронбаха и омега МакДональда (α, ω), которые у согласованных шкал должны быть выше 0,7. Конвергентная валидность проанализирована по показателю извлеченной средней дисперсии (AVE). Обычно считается, что AVE выше 0,50 эмпирически подтверждает конвергентную валидность. Однако согласно недавним исследованиям [Hair et al., 2021], если AVE больше 0,4, а надежность выше 0,6, такие показатели конвергентной валидности считаются приемлемыми [Lam, 2012]. Дискриминантная валидность рассчитана для субшкал внутри инструментария по показателю соотношения гетеропризнаков и монопризнаков (*heterotrait-monotrait ratio*, HTMT). Показатель ниже 0,90 свидетельствует о дискриминантной валидности субшкал.

Далее были рассчитаны дескриптивные статистики и показатели альфа Кронбаха по полной выборке по всем субшкалам, проведены межгрупповые и внутригрупповые сравнения, а также проанализированы коэффициенты корреляции показателей шкал разработанного инструментария с другими показателями, характеризующими успешность межкультурного обучения: шкалой культурного интеллекта и двумя шкалами поддержки межгрупповых идеологий — шкалой мультикультурализма и шкалой ассимиляционизма. Данные расчеты проведены в программе SPSS (22-я версия).

3. Результаты

По результатам параллельного анализа и EFA извлечены пять факторов, которые объяснили 58,11% дисперсии, мера адекватности выборки Кайзера — Мейера — Олкина (0,96) и тест сферичности Бартлетта ($\chi^2 = 46316{,}64$, $df = 7260$, $p < 0{,}001$) свиде-

тельствуют о состоятельности факторного анализа. Однако при изучении факторных нагрузок оказалось, что пункты разделились на факторы не только по содержанию, но и в зависимости от того, как они были сформулированы: были это прямые или обратные вопросы. Так, первый фактор включает пункты относительно погружения в другую культуру и относительно обучения в процессе общения (21,18% дисперсии). Во второй фактор вошли почти все обратные вопросы, за исключением двух вопросов по неформальному обучению (17,15% дисперсии). Третий фактор составили пункты относительно неформального обучения на внеучебных мероприятиях (9,67% дисперсии). Четвертый фактор объединил пункты о межкультурных курсах (6,67% дисперсии). Пятый фактор составили два пункта об обучении в процессе общения с высокой кросс-нагрузкой в первый фактор (3,43% дисперсии).

После исключения 37 обратных вопросов в процессе EFA идентифицированы четыре фактора, объясняющие 57,25% дисперсии, при этом мера адекватности выборки Кайзера — Мейера — Олкина (0,97) и показатель теста сферичности Бартлетта (χ^2 = 30341,36, df = 3486, p < 0,001) свидетельствуют о состоятельности факторного анализа. Первый фактор составили пункты о погружении в другую культуру (20,4% дисперсии). Во второй фактор вошли пункты об обучении в процессе общения и один пункт о формальном обучении, однако он имел низкую факторную нагрузку и в дальнейшем был исключен (13,29% дисперсии). Третий фактор объединил пункты о неформальном обучении на внеучебных мероприятиях (12,69% дисперсии). В четвертый фактор вошли пункты о межкультурных курсах (10,86% дисперсии). Нагрузки элементов оцениваются в пределах от 0,48 до 0,89, все статистически значимые (p < 0,01).

Таким образом, для дальнейшего анализа остались 83 пункта. CFA реализован на основании этих пунктов на полной выборке. В соответствии с теоретической моделью построены факторные модели третьего уровня — с таким расчетом, чтобы каждый пункт вносил вклад в соответствующий компонент установки, каждый компонент установки вносил вклад в определенный источник обучения, а каждый источник обучения вносил вклад в общий показатель установок к межкультурному обучению (табл. 1). Без модификаций показатели соответствия у данной модели оказались недостаточными. В ходе модификации, в частности, учтены корреляции остатков и последовательно удалены пункты с наименьшими факторными нагрузками — это оказались пункты, касающиеся преодоления негативных эмоций при изучении межкультурных курсов, а также пункты, наиболее коррелированные с другими пунктами и факторами.

Таблица 1. **Показатели качества моделей шкалы установок к формальному межкультурному обучению**

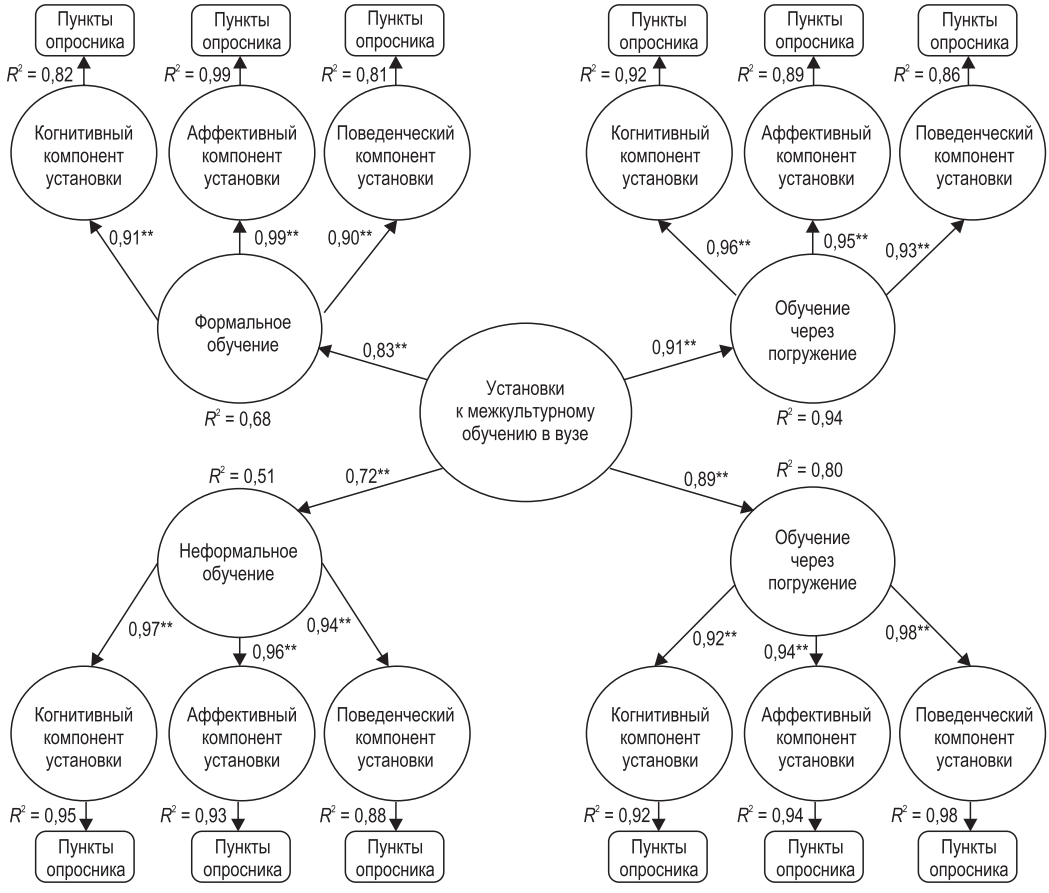
Модель	CMIN /*df*	CFI	RMSEA	AIC	BIC
Модель третьего уровня четырехфакторная (83 элемента)	3,57	0,58	0,08	25975,79	27000,95
Модель третьего уровня четырехфакторная модифицированная (56 элементов)	2,05	0,91	0,05	3272,92	3843,34

Итоговый вариант опросника после CFA состоит из 56 пунктов, сгруппированных в четыре субшкалы, соответствующие четырем источникам обучения. Факторный анализ показал, что каждая субшкала вносит значимый вклад в общий показатель. Надежность единой шкалы составила α = 0,98, ω = 0,98. Конвергентная валидность по единой шкале AVE = 0,73. Дискриминантная валидность рассчитана для четырехфакторной модели: значение по установкам к формальному и установкам к неформальному обучению HTMT = 0,76, по установкам к формальному обучению и обучению через общение HTMT = 0,74, по установкам к формальному обучению и обучению через погружение HTMT = 0,80, по установкам к неформальному обучению и обучению через общение HTMT = 0,64, по установкам к неформальному обучению и обучению через погружение HTMT = 0,69; по установкам к обучению через общение и обучению через погружение HTMT = 0,86. Таким образом, четырехфакторная модель шкалы (CFA третьего уровня) имеет приемлемые показатели надежности, конвергентной и дискриминантной валидности. Рассмотрим далее показатели надежности и валидности каждой из четырех субшкал, представленных тремя компонентами установки — когнитивным, аффективным и поведенческим.

Факторные нагрузки пунктов на субшкалы (компоненты установки к межкультурному обучению в рамках конкретного источника обучения), представленные в табл. 2–5, отражают результаты, полученные по общей факторной модели третьего уровня (рис. 1). Для проверки вклада каждого пункта в общий фактор дополнительно построена факторная модель второго уровня, в которой каждый пункт вносит вклад в соответствующий компонент установки, а каждый компонент установки — в определенный источник обучения, при этом общий фактор шел отдельной латентной переменной, определяемой только пунктами опросника. Соответственно в табл. 2–5 представлены нагрузки из факторной модели третьего уровня (когнитивный компонент, аффективный компонент и поведенческий компонент установки к обучению в рамках определенного источни-

ка обучения), а также нагрузки на установки к межкультурному обучению в целом из факторной модели второго уровня (общий фактор — установки к межкультурному обучению без разделения по источникам обучения).

Рис. 1. **Шкала установок к межкультурному обучению (результаты EFA)**



Примечание: Факторные нагрузки по субшкалам и их состав по пунктам представлены в табл. 2–5.

Факторные нагрузки по субшкале формального обучения и ее компонентам (как часть общей модели) представлены в табл. 2. Оказалось, что в данной субшкале представлены пункты, описывающие только четыре из пяти процессов обучения: пункты, касающиеся преодоления негативных эмоций, имели низкие факторные нагрузки и были удалены из модели. То есть для студентов вузов данный вопрос не актуален, они не испытывают эмоциональных трудностей при получении новых знаний на межкультурных курсах — вероятно, потому, что межкультурные курсы являются инструктивным обучением. Показатели надеж-

Таблица 2. **Факторные нагрузки по пунктам субшкалы установок к формальному межкультурному обучению**

Пункты	Факторы (модель третьего уровня)			Общий фактор (модель второго уровня)
	Когнитивный компонент	Аффективный компонент	Поведенческий компонент	
Это полезно — выдвигать собственные предположения о других культурах на основании знаний, полученных на межкультурном курсе	0,71			0,55
Это важно — применять знания, получаемые на межкультурном курсе, на практике	0,88			0,66
Это важно — проходить курсы о культурах народов мира и межкультурной коммуникации	0,85			0,74
Необходимо обдумывать, что нового я узнаю на межкультурных курсах	0,68			0,61
Я радуюсь, когда благодаря межкультурному курсу начинаю лучше понимать другие культуры		0,87		0,81
Я был(а) бы рад(а) использовать знания и навыки, полученные на межкультурном курсе		0,83		0,79
Я бы порадовался(ась) возможности прослушать межкультурный курс		0,80		0,73
Мне нравится строить догадки о других культурах во время обучения на межкультурном курсе		0,67		0,68
Я часто пытаюсь предположить, как и почему ведут себя люди из разных культур, основываясь на информации с межкультурного курса			0,71	0,66
Мне хотелось бы использовать знания и навыки, полученные на межкультурном курсе			0,82	0,79
Я намерен(а) пройти как можно больше межкультурных курсов во время учебы в университете			0,66	0,63
Я часто думаю о том, как проходит мое обучение на межкультурных курсах			0,45	0,40

ности по субшкале установок к формальному межкультурному обучению составили α = 0,93, ω = 0,93, по когнитивному компоненту установок α = 0,86, ω = 0,86, по аффективному компоненту установок α = 0,87, ω = 0,87, по поведенческому компоненту установок α = 0,80, ω = 0,80. Конвергентная валидность установлена по единой субшкале и трем компонентам: по еди-

ной субшкале AVE = 0,87, по когнитивному компоненту установок AVE = 0,62, по аффективному компоненту установок AVE = = 0,63, по поведенческому компоненту установок AVE = 0,46. Дискриминантная валидность рассчитана для трех компонентов установок внутри данной субшкалы: значение по когнитивному и аффективному компонентам HTMT = 0,73, по когнитивному и поведенческому компонентам HTMT = 0,68, по аффективному и поведенческому компонентам HTMT = 0,55. Таким образом, субшкала установок к формальному межкультурному обучению при условии разделения на три компонента установок имеет приемлемые показатели надежности, конвергентной и дискриминантной валидности.

Факторные нагрузки по субшкале неформального межкультурного обучения и ее компонентам представлены в табл. 3. Содержательно в данной субшкале представлены все пять процессов обучения. Показатели надежности по субшкале установок к неформальному межкультурному обучению α = 0,93, ω = 0,93, по когнитивному компоненту установок α = 0,86, ω = 0,86, по аффективному компоненту установок α = 0,80, ω = 0,81, по поведенческому компоненту установок α = 0,84, ω = 0,84. Конвергентная валидность установлена по единой субшкале и трем компонентам: по единой субшкале AVE = 0,92, по когнитивному компоненту установок AVE = 0,54, по аффективному компоненту установок AVE = 0,56, по поведенческому компоненту установок AVE = 0,52. Дискриминантная валидность рассчитана для трех компонентов установок внутри данной субшкалы: значение по когнитивному и аффективному компонентам HTMT = 0,74, по когнитивному и поведенческому компонентам HTMT = 0,68, по аффективному и поведенческому компонентам HTMT = 0,70. Таким образом, субшкала установок к неформальному межкультурному обучению при условии разделения на три компонента установок имеет приемлемые показатели надежности, конвергентной и дискриминантной валидности.

Факторные нагрузки по субшкале межкультурного обучения через общение и ее компонентам показаны в табл. 4. Содержательно в данной субшкале представлены все пять процессов обучения. Показатели надежности по субшкале установок к межкультурному обучению через общение α = 0,95, ω = 0,95, по когнитивному компоненту установок α = 0,88, ω = 0,88, по аффективному компоненту установок α = 0,90, ω = 0,90, по поведенческому компоненту установок α = 0,86, ω = 0,86. Конвергентная валидность установлена по единой субшкале и трем компонентам: по единой субшкале AVE = 0,95, по когнитивному компоненту установок AVE = 0,57, по аффективному компоненту установок AVE = 0,64, по поведенческому компоненту установок AVE = 0,56. Дискриминантная валидность рассчитана для трех

Таблица 3. **Факторные нагрузки по пунктам субшкалы установок к неформальному межкультурному обучению**

Пункты	Факторы (модель третьего уровня)			Общий фактор (модель второго уровня)
	Когнитивный компонент	Аффективный компонент	Поведенческий компонент	
Это важно — обдумывать, что нового я узнал(а) на внеучебных межкультурных мероприятиях	0,79			0,80
Это правильно — по-своему интерпретировать знания, полученные на внеучебных межкультурных мероприятиях	0,77			0, 32
Это полезно — посещать внеучебные межкультурные мероприятия	0,78			0,53
Это полезно — применять в реальной жизни знания, полученные на внеучебных межкультурных мероприятиях	0,57			0,65
Это правильно — преодолевать сильные эмоции, которые могут возникнуть во время внеучебных межкультурных мероприятий из-за неожиданной информации о разных культурах	0,72			0,57
Это интересно — участвовать во внеучебных межкультурных мероприятиях		0,42		0,52
Мне нравится вспоминать, как я учился(ась) чему-то на внеучебных межкультурных мероприятиях		0,84		0,54
Я рад(а) использовать на практике навыки и знания о разных культурах, полученные на внеучебных мероприятиях		0,81		0,68
Мне страшно, что во время участия во внеучебных мероприятиях я могу узнать какие-то неприятные факты о моей и других культурах		0,84		0,28
Даже если я почувствую себя не в своей тарелке во время внеучебных межкультурных мероприятий, я смогу справиться с дискомфортом и научиться чему-то новому			0,65	0,46
Я планирую посетить максимально возможное количество внеучебных межкультурных мероприятий			0,66	0,58
Обычно я стараюсь обдумывать, чему я научился(ась) на межкультурных внеучебных мероприятиях			0,73	0,55
Я пытаюсь разобраться в особенностях разных культур, посещая внеучебные межкультурные мероприятия			0,76	0,64
Я хочу использовать то, чему научился(ась) на внеучебных межкультурных мероприятиях, в реальной жизни			0,76	0,64

Таблица 4. **Факторные нагрузки по пунктам субшкалы установок к межкультурному обучению через общение**

Пункты	Факторы (модель третьего уровня)			Общий фактор (модель второго уровня)
	Когнитивный компонент	Аффективный компонент	Поведенческий компонент	
Если быть внимательным к представителям других народов, можно лучше понять их культуры, ценности и традиции	0,83			0,65
Это полезно — строить догадки об особенностях разных культур на основании моего общения с представителями других народов	0,62			0,55
После того как я узнаю что-то новое от представителей других культур — например, как правильно приветствуют в их стране, — целесообразно использовать это в дальнейшем общении	0,77			0,60
Это правильно — пытаться научиться чему-то новому в межкультурном общении, даже если я буду испытывать дискомфорт из-за культурных различий	0,73			0,48
Это важно — учиться новому о разных народах в межкультурном общении	0,80			0,63
Я очень доволен(ьна), когда могу применить новые знания о других культурах в межкультурном общении		0,82		0,70
Я всегда радуюсь, когда узнаю что-то новое о разных культурах в межкультурном общении		0,87		0,68
Мне интересно размышлять о том, как проходит мое общение с представителями других народов		0,86		0,64
Мне интересно думать о том, как культурные различия могут сказываться на моем общении с представителями разных народов		0,79		0,65
Я не испытываю или легко перебарываю чувство дискомфорта, когда пытаюсь узнать что-то новое о других культурах через общение		0,63		0,48
Я часто выдвигаю гипотезы о других культурах на основании моего общения с представителями разных народов			0,64	0,48

Пункты	Факторы (модель третьего уровня)			Общий фактор (модель второго уровня)
	Когнитивный компонент	Аффективный компонент	Поведенческий компонент	
Если я почувствую себя неловко при общении с представителями других народов, это не помешает мне узнать что-то новое о других культурах			0,60	0,49
Я стараюсь учиться чему-то, связанному с разными культурами, когда общаюсь с представителями других народов			0,84	0,66
Я стараюсь понять, что именно идет не так, если кто-то из знакомых мне представителей других народов ведет себя странно			0,80	0,61
Я использую знания о других культурах и навыки, полученные при общении с представителями других народов, в реальной жизни			0,84	0,69

компонентов установок внутри данной субшкалы: значение по когнитивному и аффективному компонентам HTMT = 0,72, по когнитивному и поведенческому компонентам HTMT = 0,68, по аффективному и поведенческому компонентам HTMT = 0,71. Таким образом, субшкала установок к межкультурному обучению через общение при условии разделения на три компонента установок имеет приемлемые показатели надежности, конвергентной и дискриминантной валидности.

Факторные нагрузки по субшкале межкультурного обучения через погружение в другие культуры и ее компонентам приведены в табл. 5. Содержательно в данной субшкале представлены все пять процессов обучения. Показатели надежности по субшкале установок к обучению через погружение α = 0,96, ω = 0,96, по когнитивному компоненту установок α = 0,93, ω = = 0,93, по аффективному компоненту установок α = 0,90, ω = 0,90, по поведенческому компоненту установок α = 0,87, ω = 0,88. Конвергентная валидность установлена по единой субшкале и трем компонентам: по единой субшкале AVE = 0,89, по когнитивному компоненту установок AVE = 0,73, по аффективному компоненту установок AVE = 0,64, по поведенческому компоненту установок AVE = 0,59. Дискриминантная валидность рассчитана для трех компонентов установок внутри данной субшкалы: значение по когнитивному и аффективному компонентам HTMT = 0,80, по когнитивному и поведенческому компонентам HTMT = 0,80, по аффективному и поведенческому компонентам HTMT = 0,78. Таким образом, субшкала установок к межкультурному обучению через погружение в другие культу-

ры при условии разделения на три компонента установок имеет приемлемые показатели надежности, конвергентной и дискриминантной валидности.

Таблица 5. **Факторные нагрузки по пунктам субшкалы установок к межкультурному обучению через погружение в другие культуры**

Пункты	Факторы (модель третьего уровня)			Общий фактор (модель второго уровня)
	Когнитивный компонент	Аффективный компонент	Поведенческий компонент	
Это важно — замечать, как я узнаю что-то новое о других культурах на личном опыте во время учебных поездок	0,85			0,62
Это полезно — искать объяснения поступкам местных жителей, когда учишься в другой стране или регионе	0,83			0,59
Нужно использовать на практике навыки и знания о других культурах в межкультурных поездках	0,90			0,67
Это важно — узнавать новое о других культурах во время учебных поездок	0,83			0,65
Это правильно — стараться узнавать новое о разных народах в учебной поездке, даже если чужая культура кажется непонятной	0,88			0,61
Меня радует, когда я догадываюсь о причинах поведения местных людей во время учебной поездки		0,82		0,68
Мне было бы приятно изменять свое поведение, используя знания о других культурах, во время учебной поездки		0,73		0,62
Это интересно — изучать культурные особенности разных стран и регионов через учебные поездки		0,84		0,62
Это интересно — вести дневник и размышлять о новом опыте во время межкультурной поездки		0,82		0,56
Я бы гордился(ась) собой, если бы смог(ла) узнать что-то новое о разных культурах даже во время дискомфортной учебной поездки в другую страну или регион		0,79		0,64
Я бы старался(ась) быть внимательным(ой) к культурным особенностям при обучении в другой стране или регионе			0,86	0,67
Во время учебной поездки я бы постарался(ась) сформировать собственное представление о том, почему в местной культуре такие правила и нормы, традиции и обычаи			0,83	0,68

Пункты	Факторы (модель третьего уровня)			Общий фактор (модель второго уровня)
	Когнитивный компонент	Аффективный компонент	Поведенческий компонент	
Если я почувствую себя неловко при попытке узнать что-то новое о другой культуре во время учебной поездки, я смогу справиться с этим и продолжить с интересом учиться			0,75	0,54
Я хочу узнать больше о разных культурах через учебные поездки			0,69	0,57
При обучении за рубежом или в другом регионе я бы часто пробовал(а) отработать на практике навыки и знания о разных культурах			0,68	0,59

3.1. Описательные статистики и конструктная валидность

Описательные статистики и корреляции между разработанными шкалами, а также корреляции Шкалы установок к межкультурному обучению и Шкалы культурного интеллекта представлены в табл. 6 и 7. Рассмотрим их подробнее, а также проведем внутригрупповые сравнения установок к межкультурному обучению по источникам обучения и компонентам установок.

Таблица 6. **Описательные статистики по Шкале установок к межкультурному обучению**

	Среднее	Стандартное отклонение
Установки к межкультурному обучению	4,7379	1,03658
К формальному межкультурному обучению	4,6464	1,16949
Когнитивный компонент	4,8766	1,29333
Аффективный компонент	4,7105	1,39155
Поведенческий компонент	4,3521	1,22336
К неформальному межкультурному обучению	4,669	1,1191
Когнитивный компонент	4,8371	1,20227
Аффективный компонент	4,5689	1,25073
Поведенческий компонент	4,5810	1,22611
К обучению через общение	4,7576	1,12646
Когнитивный компонент	4,9742	1,19390
Аффективный компонент	4,6029	1,33324
Поведенческий компонент	4,6957	1,13846
К обучению через погружение в другие культуры	4,8558	1,29852
Когнитивный компонент	5,0025	1,40791
Аффективный компонент	4,7218	1,40040
Поведенческий компонент	4,8431	1,29014

Таблица 7. **Корреляции между установками к межкультурному обучению и компонентами культурного интерректа**

	Установки к межкультурному обучению в вузе			
	Формальное обучение	Неформальное обучение	Обучение через общение	Обучение через погружение
Метакогнитивный культурный интеллект	0,27**	0,22**	0,33**	0,31**
Когнитивный культурный интеллект	0,35**	0,31**	0,42**	0,36**
Мотивационный культурный интеллект	0,48**	0,40**	0,58**	0,48**
Поведенческий культурный интеллект	0,24**	0,23**	0,31**	0,26**
Мультикультурализм	0,19**	0,15**	0,18**	0,16**
Ассимиляционизм	0,06	0,03	0,03	0,04

Примечание: ** $p < 0,01$.

Все компоненты установок к межкультурному обучению в вузе оценены респондентами выше среднего (по шкале от 1 до 7), что свидетельствует о позитивном отношении к межкультурному обучению в целом. При этом применение *t*-критерия Стьюдента для связанных выборок позволило выявить значимые различия в установках к межкультурному обучению по источникам обучения — а значит, их иерархию. Так, установки к межкультурному обучению в процессе погружения в другую культуру выражены сильнее, чем установки к обучению в процессе общения ($t = 2,58$, $p < 0,05$), установки к неформальному межкультурному обучению в процессе внеучебных мероприятий ($t = 3,33$, $p < 0,01$) и установки к формальному межкультурному обучению на курсах ($t = 4,25$, $p < 0,01$). На втором месте по позитивности установки к межкультурному обучению во время общения и установки к межкультурному обучению в процессе внеучебных мероприятий: они выражены примерно одинаково ($t = 1,84$, $p > 0,05$) и сильнее, чем установки к формальному межкультурному обучению на курсах ($t = 2,65$, $p < 0,01$).

Внутри всех субшкал наиболее сильно выражен когнитивный компонент установки, при этом приоритетность аффективного или поведенческого компонентов зависит от источника обучения. В субшкале установок к формальному межкультурному обучению когнитивный компонент выражен сильнее, чем аффективный ($t = 3,82$, $p < 0,01$) и поведенческий ($t = 9,27$, $p < 0,01$), а аффективный компонент сильнее, чем поведенческий ($t = 7,24$, $p < 0,01$). В субшкале установок к неформальному межкультурному обучению когнитивный компонент сильнее аффективного ($t = 6,73$, $p < 0,01$) и поведенческого ($t = 5,67$, $p < 0,01$), которые выражены примерно одинаково ($t = –0,27$, $p > 0,05$). В субшкале установок к межкультурному обучению через общение когнитивный компонент также силь-

нее аффективного (t = 8,52, p < 0,01) и поведенческого (t = 6,98, p < 0,01), но здесь аффективный компонент слабее поведенческого (t = 2,24, p < 0,05). Аналогичные результаты получены для субшкалы установок к межкультурному обучению через погружение в другие культуры: когнитивный компонент сильнее аффективного (t = 7,56, p < 0,01) и поведенческого (t = 4,33, p < 0,01), а аффективный компонент слабее поведенческого (t = 3,24, p < 0,01).

Не обнаружено статистически значимых различий в выраженности изучаемых показателей между мужчинами и женщинами, а также между теми, кто имел, и теми, кто не имел опыта обучения на межкультурных учебных курсах. У студентов, имеющих опыт обучения и/или проживания за рубежом, сильнее выражены установки к межкультурному обучению через погружение в другую культуру (t = 3,36, p < 0,01) и в процессе обучения на курсах (t = 2,50, p < 0,05). У студентов со значимым опытом межкультурного общения выше показатели установок к межкультурному обучению в процессе погружения в другую культуру (t = 2,29, p < 0,01). У студентов, не имеющих опыта участия во внеучебных межкультурных мероприятиях, выше, чем в других подгруппах, показатели установок к межкультурному обучению на курсах (t = 2,23, p < 0,05), через общение (t = 2,60, p < 0,05) и при погружении в другую культуру (t = 3,26, p < 0,01).

Установки к межкультурному обучению по четырем видам обучения в вузе оказались позитивно связаны между собой. Более того, каждая из разработанных шкал позитивно коррелирует с субшкалами опросника на культурный интеллект. Самые высокие коэффициенты корреляции установок к межкультурному обучению получены с субшкалой «мотивация», которая отражает, в частности, готовность человека тратить время и силы на изучение культур и функционирование в культурно гетерогенной среде. Также все четыре субшкалы установок к межкультурному обучению позитивно связаны с поддержкой идеологии мультикультурализма и не связаны с поддержкой идеологии ассимиляционизма. Данные паттерны взаимосвязи свидетельствуют о конструктной валидности разработанного инструментария и о том, что он может быть использован в дальнейших исследованиях и на практике.

4. Обсуждение результатов

Межкультурное обучение является основой формирования кросс-культурной компетентности [Barret, 2012], и глубокое понимание того, как именно оно происходит, необходимо современным исследователям и практикам системы высшего образования. Однако в большинстве исследований такое обучение операционализируется через включение в определенный тип

межкультурного опыта (например, проживание в другой стране, прохождение межкультурных курсов или общение с представителями других культур как неформальное обучение) или через компетентность как его результат. В данном исследовании апробирован качественно новый инструментарий для оценки установок студентов к межкультурному обучению в вузе. Создание такого опросника позволяет сместить акценты в оценивании в психологическую плоскость и проанализировать не только средовые детерминанты развития межкультурной компетентности, но и связанные с ним личностные характеристики, в частности установки.

По итогам исследования валидизирован на выборке российских студентов опросник установок к межкультурному обучению. Полная Шкала установок к межкультурному обучению содержит 56 пунктов. Субшкала установок к межкультурному формальному обучению в рамках учебных курсов включает 12 утверждений, по четыре утверждения на каждый из трех компонентов установки. Формальное обучение оказалось единственным источником обучения, для которого не важен процесс преодоления негативных эмоциональных реакций. Шкала установок к неформальному межкультурному обучению на внеучебных мероприятиях включает 14 утверждений: по пять утверждений на когнитивный и поведенческий компоненты и четыре на аффективный. Шкалы установок к межкультурному обучению через общение и установок к межкультурному обучению через погружение в другие культуры содержат по 15 пунктов, по пять утверждений на каждый из трех компонентов установки. Надежность и валидность каждой шкалы подтверждены релевантными статистическими процедурами.

Наиболее сильными у студентов оказались установки к межкультурному обучению через погружение в другую культуру и через общение с представителями других культур (в своем университете). Предыдущие исследования показали, что интенсивные дружеские контакты с представителями культур, отличных от собственной, важны для процесса обучения, так как повышают культурную осведомленность и желание учиться у других [Bazron, Osher, Fleishmann, 2005]. Наиболее выраженным по всем субшкалам оказался когнитивный компонент установок. Вероятно, его приоритетность отражает осознание современными студентами необходимости наработки межкультурной компетентности. С другой стороны, для наиболее интерактивных источников обучения, таких как общение или погружение в другую культуру, сравнительно низок (тем не менее выше среднего по шкале) аффективный компонент установок, что может говорить о наличии у студентов страхов перед межкультурной коммуникацией.

В процессе валидизации инструментария обнаружены позитивные взаимосвязи установок к межкультурному обучению в вузе с уровнем культурного интеллекта студентов. В целом эти данные соответствуют представлениям о позитивных установках как факторе, способствующем включенности в обучение [Kara, 2010], а межкультурное обучение приводит к развитию межкультурной компетентности [Bennett, Salonen, 2007; Fenech, Baguant, Abdelwahed, 2020; Lane, 2012]. При этом самыми сильными оказались взаимосвязи установок к межкультурному обучению с мотивационным культурным интеллектом. Действительно, в соответствии с саморегуляционной моделью обучения [Pintrich, 2003; Boekaerts, Pintrich, Zeidner, 1999] люди по-разному мотивированы в межкультурном обучении, и именно их цели, самоэффективность и внутренняя мотивация к межкультурному обучению определяют успешность процесса обучения [Strohmeier, Gradinger, Wagner, 2017].

Установки к межкультурному обучению в вузе оказались позитивно связаны с поддержкой идеологии мультикультурализма и не связаны с поддержкой идеологии ассимиляционизма. Данный результат, вероятно, объясняется содержанием каждой из идеологий. В то время как мультикультурализм признает и приветствует различия, ассимиляционизм отрицает значимость культурных различий и является антиэгалитарным по своей сути [Дубров, Григорьев, 2019]. При этом люди, позитивно относящиеся к культурным различиям, более восприимчивы к поступающей от представителей других культур новой информации, что способствует формированию у них более глубоких культурно специфических знаний [Matusitz, 2012].

Новизна полученных в данном исследовании результатов заключается в психологизации процесса межкультурного обучения, их методологическая значимость состоит в разработке и валидизации Шкалы установок к межкультурному обучению. Практическое применение результаты исследования могут найти при планировании современными вузами политики интернационализации и выстраивании ими академического климата, который способствовал бы межкультурному обучению. Хотя исследование ориентировано на образовательную среду, закономерности межкультурного обучения универсальны и результаты исследования могут использоваться в любых других ситуациях межкультурного взаимодействия. Поэтому перспективной видится адаптация разработанного инструментария в других контекстах, например в бизнес-среде, и применительно к другим возрастным группам, при этом, безусловно, должны быть учтены особенности обучения детей и взрослых, а также то, что с возрастом людям сложнее преодолевать стресс аккультурации и открываться новому опыту [Donnellan, Lucas, 2008; Cheung, Chudek, Heine, 2011].

Используя результаты данной работы, важно иметь в виду ряд присущих ей ограничений. Во-первых, выборка была недостаточно большой для проведения факторного анализа на полном наборе пунктов шкалы. При этом для последующего использования итогового варианта опросника рекомендуется адаптировать и упростить инструкции к каждой из субшкал анкеты. В целях данного исследования инструкции были предварительно протестированы, но некоторые формулировки могут требовать уточнений или разъяснений для другой выборки и в другом контексте. Во-вторых, в данном исследовании приняли участие студенты, относящиеся к этнокультурному большинству: почти 90% респондентов считают себя русскими. При этом люди по-разному воспринимают и оценивают межкультурный климат в зависимости от того, относятся они к этнокультурному большинству или меньшинству [Harper, Hurtado, 2007], — а значит, и процессы межкультурного обучения у них могут протекать по-разному. Проведение аналогичного исследования на студентах из других этнокультурных групп помогло бы оценить, насколько универсален созданный инструментарий. Кроме того, исследование проводилось в России в период, когда границы по большей части были закрыты, возможности академической мобильности были существенно ограничены и социальная ситуация характеризовалась высоким уровнем неопределенности. Поэтому необходима дополнительная проверка инструментария в других социокультурных контекстах. Процессы индивидуального культурного обучения могут протекать эффективно только тогда, когда студенты чувствуют себя в безопасности настолько, чтобы интересоваться культурными различиями и исследовать их [King, Perez, Shim, 2013]. Исследования показывают, что в разных странах межкультурное обучение происходит по-разному [Hong, Snell, 2008]. Более того, существуют кросс-культурные различия в стилях обучения в целом [Brown et al., 2007]. Проведение кросс-культурного исследования могло бы выявить культурно обусловленные особенности установок к межкультурному обучению. На протекание этого процесса может оказывать влияние формирование чувства принадлежности к учебной организации и общей идентичности, а также формирование соответствующих установок [Maramba, Museus, 2013]. Общая идентичность способствует общению и обучению друг у друга, преодолению возможных предрассудков — и в результате облегчает протекание межкультурного обучения. Поэтому в будущих исследованиях целесообразно принять в рассмотрение социальные идентичности респондентов и степень их идентификации с вузом. Более того, в этнически и культурно гетерогенных обществах, таких как российское, имеет смысл выделять среди инокультур-

ных студентов внутренних и внешних мигрантов. При проведении дальнейших исследований целесообразно проанализировать, как русские студенты воспринимают представителей той или иной культуры, т.е. кого они определяют в качестве инокультурных, а кого не относят к этой группе из-за отсутствия явных идентифицирующих признаков.

5. Выводы

По итогам исследования разработан опросник установок к межкультурному обучению. В нем выделены четыре субшкалы, каждая из которых представлена тремя компонентами установок: когнитивным, аффективным и поведенческим. Теоретически данные субшкалы базируются на трех ключевых основаниях: на выделении нескольких источников обучения, на учете оценочного отношения студентов к межкультурному обучению по трем компонентам установок и на понимании содержания межкультурного обучения в парадигме теорий экспириентального [Kolb, Kolb, 2022] и трансформационного обучения [Mezirow, 2003]. Эти предпосылки соответствуют структурно-оценочной модели мультикультурного опыта, учитывающей три характеристики опыта — глубину, ширину и общую позитивность оценки [Maddux et al., 2021]. Разработанная шкала характеризуется приемлемыми показателями надежности и валидности и может быть использована для проведения дальнейших исследований и оценки установок к межкультурному обучению в практических целях. Однако для ее применения в других контекстах и возрастных группах может потребоваться адаптация инструкций и пунктов шкалы.

Благодарности

Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ.

Приложение

Инструкции к Шкале установок к межкультурному обучению

1. Субшкала установок к формальному межкультурному обучению.

В университете вам могут быть предложены учебные курсы, посвященные культурам разных народов, а также межкультурной коммуникации в разных сферах деятельности. Далее будем называть их межкультурными курсами. В рамках таких межкультурных курсов, как правило, необходимо прослушать ряд лекций и поучаствовать в практических занятиях, предполагающих различные активности. Практические занятия часто включают дискуссии, ролевые игры, подготовку презентаций о разных культурах, элемен-

ты тренинга, решение кейсов, связанных с проблемами межкультурного взаимодействия в реальной жизни.

Утверждения ниже касаются вашего отношения к обучению на таких курсах при условии, что они доступны или были бы доступны для прохождения в рамках вашей образовательной программы. Оцените, пожалуйста, степень согласия с утверждениями ниже по шкале от 1 (абсолютно не согласен) до 7 (полностью согласен).

2. Субшкала установок к неформальному межкультурному обучению.

В университете вы можете поучаствовать не только в учебных курсах, но и в различных дополнительных (необязательных) внеучебных мероприятиях, посвященных культурам разных народов мира, а также особенностям межкультурной коммуникации в разных сферах деятельности. Далее будем называть их внеучебными межкультурными мероприятиями. К таким мероприятиям можно отнести специальные события, посвященные разным культурам, такие как праздники, выставки, просмотр и обсуждение фильмов и спектаклей, культурные воркшопы, мастер-классы и т.п.

Утверждения ниже касаются вашего отношения к участию в подобных мероприятиях при условии, что они доступны или были бы доступны в вашем университете. Оцените, пожалуйста, степень согласия с утверждениями ниже по шкале от 1 (абсолютно не согласен) до 7 (полностью согласен).

3. Субшкала установок к межкультурному обучению через общение.

Люди, с которыми вы общаетесь в рамках обучения в вузе, также могут выступать источником новых знаний. Это могут быть представители любых народов, чья культура отличается от вашей, с которыми вы встречаетесь в университете. Например, преподаватели или студенты из других стран или россияне из разных этнокультурных групп (татары, казахи, чуваши и т.п.), с которыми вы вместе учитесь, выполняете групповые домашние задание или дружески общаетесь.

Утверждения ниже касаются вашего отношения к получению новых знаний о разных культурах и навыков общения с представителями этих культур в процессе межкультурного общения в университете. Оцените, пожалуйста, степень согласия с утверждениями ниже по шкале от 1 (абсолютно не согласен) до 7 (полностью согласен).

4. Субшкала установок к межкультурному обучению через погружение в другие культуры (поездки в инокультурные страны и регионы).

Многие вузы предоставляют своим студентам возможность глубокого погружения в другую культуру, организуя различные

учебные поездки. Это могут быть экскурсии и экспедиции, краткосрочные и долгосрочные стажировки, программы студенческого обмена. Возможны поездки как за рубеж, так и внутри своей страны — в регион с культурой, отличной от вашей (например, Татарстан или Чувашия для этнически русского студента).

Утверждения ниже касаются вашего отношения к участию в подобных учебных поездках в другую страну или регион при условии, что они доступны или были бы доступны для вас в университете. При ответе на вопросы данного блока опирайтесь, пожалуйста, на ваше мнение о полезности, приятности и желательности подобного опыта и не учитывайте геополитические, финансовые и иные ограничения. Оцените, пожалуйста, степень согласия с утверждениями ниже по шкале от 1 (абсолютно не согласен) до 7 (полностью согласен).

Литература

1. Беловол Е.В., Шкварило К.А., Хворова Е.М. (2012) Адаптация опросника «Шкала культурного интеллекта» К. Эрли и С. Анга на русскоязычной выборке. *Вестник Российского университета дружбы народов. (Психология и педагогика*), № 4, сс. 5–14.
2. Бульцева М.А. (2020) *Межкультурные контакты и кросс-культурная компетентность как факторы креативности российских студентов*: дис. ... канд. психол. наук. М.: НИУ ВШЭ. Доступно по ссылке: https://www.hse.ru/data/xf/475/666/1573/1%D0%94%D0%B8%D1%81%D1%81%D0%B5%D1%80%D1%82%D0%B0%D1%86%D0%B8%D1%8F.pdf (дата обращения 12.09.2023).
3. Григорьев Д.С., Батхина А.А., Дубров Д.И. (2018) Ассимиляционизм, мультикультурализм, этнический дальтонизм и поликультурализм в российском контексте. *Культурно-историческая психология,* т. 14*,* № 2, сс. 53–65. https://doi.org/10.17759/chp.2018140206
4. Дубров Д.И., Григорьев Д.С. (2019) Современные исследования межгрупповых идеологий: ассимиляционизм, этнический дальтонизм, мультикультурализм, поликультурализм. *Общественные науки и современность,* № 1, сс. 143–155. https://doi.org/10.31857/S086904990002755-2
5. Мацумото Д. (2003) *Психология и культура*. СПб.: Питер.
6. Abirin S.G. (2023) Students' Learning Attitudes toward Home-Based Education Amidst the COVID-19 Pandemic. *Russian Law Journal*, vol. 11, no 3, pp. 1002–1009. https://doi.org/10.52783/rlj.v11i3.1389
7. Ang S., van Dyne L., Koh C., Ng K.Y., Templer K.J., Tay C., Chandrasekar N.A. (2007) Cultural Intelligence Scale (CQS). *APA PsycTests*. https://doi.org/10.1037/t24375-000
8. Ari L.L., Laron D. (2013) Intercultural Learning in Graduate Studies at an Israeli College of Education: Attitudes toward Multiculturalism among Jewish and Arab Students. *Higher Education*, vol. 68, no 2, pp. 243–262. https://doi.org/10.1007/s10734-013-9706-9
9. Barrett M. (2012) Intercultural Competence. *EWC Statement Series.* Oslo: The European Wergeland Centre, pp. 23–27. Available at: https://theewc.org/content/uploads/2020/02/EWC-Statement-Series-2012.pdf (accessed 11 September 2023).
10. Bazron B., Osher D., Fleischman S. (2005) Creating Culturally Responsive Schools. *American Educator*, vol. 11, no 1, pp. 83–84.

11. Bennett J.M., Salonen R. (2007) Intercultural Communication and the New American Campus. *Change: The Magazine of Higher Learning*, vol. 39, no 2, pp. 46–50. https://doi.org/10.3200/chng.39.2.46-c4
12. Binkley E.E., Minor A.J. (2020) Constructivist Pedagogy to Promote Cultural Learning in Counselor Education. *Journal of Creativity in Mental Health*, vol. 16, no 3, pp. 348–359. https://doi.org/10.1080/15401383.2020.1763222
13. Boekaerts M., Pintrich P.R., Zeidner M. (1999) Self-Regulation: An Introductory Overview. *Handbook of Self-Regulation* (eds M. Boekaerts, M. Zeidner, P.R. Pintrich), San Diego, CA: Academic Press, pp. 1–9. https://doi.org/10.1016/b978-012109890-2/50030-5
14. Brown M.B., Aoshima M., Bolen L.M., Chia R., Kohyama T. (2007) Cross-Cultural Learning Approaches in Students from the USA, Japan and Taiwan. *School Psychology International*, vol. 28, no 5, pp. 592–604. https://doi.org/10.1177/0143034307085660
15. Buschenhofen P. (1998) English Language Attitudes of Final-Year High School and First-Year University Students in Papua New Guinea. *Asian Journal of English Language Teaching*, vol. 8, pp. 93–116. Available at: http://www.cuhk.edu.hk/ajelt/vol8/rep2.htm (accessed 11 September 2023).
16. Chan E.A., Lai T., Wong A., Ho S., Chan B., Stenberg M., Carlson E. (2017) Nursing Students' Intercultural Learning via Internationalization at Home: A Qualitative Descriptive Study. *Nurse Education Today*, vol. 52, May, pp. 34–39. https://doi.org/10.1016/j.nedt.2017.02.003
17. Chen J.R., Ju T.L. (2005) Continuing Professional Development: A Cross-Cultural Learning Approach. International Journal of Innovation and Learning, vol. 2, no 3, pp. 283–302. https://doi.org/10.1504/IJIL.2005.006371
18. Cheung B.Y., Chudek M., Heine S.J. (2010) Evidence for a Sensitive Period for Acculturation. *Psychological Science*, vol. 22, no 2, pp. 147–152. https://doi.org/10.1177/0956797610394661
19. Chinnappan M., McKenzie B., Fitzsimmons P. (2013) Pre-Service Teachers' Attitudes towards Overseas Professional Experience: Implications for Professional Practice. *Australian Journal of Teacher Education*, vol. 38, no 12, pp. 36–54. https://doi.org/10.14221/ajte.2013v38n12.5
20. Chiu C., Cheng S.Y. (2007) Toward a Social Psychology of Culture and Globalization: Some Social Cognitive Consequences of Activating Two Cultures Simultaneously. *Social and Personality Psychology Compass*, vol. 1, no 1, pp. 84–100. https://doi.org/10.1111/j.1751-9004.2007.00017.x
21. Chwialkowska A. (2020) Maximizing Cross-Cultural Learning from Exchange Study Abroad Programs: Transformative Learning Theory. *Journal of Studies in International Education*, vol. 24, no 5, pp. 535–554. https://doi.org/10.1177/1028315320906163
22. Constantin E.C., Cohen-Vida M.I., Popescu A.V. (2015) Developing Cultural Awareness. *Procedia — Social and Behavioral Sciences,* vol. 191, June, pp. 696–699. https://doi.org/10.1016/j.sbspro.2015.04.228
23. Cranton P. (2023) Transformative Learning Theory as an Integrated Perspective. *Understanding and Promoting Transformative Learning,* New York, NY: Routledge, pp. 30–45. https://doi.org/10.4324/9781003448433-3
24. Crisp R.J., Turner R.N. (2011) Cognitive Adaptation to the Experience of Social and Cultural Diversity. *Psychological Bulletin*, vol. 137, no 2, pp. 242–266. https://doi.org/10.1037/a0021840
25. Czaika M., de Haas H. (2014) The Globalization of Migration: Has the World Become More Migratory? *International Migration Review*, vol. 48, no 2, pp. 283–323. https://doi.org/10.1111/imre.12095
26. Dai K., Garcia J. (2019) Intercultural Learning in Transnational Articulation Programs. *Journal of International Students*, vol. 9, no 2, pp. 362–383. https://doi.org/10.32674/jis.v9i2.677

27. DeRobertis E.M., Bland A.M. (2020) From Personal Threat to Cross-Cultural Learning: An Eidetic Investigation. *Journal of Phenomenological Psychology,* vol. 51, no 1, pp. 1–15. https://doi.org/10.1163/15691624-12341368
28. Donnellan M.B., Lucas R.E. (2008) Age Differences in the Big Five across the Life Span: Evidence from Two National Samples. *Psychology and Aging*, vol. 23, no 3, pp. 558–566. https://doi.org/10.1037/a0012897
29. Dyne van L., Ang S., Koh C. (2009) Cultural Intelligence: Measurement and Scale Development. *Contemporary Leadership and Intercultural Competence: Exploring the Cross-Cultural Dynamics within Organizations* (ed. M.A. Moodian), Los Angeles: Sage, pp. 233–254. https://doi.org/10.4135/9781452274942.n18
30. Fenech R., Baguant P., Abdelwahed I. (2020) Cultural Learning in the Adjustment Process of Academic Expatriates. *Cogent Education*, vol. 7, no 1, Article no 1830924. https://doi.org/10.1080/2331186x.2020.1830924
31. Gawronski B. (2007) Attitudes Can Be Measured! But What Is an Attitude? *Social Cognition,* vol. 25, no 5, pp. 573–581. https://doi.org/10.1521/soco.2007.25.5.573
32. Gondra A., Czerwionka L. (2018) Intercultural Knowledge Development during Short-Term Study Abroad in the Basque Country: A Cultural and Linguistic Minority Context. *Frontiers: The Interdisciplinary Journal of Study Abroad*, vol. 30, no 3, pp. 119–146. https://doi.org/10.36366/frontiers.v30i3.427
33. Gregersen-Hermans J. (2017) Intercultural Competence Development in Higher Education. *Intercultural Competence in Higher Education: International Approaches, Assessment and Application* (eds D.K. Deardorff, L.A. Arasaratnam-Smith), London: Routledge, pp. 67–82. https://doi.org/10.4324/9781315529257-7
34. Griffiths K., Kopanidis F., Steel M. (2018) Investigating the Value of a Peer-to-Peer Mentoring Experience. *Australasian Marketing Journal*, vol. 26, no 2, pp. 92–98. https://doi.org/10.1016/j.ausmj.2018.05.006
35. Gudykunst W.B., Ting-Toomey S. (1988) Culture and Affective Communication. *American Behavioral Scientist*, vol. 31, no 3, pp. 384–400. https://doi.org/10.1177/000276488031003009
36. Haddock G., Maio G.R. (2008) Attitudes: Content, Structure and Functions. *Introduction to Social Psychology: A European Perspective* (eds M. Hewstone, W. Stroebe, K. Jonas), Malden, MA; Oxford: BPS Blackwell, pp. 112–133. Available at: https://www.blackwellpublishing.com/content/hewstonesocialpsychology/chapters/cpt6.pdf (accessed 11 September 2023).
37. Hair Jr. J.F., Hult G.T.M., Ringle C.M., Sarstedt M. (2021) *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM).* Los Angeles: Sage.
38. Harper S.R., Hurtado S. (2007) Nine Themes in Campus Racial Climates and Implications for Institutional Transformation. *New Directions for Student Services*, no 120, pp. 7–24. https://doi.org/10.1002/ss.254
39. Hong J.F., Snell R.S. (2008) Power Inequality in Cross-Cultural Learning: The Case of Japanese Transplants in China. *Asia Pacific Business Review*, vol. 14, no 2, pp. 253–273. https://doi.org/10.1080/13602380701314750
40. Horn J.L. (1965) A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika,* vol. 30, no 2, pp. 179–185. https://doi.org/10.1007/BF02289447
41. Ippolito K. (2007) Promoting Intercultural Learning in a Multicultural University: Ideals and Realities. *Teaching in Higher Education*, vol. 12, no 5–6, pp. 749–763. https://doi.org/10.1080/13562510701596356
42. Jarosiński M., Kozma M., Sekliuckiene J. (2021) Comparison of Explicit and Implicit Methods of Cross-Cultural Learning in an International Classroom. *Sustainability*, vol. 13, no 19, Article no 10641. https://doi.org/10.3390/su131910641
43. Jin L., Cortazzi M. (2016) Practising Cultures of Learning in Internationalising Universities. *Journal of Multilingual and Multicultural Development*, vol. 38, no 3, pp. 237–250. https://doi.org/10.1080/01434632.2015.1134548

44. Kara A. (2010) The Development of the Scale of Attitudes towards Learning. *Elektronic Journal of Social Sciences*, vol. 9, no 32, pp. 49–62.
45. King P.M., Perez R.J., Shim W. (2013) How College Students Experience Intercultural Learning: Key Features and Approaches. *Journal of Diversity in Higher Education*, vol. 6, no 2, pp. 69–83. https://doi.org/10.1037/a0033243
46. Klak T., Martin P. (2003) Do University-Sponsored International Cultural Events Help Students to Appreciate "Difference"? *International Journal of Intercultural Relations,* vol. 27, no 4, pp. 445–465. https://doi.org/10.1016/s0147-1767(03)00033-6
47. Kolb A.Y., Kolb D.A. (2022) Experiential Learning Theory as a Guide for Experiential Educators in Higher Education. *Experiential Learning and Teaching in Higher Education*, vol. 1, no 1, pp. 7–44. https://doi.org/10.46787/elthe.v1i1.3362
48. Kolb A.Y., Kolb D.A. (2005) Learning Styles and Learning Spaces: Enhancing Experiential Learning in Higher Education. *Academy of Management Learning & Education*, vol. 4, no 2, pp. 193–212. https://doi.org/10.5465/amle.2005.17268566
49. Kulich S., Wang Y. (2015) Intercultural Communication in China. *The SAGE Encyclopedia of Intercultural Communication* (ed. J.M. Bennett), Thousand Oaks, CA: Sage, pp. 458–469. https://doi.org/10.4135/9781483346267
50. Kurt M., Benzer S. (2020) An Investigation on the Effect of STEM Practices on Sixth Grade Students' Academic Achievement, Problem Solving Skills, and Attitudes towards STEM. *Journal of Science Learning*, vol. 3, no 2, pp. 79–88. https://doi.org/10.17509/jsl.v3i2.21419
51. Lam L.W. (2012) Impact of Competitiveness on Salespeople's Commitment and Performance. *Journal of Business Research*, vol. 65, no 9, pp. 1328–1334. https://doi.org/10.1016/j.jbusres.2011.10.026
52. Lane H.C. (2012) Intercultural Learning. *Encyclopedia of the Sciences of Learning* (ed. N.M. Seel), New York, NY: Springer, pp. 1618–1620. https://doi.org/10.1007/978-1-4419-1428-6_242
53. Leask B. (2009) Using Formal and Informal Curricula to Improve Interactions between Home and International Students. *Journal of Studies in International Education*, vol. 13, no 2, pp. 205–221. https://doi.org/10.1177/1028315308329786
54. Lee A., Williams R.D., Shaw M.A., Jie Y. (2014) First-Year Students' Perspectives on Intercultural Learning. *Teaching in Higher Education,* vol. 19, no 5, pp. 543–554. https://doi.org/10.1080/13562517.2014.880687
55. Leung A.K., Chiu C. (2010) Multicultural Experience, Idea Receptiveness, and Creativity. *Journal of Cross-Cultural Psychology*, vol. 41, no 5–6, pp. 723–741. https://doi.org/10.1177/0022022110361707
56. Lin X., Shen G.Q. (2019) How Formal and Informal Intercultural Contacts in Universities Influence Students' Cultural Intelligence? *Asia Pacific Education Review*, vol. 21, no 2, pp. 245–259. https://doi.org/10.1007/s12564-019-09615-y
57. Lu J.G., Hafenbrack A.C., Eastwick P.W., Wang D.J., Maddux W.W., Galinsky A.D. (2017) "Going Out" of the Box: Close Intercultural Friendships and Romantic Relationships Spark Creativity, Workplace Innovation, and Entrepreneurship. *Journal of Applied Psychology*, vol. 102, no 7, pp. 1091–1108. https://doi.org/10.1037/apl0000212
58. Maddux W.W., Lu J.G., Affinito S.J., Galinsky A.D. (2021) Multicultural Experiences: A Systematic Review and New Theoretical Framework. *Academy of Management Annals*, vol. 15, no 2, pp. 345–376. https://doi.org/10.5465/annals.2019.0138
59. Mahoney S.L., Schamber J.F. (2004) Exploring the Application of a Developmental Model of Intercultural Sensitivity to a General Education Curriculum on Diversity. *The Journal of General Education*, vol. 53, no 3, pp. 311–334. https://doi.org/10.1353/jge.2005.0007

60. Mašić A., Bećirović S. (2021) Attitudes towards Learning English as a Foreign Language. *JoLIE*, no 14, pp. 85–105. https://doi.org/10.29302/jolie.2021.2.5
61. Maramba D.C., Museus S.D. (2013) Examining the Effects of Campus Climate, Ethnic Group Cohesion, and Cross-Cultural Interaction on Filipino American Students' Sense of Belonging in College. *Journal of College Student Retention: Research, Theory & Practice*, vol. 14, no 4, pp. 495–522. https://doi.org/10.2190/cs.14.4.d
62. Matusitz J. (2012) Relationship between Knowledge, Stereotyping, and Prejudice in Interethnic Communication. *PASOS Revista de Turismo y Patrimonio Cultural*, vol. 10, no 1, pp. 89–98. https://doi.org/10.25145/j.pasos.2012.10.008
63. McKay J., O'Neill D., Petrakieva L. (2016) CAKES (Cultural Awareness and Knowledge Exchange Scheme): A Holistic and Inclusive Approach to Supporting International Students. *Journal of Further and Higher Education*, vol. 42, no 2, pp. 276–288. https://doi.org/10.1080/0309877x.2016.1261092
64. Mezirow J. (2003) Transformative Learning as Discourse. *Journal of Transformative Education*, vol. 1, no 1, pp. 58–63. https://doi.org/10.1177/1541344603252172
65. Mitchell L., Paras A. (2018) When Difference Creates Dissonance: Understanding the 'Engine' of Intercultural Learning in Study Abroad. *Intercultural Education*, vol. 29, no 3, pp. 321–339. https://doi.org/10.1080/14675986.2018.1436361
66. Mya K.S., Zaw K.K., Mya K.M. (2021) Developing and Validating a Questionnaire to Assess an Individual's Perceived Risk of Four Major Non-Communicable Diseases in Myanmar. *PLOS One*, vol. 16, no 4, Article no e0234281. https://doi.org/10.1371/journal.pone.0234281
67. O'Brien B., Tuohy D., Fahy A., Markey K. (2019) Home Students' Experiences of Intercultural Learning: A Qualitative Descriptive Design. *Nurse Education Today*, vol. 74, December, pp. 25–30. https://doi.org/10.1016/j.nedt.2018.12.005
68. Pence H.M., Macgillivray I.K. (2008) The Impact of an International Field Experience on Preservice Teachers. *Teaching and Teacher Education,* vol. 24, no 1, pp. 14–25. https://doi.org/10.1016/j.tate.2007.01.003
69. Pintrich P.R. (2003) A Motivational Science Perspective on the Role of Student Motivation in Learning and Teaching Contexts. *Journal of Educational Psychology*, vol. 95, no 4, pp. 667–686. https://doi.org/10.1037/0022-0663.95.4.667
70. Rahman A.R., Jalaluddin I., Mohd Kasim Z., Darmi R. (2021) Attitudes towards Learning English among the Aliya Madrasah Students in Bangladesh. *Indonesian Journal of Applied Linguistics*, vol. 11, no 2, pp. 269–280. https://doi.org/10.17509/ijal.v11i2.34121
71. Ramirez S. (2021) Cultural Exposure as a Creative Experiential Learning Intervention. *Journal of Creativity in Mental Health,* vol. 18, no 1, pp. 118–133. https://doi.org/10.1080/15401383.2021.1949420
72. Rapanta C., Trovão S. (2021) Intercultural Education for the Twenty-First Century: A Comparative Review of Research. *Dialogue for Intercultural Understanding. Placing Cultural Literacy at the Heart of Learning* (eds F. Maine, M. Vrikki), Cham: Springer Nature, pp. 9–26. https://doi.org/10.1007/978-3-030-71778-0_2
73. Rosenthal L., Levy S.R. (2012) The Relation between Polyculturalism and Intergroup Attitudes among Racially and Ethnically Diverse Adults. *Cultural Diversity and Ethnic Minority Psychology*, vol. 18, no 1, pp. 1–16. https://doi.org/10.1037/a0026490
74. Rosenthal L., Levy S.R. (2010) The Colorblind, Multicultural, and Polycultural Ideological Approaches to Improving Intergroup Attitudes and Relations. *Social Issues and Policy Review*, vol. 4, no 1, pp. 215–246. https://doi.org/10.1111/j.1751-2409.2010.01022.x
75. Santoro N., Major J. (2012) Learning to Be a Culturally Responsive Teacher through International Study Trips: Transformation or Tourism? *Teaching Education*, vol. 23, no 3, pp. 309–322. https://doi.org/10.1080/10476210.2012.685068

76. Şen H. (2013) The Attitudes of University Students towards Learning. *Procedia — Social and Behavioral Sciences*, vol. 83, July, pp. 947–953. https://doi.org/10.1016/j.sbspro.2013.06.177
77. Sizoo S., Serrie H. (2004) Developing Cross-Cultural Skills of International Business Students: An Experiment. *Journal of Instructional Psychology*, vol. 31, no 2, pp. 160–166.
78. Shao Y., Crook C. (2015) The Potential of a Mobile Group Blog to Support Cultural Learning among Overseas Students. *Journal of Studies in International Education*, vol. 19, no 5, pp. 399–422. https://doi.org/10.1177/1028315315574101
79. Spitzberg B.H., Changnon G. (2009) Conceptualizing Intercultural Competence. *The SAGE Handbook of Intercultural Competence* (ed D.K. Deardorff), Thousand Oaks, CA: Sage, pp. 2–52. https://doi.org/10.1177/0021886308314460
80. Stern H.H. (1983) *Fundamental Concepts of Language Teaching: Historical and Interdisciplinary Perspectives on Applied Linguistic Research.* Oxford: Oxford University.
81. Stevens F.G., Plaut V.C., Sanchez-Burks J. (2008) Unlocking the Benefits of Diversity: All-Inclusive Multiculturalism and Positive Organizational Change. *The Journal of Applied Behavioral Science,* vol. 44, no 1, pp. 116–133. http://dx.doi.org/10.1177/0021886308314460
82. Strohmeier D., Gradinger P., Wagner P. (2017) Intercultural Competence Development among University Students from a Self-Regulated Learning Perspective. *Zeitschrift für Psychologie*, vol. 225, no 1, pp. 85–94. https://doi.org/10.1027/2151-2604/a000282
83. Syahrial S., Asrial A., Kurniawan D.A., Nugroho P., Septiasari R., Pratama R.A., Perdana R. (2019) Increased Behavior of Students' Attitudes to Cultural Values Using the Inquiry Learning Model Assisted by Ethnoconstructivism. *Journal of Educational Science and Technology (EST)*, vol. 5, no 2, pp. 166–175. https://doi.org/10.26858/est.v5i2.9670
84. Tahaneh Y., Hana D. (2013) Jordanian Undergraduates Motivations and Attitudes towards Learning English in EFL Context. *International Review of Social Sciences and Humanities,* vol. 4, no 2, pp. 159–180.
85. Tomasello M., Kruger A.C., Ratner H.H. (1993) Cultural Learning. *Behavioral and Brain Sciences,* vol. 16, no 3, pp. 495–511. https://doi.org/10.1017/s0140525x0003123x
86. Tran T.Q., Seepho S. (2022) Intercultural Language Education: Supportive Factors and Constraints on EFL Learners' Intercultural Communicative Competence Development. *Suranaree Journal of Social Science,* vol. 11, no 1, pp. 1–28. https://doi.org/10.55766/vgwq2509
87. Uyar A. (2023) Exploring the Students' Attitudes towards e-Learning at Territory Level: A Focus on Türkiye. *International Journal of Curriculum and Instruction,* vol. 15, no 2, pp. 1327–1353.
88. Ward C., Okura Y., Kennedy A., Kojima T. (1998) The U-Curve on Trial: A Longitudinal Study of Psychological and Sociocultural Adjustment during Cross-Cultural Transition. *International Journal of Intercultural Relations,* vol. 22, no 3, pp. 277–291. https://doi.org/10.1016/s0147-1767(98)00008-x
89. Wilson J., Ward C., Fischer R. (2013) Beyond Culture Learning Theory. *Journal of Cross-Cultural Psychology,* vol. 44, no 6, pp. 900–927. https://doi.org/10.1177/0022022113492889
90. Xu X., Chen X. (2017) Unlocking Expatriates' Job Creativity: The Role of Cultural Learning, and Metacognitive and Motivational Cultural Intelligence. *Management and Organization Review*, vol. 13, no 4, pp. 767–794. https://doi.org/10.1017/mor.2017.50
91. Yamazaki Y., Kayes D.C. (2004) An Experiential Approach to Cross-Cultural Learning: A Review and Integration of Competencies for Successful Expa-

triate Adaptation. *Academy of Management Learning & Education,* vol. 3, no 4, pp. 362–379. https://doi.org/10.5465/amle.2004.15112543

92. Zhu Y., Bargiela-Chiappini F. (2013) Balancing Emic and Etic: Situated Learning and Ethnography of Communication in Cross-Cultural Management Education. *Academy of Management Learning & Education,* vol. 12, no 3, pp. 380–395. https://doi.org/10.5465/amle.2012.0221
93. Zhu Y., Okimoto T.G., Roan A., Xu H. (2017) Developing Management Student Cultural Fluency for the Real World. *Education + Training,* vol. 59, no 4, pp. 353–373. https://doi.org/10.1108/et-03-2016-0059

References

Abirin S.G. (2023) Students' Learning Attitudes toward Home-Based Education Amidst the COVID-19 Pandemic. *Russian Law Journal*, vol. 11, no 3, pp. 1002–1009. https://doi.org/10.52783/rlj.v11i3.1389

Ang S., van Dyne L., Koh C., Ng K.Y., Templer K.J., Tay C., Chandrasekar N.A. (2007) Cultural Intelligence Scale (CQS). *APA PsycTests*. https://doi.org/10.1037/t24375-000

Ari L.L., Laron D. (2013) Intercultural Learning in Graduate Studies at an Israeli College of Education: Attitudes toward Multiculturalism among Jewish and Arab Students. *Higher Education*, vol. 68, no 2, pp. 243–262. https://doi.org/10.1007/s10734-013-9706-9

Barrett M. (2012) Intercultural Competence. *EWC Statement Series.* Oslo: The European Wergeland Centre, pp. 23–27. Available at: https://theewc.org/content/uploads/2020/02/EWC-Statement-Series-2012.pdf (accessed 11 September 2023).

Bazron B., Osher D., Fleischman S. (2005) Creating Culturally Responsive Schools. *American Educator*, vol. 11, no 1, pp. 83–84.

Belovol E.V., Shkvarilo K.A., Khvorova E.M. (2012) Adaptatsiya oprosnika "Shkala kul´turnogo intellekta" K. Earley i S. Anga na russkoyazychnoy vyborke [Russian-Language Verification of P.C. Earley and S. Ang's "Cultural Intelligence Scale"]. *RUDN Journal of Psychology and Pedagogics*, no 4, pp. 5–14.

Bennett J.M., Salonen R. (2007) Intercultural Communication and the New American Campus. *Change: The Magazine of Higher Learning*, vol. 39, no 2, pp. 46–50. https://doi.org/10.3200/chng.39.2.46-c4

Binkley E.E., Minor A.J. (2020) Constructivist Pedagogy to Promote Cultural Learning in Counselor Education. *Journal of Creativity in Mental Health*, vol. 16, no 3, pp. 348–359. https://doi.org/10.1080/15401383.2020.1763222

Boekaerts M., Pintrich P.R., Zeidner M. (1999) Self-Regulation: An Introductory Overview. *Handbook of Self-Regulation* (eds M. Boekaerts, M. Zeidner, P.R. Pintrich), San Diego, CA: Academic Press, pp. 1–9. https://doi.org/10.1016/b978-012109890-2/50030-5

Brown M.B., Aoshima M., Bolen L.M., Chia R., Kohyama T. (2007) Cross-Cultural Learning Approaches in Students from the USA, Japan and Taiwan. *School Psychology International*, vol. 28, no 5, pp. 592–604. https://doi.org/10.1177/0143034307085660

Bultseva M.A. (2020) *Mezhkul´turnye kontakty i cross-kul´turnaya kompetentnost´ kak factory kreativnosti rossijskikh studentov* [Intercultural Contacts and Cross-Cultural Competence as Factors of Creativity among Russian Students] (PhD Thesis). Moscow: HSE. Available at: https://www.hse.ru/data/xf/475/666/1573/1%D0%94%D0%B8%D1%81%D1%81%D0%B5%D1%80%D1%82%D0%B0%D1%86%D0%B8%D1%8F.pdf (accessed 20 September 2023).

Buschenhofen P. (1998) English Language Attitudes of Final-Year High School and First-Year University Students in Papua New Guinea. *Asian Journal of English Language Teaching*, vol. 8, pp. 93–116. Available at: http://www.cuhk.edu.hk/ajelt/vol8/rep2.htm (accessed 11 September 2023).

Chan E.A., Lai T., Wong A., Ho S., Chan B., Stenberg M., Carlson E. (2017) Nursing Students' Intercultural Learning via Internationalization at Home: A Qualitative Descriptive Study. *Nurse Education Today*, vol. 52, May, pp. 34–39. https://doi.org/10.1016/j.nedt.2017.02.003

Chen J.R., Ju T.L. (2005) Continuing Professional Development: A Cross-Cultural Learning Approach. *International Journal of Innovation and Learning*, vol. 2, no 3, pp. 283–302. https://doi.org/10.1504/IJIL.2005.006371

Cheung B.Y., Chudek M., Heine S.J. (2010) Evidence for a Sensitive Period for Acculturation. *Psychological Science*, vol. 22, no 2, pp. 147–152. https://doi.org/10.1177/0956797610394661

Chinnappan M., McKenzie B., Fitzsimmons P. (2013) Pre-Service Teachers' Attitudes towards Overseas Professional Experience: Implications for Professional Practice. *Australian Journal of Teacher Education*, vol. 38, no 12, pp. 36–54. https://doi.org/10.14221/ajte.2013v38n12.5

Chiu C., Cheng S.Y. (2007) Toward a Social Psychology of Culture and Globalization: Some Social Cognitive Consequences of Activating Two Cultures Simultaneously. *Social and Personality Psychology Compass*, vol. 1, no 1, pp. 84–100. https://doi.org/10.1111/j.1751-9004.2007.00017.x

Chwialkowska A. (2020) Maximizing Cross-Cultural Learning from Exchange Study Abroad Programs: Transformative Learning Theory. *Journal of Studies in International Education*, vol. 24, no 5, pp. 535–554. https://doi.org/10.1177/1028315320906163

Constantin E.C., Cohen-Vida M.I., Popescu A.V. (2015) Developing Cultural Awareness. *Procedia — Social and Behavioral Sciences,* vol. 191, June, pp. 696–699. https://doi.org/10.1016/j.sbspro.2015.04.228

Cranton P. (2023) Transformative Learning Theory as an Integrated Perspective. *Understanding and Promoting Transformative Learning,* New York, NY: Routledge, pp. 30–45. https://doi.org/10.4324/9781003448433-3

Crisp R.J., Turner R.N. (2011) Cognitive Adaptation to the Experience of Social and Cultural Diversity. *Psychological Bulletin*, vol. 137, no 2, pp. 242–266. https://doi.org/10.1037/a0021840

Czaika M., de Haas H. (2014) The Globalization of Migration: Has the World Become More Migratory? *International Migration Review*, vol. 48, no 2, pp. 283–323. https://doi.org/10.1111/imre.12095

Dai K., Garcia J. (2019) Intercultural Learning in Transnational Articulation Programs. *Journal of International Students*, vol. 9, no 2, pp. 362–383. https://doi.org/10.32674/jis.v9i2.677

DeRobertis E.M., Bland A.M. (2020) From Personal Threat to Cross-Cultural Learning: An Eidetic Investigation. *Journal of Phenomenological Psychology,* vol. 51, no 1, pp. 1–15*. https://doi.org/10.1163/15691624-12341368*

Donnellan M.B., Lucas R.E. (2008) Age Differences in the Big Five across the Life Span: Evidence from Two National Samples. *Psychology and Aging*, vol. 23, no 3, pp. 558–566. https://doi.org/10.1037/a0012897

Dubrov D., Grigoryev D. (2019) Sovremennye issledovaniya mezhgruppovykh ideologiy: assimilyatsionizm, etnicheskiy dal´tonizm, mul´tikul´turalizm, polikul´turalizm [Current Studies on Intergroup Ideologies: Assimilationism, Colorblindness, Multiculturalism, Polyculturalism]. *Social Sciences and Contemporary World*, no 1, pp. 143–155. https://doi.org/10.31857/S086904990002755-2

Dyne van L., Ang S., Koh C. (2009) Cultural Intelligence: Measurement and Scale Development. *Contemporary Leadership and Intercultural Competence: Exploring the Cross-Cultural Dynamics within Organizations* (ed. M.A. Moodian), Los Angeles: Sage, pp. 233–254. https://doi.org/10.4135/9781452274942.n18

Fenech R., Baguant P., Abdelwahed I. (2020) Cultural Learning in the Adjustment Process of Academic Expatriates. *Cogent Education*, vol. 7, no 1, Article no 1830924. https://doi.org/10.1080/2331186x.2020.1830924

Gawronski B. (2007) Attitudes Can Be Measured! But What Is an Attitude? *Social Cognition,* vol. 25, no 5, pp. 573–581. https://doi.org/10.1521/soco.2007.25.5.573

Gondra A., Czerwionka L. (2018) Intercultural Knowledge Development during Short-Term Study Abroad in the Basque Country: A Cultural and Linguistic Minority Context. *Frontiers: The Interdisciplinary Journal of Study Abroad*, vol. 30, no 3, pp. 119–146. https://doi.org/10.36366/frontiers.v30i3.427

Gregersen-Hermans J. (2017) Intercultural Competence Development in Higher Education. *Intercultural Competence in Higher Education: International Approaches, Assessment and Application* (eds D.K. Deardorff, L.A. Arasaratnam-Smith), London: Routledge, pp. 67–82. https://doi.org/10.4324/9781315529257-7

Griffiths K., Kopanidis F., Steel M. (2018) Investigating the Value of a Peer-to-Peer Mentoring Experience. *Australasian Marketing Journal*, vol. 26, no 2, pp. 92–98. https://doi.org/10.1016/j.ausmj.2018.05.006

Grigoryev D.S., Batkhina A.A., Dubrov D.I. (2018) Assimilyatsionizm, mul´tikul´turalizm, etnicheskiy dal´tonizm i polikul´turalizm v rossijskom kontekste [Assimilationism, Multiculturalism, Colorblindness, and Polyculturalism in the Russian Context]. *Kul'turno-istoricheskaya psikhologiya / Cultural-Historical Psychology*, vol. 14, no 2, pp. 53–65. https://doi.org/10.17759/chp.2018140206

Gudykunst W.B., Ting-Toomey S. (1988) Culture and Affective Communication. *American Behavioral Scientist*, vol. 31, no 3, pp. 384–400. https://doi.org/10.1177/000276488031003009

Haddock G., Maio G.R. (2008) Attitudes: Content, Structure and Functions. *Introduction to Social Psychology: A European Perspective* (eds M. Hewstone, W. Stroebe, K. Jonas), Malden, MA; Oxford: BPS Blackwell, pp. 112–133. Available at: https://www.blackwellpublishing.com/content/hewstonesocialpsychology/chapters/cpt6.pdf (accessed 11 September 2023).

Hair Jr. J.F., Hult G.T.M., Ringle C.M., Sarstedt M. (2021) *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM).* Los Angeles: Sage.

Harper S.R., Hurtado S. (2007) Nine Themes in Campus Racial Climates and Implications for Institutional Transformation. *New Directions for Student Services*, no 120, pp. 7–24. https://doi.org/10.1002/ss.254

Hong J.F., Snell R.S. (2008) Power Inequality in Cross-Cultural Learning: The Case of Japanese Transplants in China. *Asia Pacific Business Review*, vol. 14, no 2, pp. 253–273. https://doi.org/10.1080/13602380701314750

Horn J.L. (1965) A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika,* vol. 30, no 2, pp. 179–185. https://doi.org/10.1007/BF02289447

Ippolito K. (2007) Promoting Intercultural Learning in a Multicultural University: Ideals and Realities. *Teaching in Higher Education*, vol. 12, no 5–6, pp. 749–763. https://doi.org/10.1080/13562510701596356

Jarosiński M., Kozma M., Sekliuckiene J. (2021) Comparison of Explicit and Implicit Methods of Cross-Cultural Learning in an International Classroom. *Sustainability*, vol. 13, no 19, Article no 10641. https://doi.org/10.3390/su131910641

Jin L., Cortazzi M. (2016) Practising Cultures of Learning in Internationalising Universities. *Journal of Multilingual and Multicultural Development*, vol. 38, no 3, pp. 237–250. https://doi.org/10.1080/01434632.2015.1134548

Kara A. (2010) The Development of the Scale of Attitudes towards Learning. *Elektronic Journal of Social Sciences*, vol. 9, no 32, pp. 49–62.

King P.M., Perez R.J., Shim W. (2013) How College Students Experience Intercultural Learning: Key Features and Approaches. *Journal of Diversity in Higher Education*, vol. 6, no 2, pp. 69–83. https://doi.org/10.1037/a0033243

Klak T., Martin P. (2003) Do University-Sponsored International Cultural Events Help Students to Appreciate "Difference"? *International Journal of Intercultural Relations,* vol. 27, no 4, pp. 445–465. https://doi.org/10.1016/s0147-1767(03)00033-6

Kolb A.Y., Kolb D.A. (2005) Learning Styles and Learning Spaces: Enhancing Experiential Learning in Higher Education. *Academy of Management Learning & Education*, vol. 4, no 2, pp. 193–212. https://doi.org/10.5465/amle.2005.17268566

Kolb A.Y., Kolb D.A. (2022) Experiential Learning Theory as a Guide for Experiential Educators in Higher Education. *Experiential Learning and Teaching in Higher Education*, vol. 1, no 1, pp. 7–44. https://doi.org/10.46787/elthe.v1i1.3362

Kulich S., Wang Y. (2015) Intercultural Communication in China. *The SAGE Encyclopedia of Intercultural Communication* (ed. J.M. Bennett), Thousand Oaks, CA: Sage, pp. 458–469. https://doi.org/10.4135/9781483346267

Kurt M., Benzer S. (2020) An Investigation on the Effect of STEM Practices on Sixth Grade Students' Academic Achievement, Problem Solving Skills, and Attitudes towards STEM. *Journal of Science Learning*, vol. 3, no 2, pp. 79–88. https://doi.org/10.17509/jsl.v3i2.21419

Lam L.W. (2012) Impact of Competitiveness on Salespeople's Commitment and Performance. *Journal of Business Research*, vol. 65, no 9, pp. 1328–1334. https://doi.org/10.1016/j.jbusres.2011.10.026

Lane H.C. (2012) Intercultural Learning. *Encyclopedia of the Sciences of Learning* (ed. N.M. Seel), New York, NY: Springer, pp. 1618–1620. https://doi.org/10.1007/978-1-4419-1428-6_242

Leask B. (2009) Using Formal and Informal Curricula to Improve Interactions between Home and International Students. *Journal of Studies in International Education*, vol. 13, no 2, pp. 205–221. https://doi.org/10.1177/1028315308329786

Lee A., Williams R.D., Shaw M.A., Jie Y. (2014) First-Year Students' Perspectives on Intercultural Learning. *Teaching in Higher Education,* vol. 19, no 5, pp. 543–554. https://doi.org/10.1080/13562517.2014.880687

Leung A.K., Chiu C. (2010) Multicultural Experience, Idea Receptiveness, and Creativity. *Journal of Cross-Cultural Psychology*, vol. 41, no 5–6, pp. 723–741. https://doi.org/10.1177/0022022110361707

Lin X., Shen G.Q. (2019) How Formal and Informal Intercultural Contacts in Universities Influence Students' Cultural Intelligence? *Asia Pacific Education Review*, vol. 21, no 2, pp. 245–259. https://doi.org/10.1007/s12564-019-09615-y

Lu J.G., Hafenbrack A.C., Eastwick P.W., Wang D.J., Maddux W.W., Galinsky A.D. (2017) "Going Out" of the Box: Close Intercultural Friendships and Romantic Relationships Spark Creativity, Workplace Innovation, and Entrepreneurship. *Journal of Applied Psychology*, vol. 102, no 7, pp. 1091–1108. https://doi.org/10.1037/apl0000212

Maddux W.W., Lu J.G., Affinito S.J., Galinsky A.D. (2021) Multicultural Experiences: A Systematic Review and New Theoretical Framework. *Academy of Management Annals*, vol. 15, no 2, pp. 345–376. https://doi.org/10.5465/annals.2019.0138

Mahoney S.L., Schamber J.F. (2004) Exploring the Application of a Developmental Model of Intercultural Sensitivity to a General Education Curriculum on Diversity. *The Journal of General Education*, vol. 53, no 3, pp. 311–334. https://doi.org/10.1353/jge.2005.0007

Maramba D.C., Museus S.D. (2013) Examining the Effects of Campus Climate, Ethnic Group Cohesion, and Cross-Cultural Interaction on Filipino American Students' Sense of Belonging in College. *Journal of College Student Retention: Research, Theory & Practice*, vol. 14, no 4, pp. 495–522. https://doi.org/10.2190/cs.14.4.d

Mašić A., Bećirović S. (2021) Attitudes towards Learning English as a Foreign Language. *JoLIE*, no 14, pp. 85–105. https://doi.org/10.29302/jolie.2021.2.5

Matsumoto D. (2003) *Psikhologiya i kul´tura* [Psychology and Culture]. Saint Petersburg: Piter.

Matusitz J. (2012) Relationship between Knowledge, Stereotyping, and Prejudice in Interethnic Communication. *PASOS Revista de Turismo y Patrimonio Cultural*, vol. 10, no 1, pp. 89–98. https://doi.org/10.25145/j.pasos.2012.10.008

McKay J., O'Neill D., Petrakieva L. (2016) CAKES (Cultural Awareness and Knowledge Exchange Scheme): A Holistic and Inclusive Approach to Supporting International Students. *Journal of Further and Higher Education*, vol. 42, no 2, pp. 276–288. https://doi.org/10.1080/0309877x.2016.1261092

Mezirow J. (2003) Transformative Learning as Discourse. *Journal of Transformative Education*, vol. 1, no 1, pp. 58–63. https://doi.org/10.1177/1541344603252172

Mitchell L., Paras A. (2018) When Difference Creates Dissonance: Understanding the 'Engine' of Intercultural Learning in Study Abroad. *Intercultural Education*, vol. 29, no 3, pp. 321–339. https://doi.org/10.1080/14675986.2018.1436361

Mya K.S., Zaw K.K., Mya K.M. (2021) Developing and Validating a Questionnaire to Assess an Individual's Perceived Risk of Four Major Non-Communicable Diseases in Myanmar. *PLOS One*, vol. 16, no 4, Article no e0234281. https://doi.org/10.1371/journal.pone.0234281

O'Brien B., Tuohy D., Fahy A., Markey K. (2019) Home Students' Experiences of Intercultural Learning: A Qualitative Descriptive Design. *Nurse Education Today*, vol. 74, December, pp. 25–30. https://doi.org/10.1016/j.nedt.2018.12.005

Pence H.M., Macgillivray I.K. (2008) The Impact of an International Field Experience on Preservice Teachers. *Teaching and Teacher Education,* vol. 24, no 1, pp. 14–25. https://doi.org/10.1016/j.tate.2007.01.003

Pintrich P.R. (2003) A Motivational Science Perspective on the Role of Student Motivation in Learning and Teaching Contexts. *Journal of Educational Psychology*, vol. 95, no 4, pp. 667–686. https://doi.org/10.1037/0022-0663.95.4.667

Rahman A.R., Jalaluddin I., Mohd Kasim Z., Darmi R. (2021) Attitudes towards Learning English among the Aliya Madrasah Students in Bangladesh. *Indonesian Journal of Applied Linguistics*, vol. 11, no 2, pp. 269–280. https://doi.org/10.17509/ijal.v11i2.34121

Ramirez S. (2021) Cultural Exposure as a Creative Experiential Learning Intervention. *Journal of Creativity in Mental Health,* vol. 18, no 1, pp. 118–133. https://doi.org/10.1080/15401383.2021.1949420

Rapanta C., Trovão S. (2021) Intercultural Education for the Twenty-First Century: A Comparative Review of Research. *Dialogue for Intercultural Understanding. Placing Cultural Literacy at the Heart of Learning* (eds F. Maine, M. Vrikki), Cham: Springer Nature, pp. 9–26. https://doi.org/10.1007/978-3-030-71778-0_2

Rosenthal L., Levy S.R. (2012) The Relation between Polyculturalism and Intergroup Attitudes among Racially and Ethnically Diverse Adults. *Cultural Diversity and Ethnic Minority Psychology*, vol. 18, no 1, pp. 1–16. https://doi.org/10.1037/a0026490

Rosenthal L., Levy S.R. (2010) The Colorblind, Multicultural, and Polycultural Ideological Approaches to Improving Intergroup Attitudes and Relations. *Social Issues and Policy Review*, vol. 4, no 1, pp. 215–246. https://doi.org/10.1111/j.1751-2409.2010.01022.x

Santoro N., Major J. (2012) Learning to Be a Culturally Responsive Teacher through International Study Trips: Transformation or Tourism? *Teaching Education*, vol. 23, no 3, pp. 309–322. https://doi.org/10.1080/10476210.2012.685068

Şen H. (2013) The Attitudes of University Students towards Learning. *Procedia — Social and Behavioral Sciences*, vol. 83, July, pp. 947–953. https://doi.org/10.1016/j.sbspro.2013.06.177

Shao Y., Crook C. (2015) The Potential of a Mobile Group Blog to Support Cultural Learning among Overseas Students. *Journal of Studies in International Education*, vol. 19, no 5, pp. 399–422. https://doi.org/10.1177/1028315315574101

Sizoo S., Serrie H. (2004) Developing Cross-Cultural Skills of International Business Students: An Experiment. *Journal of Instructional Psychology*, vol. 31, no 2, pp. 160–166.

Spitzberg B.H., Changnon G. (2009) Conceptualizing Intercultural Competence. *The SAGE Handbook of Intercultural Competence* (ed D.K. Deardorff), Thousand Oaks, CA: Sage, pp. 2–52. https://doi.org/10.1177/0021886308314460

Stern H.H. (1983) *Fundamental Concepts of Language Teaching: Historical and Interdisciplinary Perspectives on Applied Linguistic Research.* Oxford: Oxford University.

Stevens F.G., Plaut V.C., Sanchez-Burks J. (2008) Unlocking the Benefits of Diversity: All-Inclusive Multiculturalism and Positive Organizational Change. *The Journal of Applied Behavioral Science,* vol. 44, no 1, pp. 116–133. http://dx.doi.org/10.1177/0021886308314460

Strohmeier D., Gradinger P., Wagner P. (2017) Intercultural Competence Development among University Students from a Self-Regulated Learning Perspective. *Zeitschrift für Psychologie*, vol. 225, no 1, pp. 85–94. https://doi.org/10.1027/2151-2604/a000282

Syahrial S., Asrial A., Kurniawan D.A., Nugroho P., Septiasari R., Pratama R.A., Perdana R. (2019) Increased Behavior of Students' Attitudes to Cultural Values Using the Inquiry Learning Model Assisted by Ethnoconstructivism. *Journal of Educational Science and Technology (EST)*, vol. 5, no 2, pp. 166–175. https://doi.org/10.26858/est.v5i2.9670

Tahaneh Y., Hana D. (2013) Jordanian Undergraduates Motivations and Attitudes towards Learning English in EFL Context. *International Review of Social Sciences and Humanities,* vol. 4, no 2, pp. 159–180.

Tomasello M., Kruger A.C., Ratner H.H. (1993) Cultural Learning. *Behavioral and Brain Sciences,* vol. 16, no 3, pp. 495–511. https://doi.org/10.1017/s0140525x0003123x

Tran T.Q., Seepho S. (2022) Intercultural Language Education: Supportive Factors and Constraints on EFL Learners' Intercultural Communicative Competence Development. *Suranaree Journal of Social Science,* vol. 11, no 1, pp. 1–28. https://doi.org/10.55766/vgwq2509

Uyar A. (2023) Exploring the Students' Attitudes towards e-Learning at Territory Level: A Focus on Türkiye. *International Journal of Curriculum and Instruction,* vol. 15, no 2, pp. 1327–1353.

Ward C., Okura Y., Kennedy A., Kojima T. (1998) The U-Curve on Trial: A Longitudinal Study of Psychological and Sociocultural Adjustment during Cross-Cultural Transition. *International Journal of Intercultural Relations,* vol. 22, no 3, pp. 277–291. https://doi.org/10.1016/s0147-1767(98)00008-x

Wilson J., Ward C., Fischer R. (2013) Beyond Culture Learning Theory. *Journal of Cross-Cultural Psychology,* vol. 44, no 6, pp. 900–927. https://doi.org/10.1177/0022022113492889

Xu X., Chen X. (2017) Unlocking Expatriates' Job Creativity: The Role of Cultural Learning, and Metacognitive and Motivational Cultural Intelligence. *Management and Organization Review*, vol. 13, no 4, pp. 767–794. https://doi.org/10.1017/mor.2017.50

Yamazaki Y., Kayes D.C. (2004) An Experiential Approach to Cross-Cultural Learning: A Review and Integration of Competencies for Successful Expatriate Adaptation. *Academy of Management Learning & Education,* vol. 3, no 4, pp. 362–379. https://doi.org/10.5465/amle.2004.15112543

Zhu Y., Bargiela-Chiappini F. (2013) Balancing Emic and Etic: Situated Learning and Ethnography of Communication in Cross-Cultural Management Education. *Academy of Management Learning & Education,* vol. 12, no 3, pp. 380–395. https://doi.org/10.5465/amle.2012.0221

Zhu Y., Okimoto T.G., Roan A., Xu H. (2017) Developing Management Student Cultural Fluency for the Real World. *Education + Training,* vol. 59, no 4, pp. 353–373. https://doi.org/10.1108/et-03-2016-0059

Роль контекста в заданиях сценарного типа при измерении универсальных навыков: применение теории генерализации

Дарья Грачева

Статья поступила в редакцию в марте 2023 г.

Грачева Дарья Александровна — младший научный сотрудник Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики». Адрес: 101000, Москва, ул. Мясницкая, 20. E-mail: dgracheva@hse.ru. ORCID: https://orcid.org/0000-0002-4646-7349

Аннотация

В современных условиях большое внимание уделяется развитию и оцениванию универсальных навыков у школьников. Для такого оценивания необходимы новые тестовые форматы, основанные на наблюдаемых действиях учащегося в цифровой среде. Один из перспективных вариантов таких инструментов — контекстные задания сценарного типа. Однако контекстное разнообразие таких заданий может затруднять сравнение результатов. В статье анализируется роль контекста сценарных заданий при измерении двух универсальных навыков: критического мышления и коммуникации. С этой целью применяются методы теории генерализации, которые позволяют установить, в какой степени согласованными являются результаты, полученные с помощью разных контекстов сценарных заданий, и как путем изменения количества индикаторов или контекстов сценариев обеспечить достаточную надежность измерения. Исследование основано на данных, которые получены при тестировании учащихся 4-х классов с помощью разных заданий сценарного типа, входящих в состав инструмента «4К». Результаты анализа показали, что поведение тестируемых в сценариях с разным контекстом различается, при этом трудности контекстов практически одинаковы. Для достижения удовлетворительной надежности рекомендуется использовать минимум два сценария с разными контекстами, а использование трех и более сценарных заданий с разными контекстами позволяет существенно сократить количество индикаторов без потери надежности. В исследовании также оценивалась роль контекста при использовании альтернативных вариантов заданий. Альтернативные варианты схожи в основной проблеме и сюжете сценария, но различаются тематическим наполнением (контентом). Изменение только контента сценария позволяет экстраполировать результаты оценивания универсальных навыков на все варианты заданий, т.е. альтернативные варианты могут использоваться как взаимозаменяемые. Проведенное исследование демонстрирует возможности использования методов теории генерализации для оптимизации разработки заданий с учетом требований к надежности измерения.

Ключевые слова теория генерализации, универсальные навыки, задания сценарного типа, контекст задания, психометрика, надежность измерений

Для цитирования Грачева Д.А. (2023) Роль контекста в заданиях сценарного типа для измерения универсальных навыков: применение теории генерализации. *Вопросы образования / Educational Studies Moscow*, № 3, сс. 62–91. https://doi.org/10.17323/vo-2023-16901

The Role of Context in Scenario-Based Tasks for Measuring Universal Skills: The Use of Generalizability Theory

Daria Gracheva

Daria A. Gracheva — Junior Research Fellow at the Center for Psychometrics and Measurements in Education, Institute of Education, National Research University Higher School of Economics. Address: 20 Myasnitskaya St, 101000 Moscow, Russian Federation. E-mail: dgracheva@hse.ru. ORCID: https://orcid.org/0000-0002-4646-7349

Abstract In education, much attention is paid to the development and evaluation of universal skills in schoolchildren. At the same time, the assessment of universal skills requires new test formats based on the observed actions of the student in the digital environment. Scenario-based contextual tasks serve as a promising format. However, the contextual diversity of such tasks can make it difficult to compare results obtained from different scenario tasks. This article aims to analyze the role of scenario task context in measuring two universal skills: critical thinking and communication. The work uses the methods of Generalizability Theory, which allows to analyze to what extent the results can be generalized for other contexts of scenario tasks, and how, by changing the number of indicators or scenario contexts, to ensure satisfied measurement reliability. The study is based on data from fourth-grade students who were tested with various scenario-based tasks of the "4K" instrument. The results of the analysis showed that the behavior of the test-takers differs in scenarios with different contexts, while the difficulties of the contexts are almost the same. To achieve satisfactory reliability, it is recommended to use at least two scenarios with different contexts, and the use of three or more scenarios with different contexts will reduce the number of indicators without loss of reliability. Also, the study evaluated the role of context when using alternative scenario-based tasks forms were used. The alternative forms were similar in the main problem and plot of the scenario, but differed in topic (content). Changing only the content of the scenario makes it possible to generalize the results across scenario forms, that is, alternative forms can be used interchangeably. This study demonstrates how Generalization Theory can be used to optimize the development of tasks, taking into account the requirements for measurement reliability.

Keywords generalizability theory, universal skills, scenario-based tasks, task context, psychometrics, reliability of measurement

For citing Gracheva D.A. (2023) Rol′ konteksta v zadaniyakh stsenarnogo tipa pri izmerenii universal′nykh navykov: primenenie teorii generalizatsii [The Role of Context in Scenario-Based Tasks for Measuring Universal Skills: The Use of Generalizability Theory]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 3, pp. 62–91. https://doi.org/10.17323/vo-2023-16901

Современное образование нацелено на развитие и оценку универсальных навыков у школьников. Согласно ФГОС, умения применять средства коммуникации и критически анализировать информацию относятся к метапредметным навыкам, которые школьники осваивают в процессе обучения. Каждый универсальный навык представляет собой не однородную структуру, а несколько связанных между собой составляющих, поэтому их называют сложными конструктами.

Оценивание таких сложных конструктов предполагает выход за пределы традиционных типов заданий, таких как задания с выбором варианта ответа, самоотчетные опросники со шкалами Ликерта. Наиболее подходящим форматом для оценки универсальных навыков являются задания в формате *performance-based*, где оценивание происходит через индикаторы наблюдаемого поведения в тестовой среде, в том числе в цифровой [Davier von, Mislevy, Hao, 2022].

К категории *performance-based* относят задания сценарного типа, в них отдельные индикаторы объединены общим контекстом (*scenario-based tasks*), поэтому такие задания принято называть контекстными [Ruiz-Primo, Li, 2015]. Контекст определяет основную проблемную ситуацию сценарного задания, ее контент (тематическое наполнение) и сюжет (последовательность этапов задания и действий персонажей).

Разработав удачный контекст задания, можно повысить мотивацию тестируемого и вовлечь его в прохождение теста. При этом контекстные задания позволяют увидеть, как респонденты применяют навыки в различных ситуациях, в том числе приближенных к реальной жизни, что особенно важно для измерения универсальных навыков. Например, при измерении коммуникации можно создать ситуацию продолжительного диалога, которая позволит оценить всю динамику коммуникативного процесса. В то же время различия в контекстах затрудняют сравнение результатов, полученных с применением разных сценарных заданий, в том числе альтернативных вариантов сценариев, которые часто используются при повторных тестированиях респондентов [Davier von, Mislevy, Hao, 2022]. Иными словами, в различия в тестовом балле вносит вклад не только оцениваемая характеристика, но и ситуация, и тема сценария.

В данной статье мы анализируем роль контекста сценарных заданий при оценке двух универсальных компетенций: критического мышления и коммуникации. Эмпирической основой исследования послужили данные по инструменту «4К», который состоит из заданий сценарного типа, реализованных в цифровой среде с автоматическим скорингом (без привлечения экспертов) для измерения универсальных навыков у млад-

ших школьников[1]. В работе используются методы теории генерализации (*generalizability theory*) [Cronbach et al., 1972], которые позволяют количественно оценить вклад контекста заданий в результаты. В рамках анализа генерализации возможно проверить предположения о том, в какой степени полученные результаты могут быть экстраполированы на другие контексты сценарных заданий и как можно изменить процедуру тестирования для повышения надежности измерения универсальных навыков. Результаты такого анализа можно применять на практике для проектирования контекстных инструментов оценивания с опорой на эмпирические доказательства их психометрического качества, а также для проверки справедливого и сопоставимого оценивания навыков в разных контекстах.

В статье поставлены следующие исследовательские вопросы.

1. Каков вклад контекста сценарного задания в результаты оценивания универсальных компетенций?

2. Какое количество контекстов сценарных заданий необходимо для надежного измерения универсальных навыков?

Статья построена следующим образом: сначала рассмотрены основные положения теории генерализации, затем представлен обзор исследований, в которых применяются методы теории генерализации. Далее описано эмпирическое исследование с применением методов теории генерализации для оценки надежности результатов измерения универсальных навыков с использованием заданий сценарного типа и вклада контекста сценария в тестовый балл.

1. Основные положения теории генерализации

Основы теории генерализации впервые были изложены в работах Л. Кронбаха и его коллег как расширение представлений о концепции надежности в рамках классической теории тестирования [Cronbach et al., 1972]. Позже идеи теории генерализации подробно раскрывались в работах Р. Шавелсона и Р. Бреннона [Shavelson, Webb, Rowley, 1992; Brennan, 1992]. Согласно определению, теория генерализации — это статистическая теория надежности (*dependability*) инструментов измерения [Shavelson, Webb, Rowley, 1992]. Надежность здесь понимается как точность экстраполяции результатов выборочной процедуры измерений на всю генеральную совокупность измерений.

[1] Инструмент «4К» для оценки универсальных навыков (критическое мышление, креативность, коммуникация и кооперация) разработан сотрудниками Центра психометрики и измерений в образовании Института образования НИУ ВШЭ в рамках договора о научно-исследовательской работе с благотворительным фондом «Вклад в будущее». Сайт инструмента доступен по ссылке: https://ioe.hse.ru/4k_test/

В теории генерализации процедуру измерения раскладывают на компоненты, каждый из которых может быть источником наблюдаемых различий в баллах. Например, различия в баллах могут объясняться способностью респондентов (объектом измерения), трудностью заданий в тесте, временем тестирования, степенью строгости экспертов. Для заданий сценарного типа таким компонентом также может быть контекст сценария. Любой компонент процедуры измерения, отличный от объекта измерения, принято называть фасетом. Каждый фасет является источником различий в баллах, которые относятся к ошибке измерения.

В соответствии с целью статьи предположим, что есть несколько сценарных заданий с разными контекстами, при этом все респонденты проходят все сценарии. Такая процедура тестирования включает следующие компоненты: респонденты (объект исследования, *p*), индикаторы сценарного задания (фасет *i*) и контекст сценария (фасет *c*). В рамках теории генерализации предполагается, что элементы фасета конкретной процедуры исследования выбраны случайно из универсума, или полного множества объектов генерализации (*universe of admissible observations*). Аналогично объекты исследования являются случайной выборкой из популяции.

Пусть любой респондент из популяции выполняет любой индикатор сценария из полного множества возможных индикаторов в любом контексте из полного множества возможных контекстов. Тогда балл респондента *p* по индикатору *i* контекста *c* можно представить в виде:

$$X_{pio} = U + v_p + v_i + v_c + v_{pi} + v_{pc} + v_{ic} + v_{pic,e}, \quad (1)$$

где U — генеральное среднее в популяции; v — независимые эффекты (компоненты), а именно: v_p — эффект респондента, v_i — эффект индикатора, v_c — эффект контекста, v_{pi} — эффект взаимодействия респондента и индикатора, v_{pc} — эффект взаимодействия респондента и контекста, v_{ic} — эффект взаимодействия индикатора и контекста, $v_{pic,e}$ — остаточный компонент, включающий эффект взаимодействия всех фасетов и компонент ошибки, т.е. случайной изменчивости и систематической изменчивости, которая не объясняется фасетами конкретной процедуры измерения.

Тогда дисперсия баллов из выражения (1) по всем респондентам из популяции и элементам фасета из полного множества объектов генерализации имеет вид:

$$\sigma^2(X_{pic}) = \sigma^2(p) + \sigma^2(i) + \sigma^2(c) + \sigma^2(pi) + \\ + \sigma^2(pc) + \sigma^2(ic) + \sigma^2(pic,e). \quad (2)$$

Таким образом, дисперсия баллов может быть разложена на независимые компоненты дисперсии, связанные с различием в истинной способности респондентов $\sigma^2(p)$ и с разными источниками ошибки измерения — различиями в трудности индикаторов $\sigma^2(i)$, контекстов $\sigma^2(c)$, эффектами взаимодействия и остаточной дисперсией $\sigma^2(pic,e)$. Это компоненты дисперсии со случайным эффектом. В теории генерализации возможно оценить компоненты дисперсии с фиксированным эффектом — в этом случае элементы конкретного фасета составляют ограниченный набор и не экстраполируются на полное множество объектов генерализации. Дизайн с фиксированными эффектами не рассматривается в данной статье.

Получить количественную оценку каждого компонента дисперсии можно в рамках дисперсионного анализа (ANOVA). Расчет независимых компонентов дисперсии по фасетам исследования становится результатом первого этапа анализа в рамках теории генерализации — G-исследования (*generalizability study*).

В рамках G-исследования определяются фасеты, которые могут оказывать влияние на результаты оценивания, а также отношения между фасетами (или дизайн G-исследования). Выражение (1) иллюстрирует перекрестный дизайн с двумя фасетами (индикаторы и контексты), который принято обозначать как *p* x *i* x *c*. При таком дизайне каждый элемент одного фасета встречается в комбинации с каждым элементом другого, т.е. каждый респондент выполняет каждый индикатор сценария в каждом контексте.

Вложенные дизайны позволяют моделировать вложенные отношения между фасетами. Фасет называется вложенным в другой фасет, если разные элементы первого фасета встречаются в комбинации с каждым элементом второго фасета. Например, дизайн *p* x (*i* : *c*) предполагает, что каждый респондент выполняет каждый индикатор, вложенный в контекст сценария (*i*, размещенное внутри *c*).

Использование перекрестного G-дизайна предпочтительно, потому что вложенный G-дизайн не позволяет оценить все комбинации фасетов между собой (например, в дизайне *p* x (*i* : *c*) не оценивается эффект взаимодействия контекстов и индикаторов). Однако перекрестные G-дизайны не всегда выполнимы на практике.

Таким образом, цель G-исследования — определить, какие фасеты важны для измерения. Чтобы количественно определить вклад фасетов, результаты G-исследования представляют в виде процентов от общей дисперсии баллов.

На основе результатов G-исследования реализуется второй этап анализа в рамках теории генерализации — D-исследование (*decision study*). Цель D-исследования — определить процедуру тестирования, которая повысит надежность за счет изменения

количества элементов фасета или связей между фасетами. Например, в рамках D-исследования возможно ответить на вопрос, как изменение количества индикаторов или контекстов сценариев скажется на надежности результатов.

Концепция надежности в рамках теории генерализации требует дополнительного пояснения, которое приведено в следующем подразделе.

1.1. Надежность в теории генерализации

В теории генерализации существуют два коэффициента для оценки надежности: коэффициент генерализации (*generalizability coefficient*, E_{p^2}) и коэффициент зависимости (*dependability coefficient*, φ).

Общая формула для коэффициентов надежности имеет следующий вид:

$$C = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(e)}, \quad (3)$$

где C — коэффициент генерализации или зависимости; $\sigma^2(p)$ — дисперсия различий в результатах, связанная с различиями в истинной способности респондентов; $\sigma^2(e)$ — дисперсия ошибки измерения.

Различия между коэффициентами надежности заключаются в определении дисперсии ошибки измерения, которая зависит от дизайна исследования и количества элементов фасета. Для перекрестного дизайна с двумя фасетами из выражения (1) дисперсия ошибки для E_{p^2} (3.1) и φ (3.2) имеет вид:

$$\sigma^2(e) = \frac{\sigma^2(pi)}{n_i} + \frac{\sigma^2(pc)}{n_c} + \frac{\sigma^2(pic,e)}{n_i n_c}; \quad (3.1)$$

$$\sigma^2(e) = \frac{\sigma^2(i)}{n_i} + \frac{\sigma^2(c)}{n_c} + \frac{\sigma^2(ic)}{n_i n_c} + \frac{\sigma^2(pi)}{n_i} + \frac{\sigma^2(pc)}{n_c} + \frac{\sigma^2(pic,e)}{n_i n_c}. \quad (3.2)$$

Для дизайна с вложенными фасетами p x (i : c) дисперсия ошибки для E_{p^2} (3.3) и φ (3.4) имеет вид:

$$\sigma^2(e) = \frac{\sigma^2(pc)}{n_c} + \frac{\sigma^2(pi, pic,e)}{n_i n_c}; \quad (3.3)$$

$$\sigma^2(e) = \frac{\sigma^2(c)}{n_c} + \frac{\sigma^2(i,\ ic)}{n_i n_c} + \frac{\sigma^2(pc)}{n_c} + \frac{\sigma^2(pi, pic,e)}{n_i n_c}, \quad (3.4)$$

где n_i — количество индикаторов; n_c — количество контекстов.

Таким образом, дисперсия ошибки в коэффициенте генерализации учитывает только компоненты, содержащие эффект респондента, поэтому ее называют дисперсией относительной ошибки измерения (*relative error variance*). Дисперсия ошиб-

ки в коэффициенте зависимости дополнительно включает основные эффекты фасетов и взаимодействия между фасетами и называется абсолютной дисперсией ошибки (*absolute error variance*). Коэффициент зависимости всегда будет меньше, чем коэффициент генерализации. Формулы для коэффициентов генерализации и надежности для иных дизайнов исследования можно найти в [Shavelson, Webb, Rowley, 1992].

Выбор коэффициента надежности зависит от типа тестирования. В нормоориентированном тестировании, когда исследователи заинтересованы в упорядочивании респондентов относительно друг друга, рассчитывается коэффициент генерализации. В критериально ориентированном тестировании, когда результаты тестируемых сравниваются с установленным пороговым баллом, рассчитывается коэффициент зависимости.

2. Применение теории генерализации в исследованиях

Методы теории генерализации применяются в исследованиях, ориентированных на изучение фасетов, которые оказывают наибольшее влияние на результаты измерений. Чаще всего это исследования эффектов эксперта и заданий теста [Hild, Gut, Brückmann, 2019].

Для изучения эффектов эксперта обычно используются письменные задания или задания *performance-based*, и эксперты оценивают созданный продукт или поведение респондента в процессе выполнения заданий. В применяемых в рамках оценки персонала заданиях типа «почтовая корзина» (*in-basket*), в которых кандидату требуется проанализировать документы компании, эффект эксперта составляет 3,4% общей дисперсии результатов [Li, Pan, Wang, 2021]. Для заданий на измерение креативности решения задач расхождения в работе экспертов зафиксированы в оценке оригинальности как составляющей креативности (15%), в то время как для других составляющих эффект эксперта ниже: 6% для полноты решения, 2% для практичности решения [Hooijdonk van et al., 2022]. Эффект эксперта на результаты тестирования можно минимизировать за счет обучения экспертов и достижения согласованности между экспертами в отношении критериев оценивания [Hild, Gut, Brückmann, 2019].

В дизайнах, где используется фасет заданий, значительная часть дисперсии результатов связана с различиями в трудности заданий и во взаимодействии тестируемого и задания. Например, Р. Шавелсон, Г. Бакстер и С. Гао рассматривают несколько заданий *performance-based*, в которых эффект взаимодействия тестируемого и задания достигает 48–82% общей дисперсии результатов, т.е. некоторые учащиеся справлялись с одними заданиями лучше, а с другими — хуже [Shavelson, Baxter, Gao, 1993]. В измерении коммуникации между студентами — будущими

стоматологами эффект заданий составлял 35,2%: такой показатель свидетельствует о том, что задания в тесте различались по трудности [Uzun et al., 2018].

При исследовании эффекта заданий в отдельных случаях вводятся дополнительные фасеты: например, оценивается эффект заданий, вложенных в содержательную область (по спецификации теста) [Keller, Clauser, Swanson, 2010], или при исследовании эффекта заданий для измерения читательских навыков оценивается вклад количества абзацев в стимульном материале [Liao, 2023]. Для заданий, предназначенных для оценки навыков письма, анализируют вклад жанра [Bouwer et al., 2015] и темы текста [Wu, Steinkrauss, Lowie, 2023] в баллы тестируемых.

Считается, что наибольший вклад в дисперсию результатов должен вносить эффект респондента. Если это так — значит, инструмент измерения успешно дифференцирует респондентов по уровню способности [Briesch et al., 2014]. На практике из-за превалирующей роли других эффектов (например, из-за сильных различий в трудности заданий) вклад респондентов в результаты оценивания может оказаться небольшим. Например, эффект респондентов составил 2,8% общей дисперсии результатов при измерении коммуникативных навыков студентов [Uzun et al., 2018], около 10% для заданий по чтению и навыку письма [Bouwer et al., 2015; Liao, 2023], 15% для компьютерных симуляций, моделирующих общение врачей и пациентов [Keller, Clauser, Swanson, 2010], от 22 до 28% при измерении креативности решения задач [Hooijdonk van et al., 2022]. Однако существуют примеры, где эффект респондентов превышает 50% [Buyukkidik, Anil, 2015].

Исследования, в которых применяются методы теории генерализации, различаются и по измеряемому конструкту, и по формату инструмента. В данном исследовании используются задания сценарного типа с системой автоматической проверки, в то время как в большинстве описанных выше исследований для оценивания респондентов привлекались эксперты. Кроме того, только в нескольких работах используются инструменты для оценивания универсальных навыков или исследуются фасеты, схожие с контекстом сценариев (например, тема текста в заданиях типа эссе).

3. Описание инструмента

3.1. Контекст как особенность заданий сценарного типа

Особенностью сценарных заданий является наличие контекста. Согласно одному из определений, контекст — это общий стимульный материал для нескольких заданий (тестлеты заданий) [Haladyna, Downing, Rodriguez, 2002]. Другие исследователи отождествляют понятие контекста с понятием сценария (*scenario*) в заданиях в видеоформате [Zhai et al., 2021]. М. Руис-Примо и М. Ли [Ruiz-Primo, Li, 2015] предложили расши-

ренную классификацию характеристик контекста, среди которых: сложность текстовых материалов, степень абстрактности контекста, тип контекста (например, школьный, профессиональный) и др. В той же статье авторы выделяют три уровня контекста: общий контекст (основная проблема сценария, связывающая все индикаторы), контекст для группы индикаторов (например, индикаторы относятся к одному тексту внутри сценария) и индивидуальный контекст индикатора (внутри сценария создается уникальный контекст для одной задачи). Согласно этой классификации индикатор может относиться сразу к нескольким контекстным уровням.

В данной статье мы рассматриваем контекст сценарного задания как стимульный материал, задающий среду тестирования, которая мотивирует респондентов на совершение действий, отражающих конструкт [Davier von, Mislevy, Hao, 2022]. В соответствии с методом доказательной аргументации при разработке тестов (*evidence-centered design*, ECD) [Mislevy, Almond, Lukas, 2003] необходимо разделять свидетельства для оценки выраженности конструкта (например, свидетельством критического мышления может быть действие «выделяет в тексте информацию, релевантную задаче») и контекст, который необходим для стимуляции выполнения нужного действия.

Таким образом, контекст определяет основную проблемную ситуацию сценарного задания (например, тестируемый решил завести домашнего питомца и найти информацию об условиях содержания питомца) и развитие ситуации (сюжета) — последовательность действий, отношения между этапами задания, персонажами и проч. [Грачева, Тарасова, 2022]. При этом одна и та же ситуация может иметь разное тематическое наполнение (например, в качестве питомца может быть кролик или собака), т.е. отличаться контентом (контент стимульного материала, *content of source*) [Homayounzadeh, Saadat, Ahmadi, 2019].

3.2. Инструмент «4К»

Для оценки универсальных навыков используются задания сценарного типа в цифровой среде из инструмента «4К», разработанного в рамках метода доказательной аргументации. Инструмент содержит автоматизированную систему проверки (без привлечения экспертов), он прошел апробацию на выборках из нескольких российских регионов и показал хорошие психометрические характеристики. Результаты валидизации инструмента представлены в нескольких статьях и докладах на научных конференциях [Брун, Орел, Угланова, 2020; Угланова, Жильцова, Лебедева, 2021].

В данной работе рассматриваются сценарии для оценки критического мышления и коммуникации. Согласно концеп-

туальной рамке инструмента, навык критического мышления включает навык работы с информацией в соответствии с целями и условиями поставленной задачи и навык формулирования собственного вывода с помощью результатов, полученных на этапе анализа. Подробнее концептуальная рамка критического мышления представлена в [Uglanova et al., 2022].

Коммуникация в данном инструменте понимается как способность успешно общаться на письме и устно, используя как вербальные, так и невербальные средства. Измерение коммуникации происходит в диалоговом общении с симуляционными аватарами или персонажами сценария (*human-to-agent approach*) [Rosen, 2017]. Таким образом, инструмент оценивает коммуникацию не как часть речевой способности, а как способность решать различные жизненные задачи в условиях сотрудничества с другими людьми — взрослыми или сверстниками. Согласно концептуальной рамке инструмента [Угланова, Жильцова, Лебедева, 2021], конструкт «коммуникация в условиях сотрудничества» включает несколько взаимосвязанных составляющих, соответствующих фазам коммуникативной деятельности:

- ориентация — способность анализировать коммуникативную ситуацию, в том числе информацию о собеседнике, чтобы распределять роли в команде, выявлять общую цель общения, а также адаптировать коммуникативные действия к ситуации общения;
- активная фаза коммуникации — способность распознавать и реализовать коммуникативные намерения в соответствии с социальными и языковыми конвенциями;
- регуляция общения — способность распознавать некорректное коммуникативное поведение собеседника и нарушение социальных норм и адекватно реагировать на них.

В исследовании используются пять заданий сценарного типа для оценки критического мышления и коммуникации. Два сценария («Аквариум» и «Динозавр») направлены на оценку критического мышления, два сценария («Спектакль» и «Торт») — на оценку коммуникации. Еще один сценарий («Путешествие») содержит индикаторы, относящиеся как к коммуникации, так и к критическому мышлению. Далее в статье эти сценарии будут упомянуты как оригинальные.

Дополнительно инструмент «4К» включает альтернативные варианты для каждого оригинального сценария. Разработка альтернативных вариантов происходила с использованием процедуры клонирования для создания заданий с единой структурой и эквивалентными психометрическими характери-

стиками в максимально похожем контексте [Грачева, Тарасова, 2022]. Альтернативные варианты отличаются от оригинальных сценариев только контентом при сохранении проблемы и сюжета. Ниже приводится описание сценарных заданий с целью продемонстрировать разнообразие контекстов инструмента.

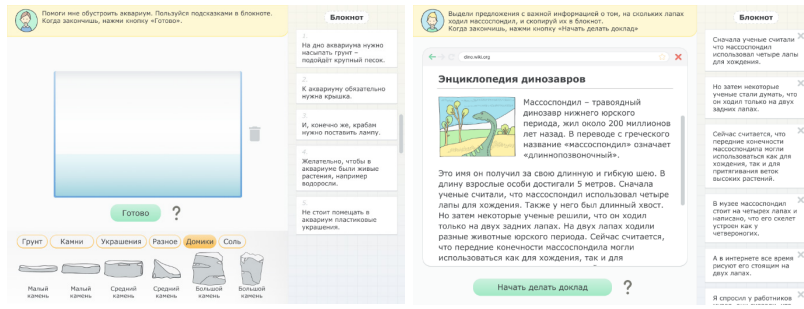
3.3. Описание контекстов сценарных заданий

3.3.1. Сценарное задание «Аквариум» для измерения критического мышления

В основе контекста сценария «Аквариум» — задача обустройства аквариума для крабов. В сценарии моделируется интернет-браузер, в котором тестируемый может изучать тексты электронных статей, выделять и сохранять информацию о предметах, которые понадобятся при обустройстве аквариума. На основе проанализированной информации тестируемый обустраивает аквариум для крабов из ограниченного набора предметов, проявляя при этом способность к формулированию собственного вывода о том, какие предметы должны быть в аквариуме для крабов, а какие нет (рис. 1а).

В альтернативном варианте сценария «Террариум» тестируемые сталкиваются с теми же этапами задания с другим контентом. Здесь главная задача — построить террариум для гекконов. Сценарии содержат по 24 индикатора критического мышления. Подробнее со сценарным заданием «Аквариум» можно ознакомиться в [Грачева, 2022].

Рис. 1. **Примеры экранов из сценарных заданий «Аквариум» и «Динозавр»**



(а) «Аквариум» (б) «Динозавр»

3.3.2. Сценарное задание «Динозавр» для измерения критического мышления

В основе контекста сценария «Динозавр» — подготовка школьного доклада про динозавров. В ходе сценария тестируемому необходимо выбрать наиболее достоверный источник информации (ссылку), проанализировать текст электронной статьи (рис. 1б), сделать вывод о том, на скольких лапах ходил динозавр, и составить слайд для презентации.

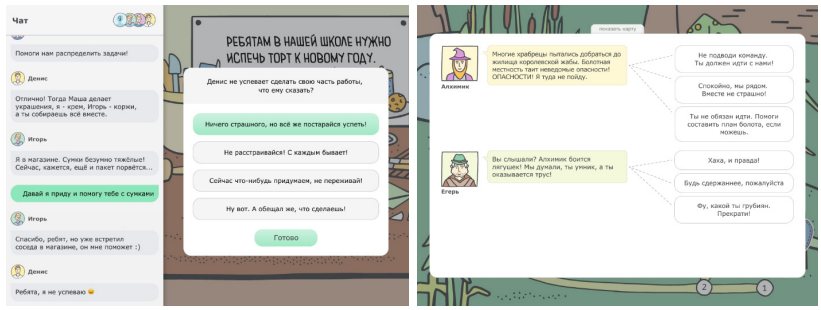
В альтернативном варианте сценария «Еж» тестируемые сталкиваются с теми же задачами с другим контентом. В нем главная цель — подготовка школьного доклада про ежей для ответа на вопрос, зачем ежики трутся иголками о предметы. Сценарии содержат по 6 индикаторов критического мышления.

3.3.3. Сценарное задание «Торт» для измерения коммуникации

Основная проблемная ситуация сценария «Торт» задается стартовым сообщением на экране: «Ребятам в нашей школе нужно испечь торт к Новому году. Помоги им, пожалуйста!». В сценарии используется симуляция чата, где тестируемый общается с персонажами, выбирая сообщение из предложенных (рис. 2а). По сюжету сценария тестируемый сталкивается с разными задачами, например распределяет роли для персонажей для приготовления торта, распознает эмоциональное состояние персонажа, который не справился с задачей, выражает недовольство при нарушении социальных норм (персонаж предлагает съесть торт до праздника).

В альтернативном варианте сценария «Плакат» тестируемые сталкиваются с теми же задачами с другим контентом, здесь главная цель — подготовка плаката к школьному празднику День весны. Сценарии содержат по 15 индикаторов коммуникации.

Рис. 2. **Примеры экранов из сценарного задания «Торт» и «Путешествие»**



(а) «Торт» (б) «Путешествие»

3.3.4. Сценарное задание «Спектакль» для измерения коммуникации

Основная задача сценария «Спектакль» — помочь школьному театру подготовить водолазный костюм для спектакля. В сценарии используется симуляция чата, в котором тестируемый общается с персонажами (аналогично чату на рис. 2а). По сюжету сценария тестируемому необходимо познакомиться с командой, совместно с персонажами выяснить, как выглядит костюм, распределить задачи, задавать уточняющие вопросы, распоз-

нать эмоциональное состояние персонажей и предлагать помощь, где необходимо.

В альтернативном варианте сценария «Конкурс» тестируемому требуется помочь школьному кружку подготовить костюм космонавта для участия в конкурсе. Сценарии содержат по 10 индикаторов коммуникации.

3.3.5. Сценарное задание «Путешествие» для измерения критического мышления и коммуникации

В сценарии «Путешествие» тестируемый оказывается в волшебном королевстве, и местные жители просят его найти ингредиенты для зелья, чтобы вылечить короля. Тестируемый отправляется в путешествие за ингредиентами вместе с командой — Егерем, Алхимиком и Воином. В сценарии тестируемый проявляет свои способности к коммуникации, выбирая реакцию в ответ на сообщение каждого персонажа (рис. 2б). Для проверки навыка критического мышления тестируемого просят сравнить два сообщения, описывающие дорогу к волшебным землям, и выделить предложения, в которых дорога описывается по-разному, либо выбрать нужный ингредиент для зелья, проанализировав информацию от напарников.

В альтернативном варианте сценария «Лабиринт» тестируемые сталкиваются с теми же задачами в другом контенте. Здесь их главная цель — найти волшебные вещи, чтобы помочь расколдовать королеву. Сценарии содержат по 8 индикаторов критического мышления и 22 индикатора коммуникации.

4. Выборка и процедура исследования

В статье используются данные, полученные осенью 2021 г. в ходе тестирования учащихся 4-х классов, которые принимали участие в исследовании универсальных навыков.

Перед началом тестирования администраторам тестирования были высланы руководства с описанием инструмента, этапов подготовки к тестированию и требований к программному обеспечению. Тестирование проходило в школах, в компьютерном классе, каждому респонденту предоставляли персональный компьютер и логин для доступа на сайт инструмента «4К». Для всех учеников получено согласие родителей на участие в исследовании.

В рамках исследования тестируемым предлагалось выполнить пять оригинальных заданий и дополнительно два альтернативных варианта заданий, которые назначались случайным образом. Альтернативный вариант сценария мог предъявляться как до соответствующего оригинального сценария, так и после него. Тестирование проходило в два дня и в сумме занимало около 80 минут.

Описанная процедура исследования позволила получить данные по всем заданиям сценарного типа (ориги-

нальные и альтернативные варианты сценариев) за ограниченное время тестирования. Данные для каждой пары сценариев получены по случайным подвыборкам респондентов: 998 респондентов — «Аквариум»/«Террариум», 1096 респондентов — «Спектакль»/«Конкурс», 1052 респондента — «Путешествие»/«Лабиринт», 868 респондентов — «Торт»/ «Плакат», 466 респондентов — «Динозавр»/«Еж».

5. Методология анализа в рамках теории генерализации

Для анализа вклада контекста сценарного задания в результаты оценивания универсальных навыков предлагаются два дизайна G-исследования. В дизайнах используются фасеты со случайным эффектом.

5.1. Дизайн G-исследования

Первый дизайн — с вложенными фасетами $p \times (i : c)$, где p — респонденты, $(i : c)$ — индикаторы, вложенные в сценарные задания. Он позволит оценить часть различий в баллах, связанных с контекстом сценария, а также с взаимодействием контекста и респондента. В анализе используются данные по оригинальным сценариям. Для оценки критического мышления предлагаются три сценария с разным контекстом: «Аквариум», «Динозавр», «Путешествие», для оценки коммуникации — три сценария с разным контекстом: «Спектакль», «Путешествие», «Торт».

Второй дизайн — перекрестный с двумя фасетами $p \times i \times v$, где p — респонденты, i — индикаторы, v — вариант сценарного задания (оригинальный или альтернативный). Данный дизайн позволит оценить компоненты дисперсии, отражающие различия в контексте вариантов заданий сценарного типа, которые были разработаны как взаимозаменяемые (только с изменением контента). Анализ проводится отдельно для каждой пары вариантов сценариев на подвыборках респондентов. Для оценки критического мышления используются три пары сценариев: «Аквариум»/«Террариум», «Динозавр»/«Еж», «Путешествие»/«Лабиринт», для оценки коммуникации — также три пары сценариев: «Путешествие»/«Лабиринт», «Торт»/«Плакат», «Спектакль/«Конкурс».

5.2. Дизайн D-исследования

Результаты G-исследований используются в D-исследованиях для оценки надежности измерений универсальных навыков. Инструмент «4К» создавался для целей упорядочивания респондентов по уровню развития универсальных навыков, поэтому для оценки надежности измерений в рамках теории генерализации наиболее подходящим является коэффициент генерализации. Тем не менее будет рассчитан и коэффициент зависимости для целей критериального тестирования.

Для первого дизайна G-исследования коэффициенты надежности отражают внутреннюю согласованность заданий

инструмента «4K» при измерении критического мышления и коммуникации. Для второго дизайна G-исследования коэффициенты надежности отражают ретестовую надежность — возможность получения одинаковых результатов у тестируемых по взаимозаменяемым вариантам сценарных заданий. Значения коэффициентов надежности ниже 0,7 свидетельствуют об неудовлетворительной надежности, от 0,7 до 0,8 — об удовлетворительной надежности, значения выше 0,8 — о высокой надежности [Engelhardt, 2009].

Одна из целей D-исследования состоит в подборе такой процедуры тестирования, которая позволит получить надежность измерения не ниже удовлетворительной (0,7). В соответствии с этой целью будет определено количество элементов фасета (индикаторов, контекстов, вариантов сценариев), достаточное для достижения удовлетворительных показателей надежности.

Анализ проведен в программной среде R с использованием пакетов *gtheory* и *lme4* [Huebner, Lucht, 2019].

6. Результаты

6.1. Описательные статистики по сценарным заданиям

Для измерения критического мышления (КМ) используются 37 индикаторов из трех заданий сценарного типа. Средний балл критического мышления на полной выборке (2255 респондентов) равен 21,35 (стандартное отклонение = 7,20). Для измерения коммуникации используются 47 индикаторов из трех заданий сценарного типа. Средний балл коммуникации (КО) на полной выборке (2015 респондентов) равен 30,69 (стандартное отклонение = 7,41). Все индикаторы сценариев дихотомизированы для удобства интерпретации результатов (индикаторы принимают значения 0 или 1).

В табл. 1 приведены описательные статистики для подвыборок респондентов, которые выполняли разные пары вариантов сценариев, с указанием количества индикаторов в каждом сценарии по каждому из навыков.

Таблица 1. **Описательные статистики по сценарным заданиям**

Сценарные задания	*N*	Навык	Среднее (стандартное отклонение)
«Путешествие»/«Лабиринт»	22	КО	15,41 (3,99) / 15,48 (4,22)
«Спектакль»/«Конкурс»	10	КО	6,83 (2,16) / 6,76 (2,06)
«Торт»/«Плакат»	15	КО	9,97 (2,84) / 9,62 (2,89)
«Путешествие»/«Лабиринт»	7	КМ	4,20 (1,63) / 4,21 (1,69)
«Аквариум»/«Террариум»	24	КМ	13,53 (5,61) / 13,40 (5,64)
«Динозавр»/«Еж»	6	КМ	2,76 (1,60) / 3,00 (1,58)

Примечание: ***N*** — количество индикаторов КМ или КО сценарного задания.

6.2. Результаты первого дизайна G-исследования

В табл. 2 представлены оцененные компоненты дисперсии для первого дизайна G-исследования *p* x (*i* : *c*). Основной эффект контекста сценария (*c*) оказался минимальным как для КМ (1%), так и для КО (0%). В среднем результаты оценки универсальных компетенций не различаются в зависимости от предлагаемого контекста сценария, т.е. уровень трудности сценарных заданий с разным контекстом практически одинаков. Тем не менее часть дисперсии результатов связана с эффектом взаимодействия тестируемого и контекста (*p* x *c*: 6,3% для КМ, 4,1% для КО). Например, тестируемый успешно справляется со сценарием «Аквариум», но результат в сценарии «Динозавр» оказывается ниже, а у другого тестируемого — наоборот.

Часть дисперсии результатов связана с различиями в трудности индикаторов внутри сценарного задания, причем индикаторы, предназначенные для оценки КМ, менее гомогенны по трудности (13,9%), чем индикаторы для оценки КО (8,1%). Анализ в рамках теории генерализации предполагает, что все индикаторы сценариев должны быть одинаковой трудности, чтобы эффект индикаторов был минимальным. Однако на практике может стоять задача разработки заданий разной трудности [Arterberry et al., 2014]. Например, легкие задания предъявляются в тестах в первую очередь, чтобы снизить тревожность респондентов перед тестированием. Нивелирование эффекта трудности индикаторов возможно за счет увеличения количества индикаторов (заданий) теста.

Таблица 2. **Оцененные компоненты дисперсии для дизайна *p* x (*i*: *c*)**

Компонент	Коммуникация		Критическое мышление	
	Дисперсия	%	Дисперсия	%
p	0,018	7,8	0,024	9,9
c	0,000	0,0	0,002	1,0
p x *c*	0,009	4,1	0,015	6,3
i : *c*	0,018	8,1	0,034	13,9
e	0,182	80	0,168	68,9

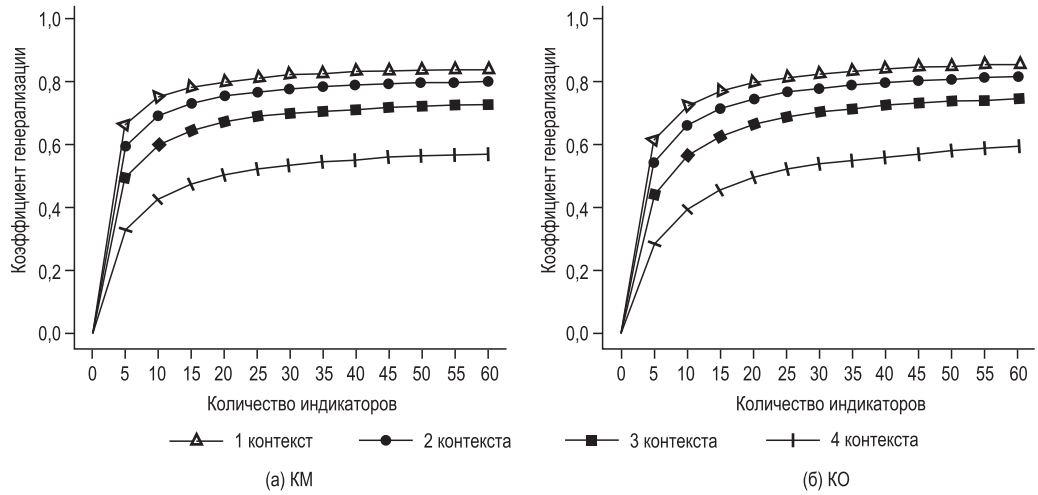
Остаточный компонент дисперсии (*e*), который включает эффект взаимодействия всех фасетов и компонент случайной или систематической ошибки измерения, вкладывается в измерения в большей степени (68,9% для КМ и 80% для КО). Полученный результат может свидетельствовать о наличии фасетов, которые не были учтены в предложенном G-дизайне.

Оцененные компоненты дисперсии по разным источникам позволяют рассчитать коэффициенты генерализации и зави-

симости. Надежность для обоих навыков удовлетворительна и близка к высоким значениям (0,8): коэффициент генерализации (E_{ρ^2}) для КМ равен 0,79 (коэффициент зависимости φ — 0,76), для КО — 0,80 (коэффициент зависимости — 0,79).

На рис. 3 представлены значения коэффициента генерализации (надежность для нормоориентированного тестирования) при разном количестве индикаторов и контекстов сценариев. Процедура тестирования, в которой все индикаторы критического мышления (38) или коммуникации (47) принадлежат одному контексту, не обеспечивает удовлетворительную надежность для нормоориентированного тестирования (ниже 0,7). При двух контекстах достижение удовлетворительного значения коэффициента генерализации (выше 0,7) возможно при не менее чем 30 индикаторах. Использование сценариев с тремя разными контекстами позволяет достичь удовлетворительных значений коэффициента генерализации при не менее чем 11 индикаторах КМ и 15 индикаторах КО и высоких значений (выше 0,8) — при не менее чем 55 индикаторах КМ и 45 индикаторах КО. Четырех контекстов с 7 индикаторами КМ и 10 индикаторами КО будет достаточно для достижения удовлетворительных значений коэффициента генерализации и с 20 индикаторами КМ и 25 индикаторами КО — для высоких значений.

Рис. 3. **Коэффициент генерализации при разных количествах индикаторов и контекстов сценарных заданий**



Таким образом, использование нескольких контекстов сценарных заданий для измерения универсальных навыков позволяет существенно повысить надежность и степень генерализации результатов, а также сократить время тестирования за счет использования меньшего количества индикаторов.

6.3. Результаты второго дизайна G-исследования

Во втором дизайне исследования анализ проводится отдельно по парам вариантов заданий сценарного типа. В табл. 3 приведены оцененные компоненты дисперсии для сценариев КО.

Таблица 3. **Оцененные компоненты дисперсии для дизайна *p* x *i* x v** (коммуникация)

Компонент	«Путешествие»/«Лабиринт»		«Торт»/«Плакат»		«Спектакль»/«Конкурс»	
	Дисперсия	%	Дисперсия	%	Дисперсия	%
p	0,024	11,3	0,023	10,1	0,025	11,4
i	0,009	4,2	0,022	10,0	0,020	9,3
v	0,000	0,0	0,000	0,1	0,000	0,0
p x *i*	0,036	17,2	0,045	19,6	0,046	20,9
p x *v*	0,004	1,7	0,001	0,6	0,003	1,2
i x *v*	0,002	1,0	0,001	0,6	0,000	0,2
e	0,136	64,7	0,135	59,0	0,125	57,1

Перекрестный дизайн исследования позволил оценить долю дисперсии результатов, связанную со взаимодействием тестируемого и индикаторов сценария: она составляет примерно пятую часть дисперсии баллов (*p* x *i*: от 17,2 до 20,9%). На основании этого показателя можно заключить, что тестируемые непоследовательны в своих ответах на разные индикаторы сценариев. С одной стороны, полученный результат может быть связан с различиями индикаторов по трудности (*i*: от 4,2 до 10%). С другой стороны, согласно теоретической рамке инструмента, коммуникация включает несколько взаимосвязанных составляющих, что могло стать причиной разного восприятия индикаторов отдельными тестируемыми.

Вариант сценария (*v*) в меньшей степени вкладывается в дисперсию результатов. Следовательно, темы сценариев в разных вариантах не различаются по трудности. Эффект взаимодействия варианта и тестируемого/индикаторов присутствует, но он невысокий — до 2%. Таким образом, варианты сценарного типа с изменением только контента могут считаться взаимозаменяемыми.

В табл. 4 приведены оцененные компоненты дисперсии для сценариев, измеряющих критическое мышление.

Процент дисперсии, связанный с взаимодействием респондентов с индикаторами, для компетенции критического мышления находится в диапазоне от 11,5 до 12,6%. Эффект различий в трудности индикаторов различается по сценариям. Например, в сценариях «Путешествие»/«Лабиринт» различия в трудности индикаторов в 2 раза меньше, чем в сценариях «Аквариум»/«Террариум». Одной из причин таких различий может быть разное количество индикаторов.

Таблица 4. **Оцененные компоненты дисперсии для дизайна *p* x *i* x *v* (критическое мышление)**

Компонент	«Путешествие»/«Лабиринт»		«Аквариум»/«Террариум»		«Динозавр»/«Еж»	
	Дисперсия	%	Дисперсия	%	Дисперсия	%
p	0,034	14,9	0,041	16,7	0,036	14,1
i	0,015	6,6	0,038	15,6	0,023	9,1
v	0,000	0,00	0,000	0,00	0,001	0,2
p x *i*	0,027	11,7	0,031	12,6	0,029	11,5
p x *v*	0,003	1,4	0,006	2,3	0,003	1,5
i x *v*	0,002	1,2	0,001	0,5	0,002	0,7
e	0,146	64,1	0,129	52,4	0,161	63,3

Вариант сценария привносит меньше дисперсии в результаты оценивания (*v*: 0–0,2%). Присутствует небольшой (0,5–2,3%) эффект взаимодействия варианта сценария с индикаторами/респондентами.

В рамках D-исследования рассчитаны коэффициенты генерализации (E_{p2}) и зависимости (φ) как меры ретестовой надежности (табл. 5).

Таблица 5. **Ретестовая надежность для двух вариантов сценариев**

Показатель	Коммуникация			Критическое мышление		
	«Путешествие»	«Торт»	«Спектакль»	«Аквариум»	«Динозавр»	«Путешествие»
(E_{p2})	0,78	0,74	0,67	0,86	0,64	0,60
φ	0,77	0,70	0,64	0,83	0,6	0,55
N	22	15	10	24	6	7

Примечание: для удобства в шапке таблицы указаны названия только оригинального варианта сценария. *N* — количество индикаторов в сценарном задании; E_{p2} — коэффициент генерализации; φ — коэффициент зависимости.

Для большинства сценариев коэффициенты ретестовой надежности удовлетворительны (выше 0,7). Неудовлетворительная ретестовая надежность (ниже 0,7) характерна для сценариев с наименьшим количеством индикаторов в каждом сценарии (до 10 включительно).

D-исследование позволило оценить минимальное количество индикаторов, необходимое для достижения удовлетворительной ретестовой надежности при нормоориентированном тестировании (табл. 6).

Судя по полученным результатам, при нормоориентированном тестировании с двумя вариантами сценариев 12–13 инди-

Таблица 6. **Минимальное количество индикаторов, необходимое для достижения удовлетворительной ретестовой надежности (коэффициента генерализации)**

Количество вариантов	Коммуникация			Критическое мышление		
	«Путешествие»	«Торт»	«Спектакль»	«Аквариум»	«Динозавр»	«Путешествие»
2	13	12	12	7	8	12
3	9	10	9	5	6	9
4	8	9	8	4	6	7

каторов в каждом варианте достаточно для достижения нижней границы ретестовой надежности. Повышение надежности возможно за счет увеличения количества вариантов теста. Однако с учетом затрат времени и ресурсов для повышения ретестовой надежности оптимальной стратегией является разработка дополнительных индикаторов (от 2 до 5) в существующие варианты сценарных заданий.

7. Обсуждение результатов

Проведено исследование с целью проанализировать роль контекста сценарных заданий при измерении универсальных навыков. Применение методов теории генерализации позволило оценить вклад контекста в результаты (G-исследование) и установить, какое количество контекстов необходимо для достижения удовлетворительной надежности измерений (D-исследование).

Исследование проводилось на данных, полученных при выполнении 4-классниками заданий методики «4К» — инструмента сценарного типа, предназначенного для измерения критического мышления и коммуникации. По результатам G-исследования для обоих навыков контексты сценарных заданий оказались практически одинаковой трудности, однако результаты тестируемых по разным сценариям различались, что проявилось в наличии эффекта взаимодействия тестируемого и контекста (4,2% для коммуникации, 6,2% для критического мышления). В предыдущих исследованиях в формате *performance-based* эффект взаимодействия тестируемого и контекста заданий был одним из наиболее сильных [Shavelson, Baxter, Gao, 1993; Hild, Gut, Brückmann, 2019]. Контекст может вовлекать и мотивировать одного тестируемого, способствуя успешному выполнению заданий, и в то же время сбить с толку другого тестируемого и исказить его результаты [Messick, 1994].

Для снижения эффекта взаимодействия тестируемого и контекста тестирование универсальных навыков необходимо проводить с использованием нескольких сценариев с разными

контекстами. D-исследование позволило оценить количество контекстов и индикаторов, необходимое для достижения удовлетворительной надежности измерений при нормоориентированном тестировании (с использованием коэффициента генерализации).

При оценивании универсальных навыков с применением одного контекста значения коэффициента генерализации оказались неудовлетворительными (ниже 0,7), при этом увеличение количества контекстов позволяет повысить значение этого коэффициента надежности. Согласно расчетам, 30 индикаторов критического мышления или коммуникации в двух контекстах обеспечивают такую же надежность для нормоориентированного тестирования (0,7), как 15 индикаторов коммуникации или 11 индикаторов критического мышления в трех разных контекстах. Проведенный анализ наглядно демонстрирует возможности D-исследования при проектировании инструмента: для оценки универсальных навыков рекомендуется использовать минимум два сценарных задания с разным контекстом — при таких условиях достигаются удовлетворительные значения коэффициента генерализации, а использование трех контекстов позволит существенно сократить количество индикаторов без потери надежности. Если для инструмента поставлены более строгие критерии надежности (0,8 и выше), рекомендуется использовать четыре сценарных задания с разными контекстами.

Повторное использование контекстных заданий неминуемо ведет к появлению эффекта запоминания — а значит, и к искажениям результатов тестирования. Для решения этой проблемы создаются альтернативные варианты заданий, которые используются как замена оригинальных. Например, в заданиях, оценивающих навыки письма, для создания сопоставимых вариантов и снижения эффекта запоминания изменяют тему стимульного материала, при этом другие характеристики задания, например жанр, сохраняются [Wu, Steinkrauss, Lowie, 2023].

Данные текущего исследования позволили оценить роль контекста для вариантов заданий сценарного типа с максимально похожими контекстами, отличающихся только темой ситуации (контентом). Согласно результатам G-исследования, для всех сценарных заданий, независимо от тестируемого навыка, вклад варианта в тестовые баллы минимален (0–0,2%). Тестируемые показывают стабильные результаты в обоих вариантах сценариев, хотя размер эффекта различается между парами сценариев (эффект взаимодействия тестируемого и варианта находится в диапазоне от 0,6 до 2,3%).

В анализе генерализации нет принятых границ для интерпретации размера эффекта. Исследователи, работающие в рамках данного подхода, рекомендуют сравнивать размеры эф-

фекта, полученного при разных условиях [Briesch et al., 2014]. Результаты текущего исследования позволяют сделать вывод, что фасет варианта сценария в меньшей степени вкладывается в результаты измерений универсальных навыков, однако этот эффект не является нулевым. Таким образом, замена только контента сценария позволяет создать в целом сопоставимые варианты заданий сценарного типа.

Для альтернативных вариантов в рамках D-исследования оценивалась ретестовая надежность — степень, в которой результаты тестируемых воспроизводятся в вариантах заданий, отдельно для пар вариантов сценариев. Ретестовая надежность оказалась неудовлетворительной для тех сценариев, где число индикаторов в каждом варианте меньше 10. Для обеспечения удовлетворительной ретестовой надежности при нормоориентированном тестировании по двум вариантам сценариев достаточно 12–13 индикаторов в каждом варианте.

Помимо надежности, которой в теории генерализации уделяется большое внимание, при оценивании универсальных навыков необходимо обеспечить валидность измерения. Использование нескольких сценариев позволяет не только уменьшить ошибку измерения, но и получить более полную картину поведения тестируемых, которое отражает целевой конструкт. Оценивание сложных навыков должно производиться в разных контекстах, чтобы респонденты могли продемонстрировать свои способности в разных, в том числе незнакомых для них, ситуациях [Wang, Liu, Hau, 2022]. Напротив, разработка альтернативных вариантов заданий сценарного типа подразумевает подбор нового контекста, в котором поведение тестируемого останется стабильным, чтобы выводы по результатам тестирования были справедливы для всех респондентов, независимо от предъявляемого варианта. Использование схожих контекстов, различающихся темами ситуации, представляется оптимальной стратегией для создания сопоставимых вариантов заданий сценарного типа.

Методы теории генерализации и результаты данного исследования могут быть полезны разработчикам при проектировании заданий сценарного типа. Разработка сценарных заданий требует немалых ресурсов [Угланова, Брун, Васин, 2018], в особенности если задания реализованы в цифровой среде. Решение о количестве контекстов и индикаторов должно приниматься с учетом цели тестирования и имеющихся ресурсов.

8. Ограничения и дальнейшие направления исследования

Результаты исследования следует воспринимать и использовать с учетом ограничений.

Рекомендации о количестве индикаторов и контекстов даны для разных дизайнов G-исследования — набора фасетов и отно-

шений между ними. При проектировании инструмента следует брать в расчет рекомендации по дизайну исследования, который будет реализован в конкретной ситуации тестирования. Например, инструмент измерения «4К» для оценки критического мышления содержит три сценария, в этом случае для оценки ретестовой надежности нужно рассматривать три сценария в каждом варианте как единое целое, а не как отдельные пары сценариев.

В данном исследовании изучался эффект общего контекста задания, а не отдельных характеристик контекста, на результаты тестирования. Изучение роли отдельных характеристик контекста (например, необходимости предметных знаний для решения задачи, приближенности контекста к реальности, сложности контекста для респондентов разного возраста, количества персонажей, ветвей сюжета и проч.) позволит глубже понять функционирование контекстных заданий. Также в будущих исследованиях целесообразно проанализировать контекстную нагруженность индикаторов в разрезе трех уровней контекста, предложенных в [Ruiz-Primo, Li, 2015]: общего контекста, контекста для группы индикаторов и индивидуального контекста индикатора. М. Руис-Примо и М. Ли анализировали среднюю трудность для групп индикаторов только с одним контекстным уровнем, двумя и тремя уровнями (например, для двух уровней: индикаторы одновременно объединены общим контекстом и специфическим групповым контекстом). Работы в этом направлении могут быть продолжены анализом эффекта контекста в зависимости от количества уровней (контекстной нагруженности) индикаторов.

С точки зрения доказательства валидности выводов по итогам тестирования особого внимания заслуживает изучение поведения респондента в сценариях, контекст которых приближен к реальной жизни. Например, отличается ли выраженность навыков коммуникации у ученика в тестовой среде, симулирующей учебную ситуацию, от реального поведения в классе? В будущих исследованиях вклад контекста с использованием теории генерализации может быть дополнительно рассмотрен в рамках концепции трансфера знаний и навыков из одного контекста в другой [Barnett, Ceci, 2002].

В данном исследовании используется ограниченное число фасетов для объяснения результатов оценивания. Другие фасеты или отношения между фасетами могут быть протестированы для сравнения результатов и снижения доли дисперсии, связанной со случайной или систематической ошибкой измерения.

В текущем исследовании каждый универсальный навык рассматривается как одномерный конструкт, в то время как теоретическая рамка инструмента подразделяет навыки на несколько составляющих. В предыдущем исследовании, осу-

ществленном в другой методологии, анализ проводился по отдельным составляющим критического мышления и на примере пары сценариев «Аквариум»/«Террариум» было показано, что оценки по вариантам значимо различаются между собой для навыка формулирования вывода и не различаются для навыка анализа информации [Грачева, 2022]. Применение многомерной теории генерализации позволит проанализировать возможности сценарных заданий в измерении отдельных составляющих навыков [Keller, Clauser, Swanson, 2010]. Кроме того, в данной работе применяются «классические» методы теории генерализации для работы с сырыми баллами тестирования. В современных работах в этом направлении предпринимаются попытки переложить идеи теории генерализации на модели структурных уравнений [Jorgensen, 2021] или байесовских сетей [Jiang, Skorupski, 2018].

9. Заключение

Измерение универсальных навыков с использованием сценарных заданий — нетривиальная задача для разработчиков и психометриков. В данной статье показано, как применение методов теории генерализации позволяет оценить вклад контекста задания в результаты тестирования для случаев, когда используются разные сценарии или сценарии с похожими контекстами (альтернативные варианты). Полученные результаты используются для предсказания надежности измерений при разном количестве контекстов или индикаторов сценария.

Таким образом, теория генерализации предлагает гибкий подход к проектированию структуры теста для разных форматов заданий и целей тестирования. Результаты анализа позволяют дать конкретные рекомендации по улучшению организации тестирования и достижению удовлетворительной надежности измерений.

Благодарности

Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ.

Автор благодарит И.Л. Угланову за комментарии и поддержку в подготовке статьи.

Литература

1. Брун И.В., Орел Е.А., Угланова И.Л. (2020) Измерение креативности и критического мышления в начальной школе. *Психологический журнал*, т. 41, № 6, сс. 96–107. https://doi.org/10.31857/S020595920011124-2
2. Грачева Д.А. (2022) Анализ сопоставимости измерения метапредметных навыков в цифровой среде. *Психологическая наука и образование*, т. 27, № 6, сс. 57–67. https://doi.org/10.17759/pse.2022270605
3. Грачева Д.А., Тарасова К.В. (2022) Подходы к разработке вариантов заданий сценарного типа в рамках метода доказательной аргументации.

Отечественная и зарубежная педагогика, т. 1, № 3, сс. 83–97. https://doi.org/10.24412/2224-0772-2022-84-83-97

4. Угланова И., Брун И., Васин Г. (2018) Методология *Evidence-Centered Design* для измерения комплексных психологических конструктов. *Современная зарубежная психология*, т. 7, № 3, сс. 18–27. https://doi.org/10.17759/jmfp.2018070302
5. Угланова И.Л., Жильцова Л.Ю., Лебедева М.Ю. (2021) Измерение навыков коммуникации и кооперации в начальной и средней школе: могут ли школьники договориться с инопланетянином? Материалы *V Международной научной конференции «Информатизация образования и методика электронного обучения: цифровые технологии в образовании» (Красноярск, 20–23 сентября 2022 г.),* Красноярск: Сибирский федеральный университет, сс. 682–686.
6. Arterberry B.J., Martens M.P., Cadigan J.M., Rohrer D. (2014) Application of Generalizability Theory to the Big Five Inventory. *Personality and Individual Differences*, vol. 69, October, pp. 98–103. https://doi.org/10.1016/j.paid.2014.05.015
7. Barnett S.M., Ceci S.J. (2002) When and Where Do We Apply What We Learn?: A Taxonomy for Far Transfer. *Psychological Bulletin*, vol. 128, no 4, pp. 612–637. https://doi.org/10.1037/0033-2909.128.4.612
8. Bouwer R., Béguin A., Sanders T., van den Bergh H. (2015) Effect of Genre on the Generalizability of Writing Scores. *Language Testing*, vol. 32, no 1, pp. 83–100. https://doi.org/10.1177/0265532214542994
9. Brennan R.L. (1992) Generalizability Theory. *Educational Measurement: Issues and Practice,* vol. 11, no 4, pp. 27–34. https://doi.org/10.1111/j.1745-3992.1992.tb00260.x
10. Briesch A.M., Swaminathan H., Welsh M., Chafouleas S.M. (2014) Generalizability Theory: A Practical Guide to Study Design, Implementation, and Interpretation. *Journal of School Psychology*, vol. 52, no 1, pp. 13–35. https://doi.org/10.1016/j.jsp.2013.11.008
11. Buyukkidik S., Anil D. (2015) Investigation of Reliability in Generalizability Theory with Different Designs on Performance-Based Assessment. *Education and Science*, vol. 40, no 117, pp. 285–296. http://dx.doi.org/10.15390/EB.2015.2454
12. Cronbach L.J., Gleser G.C., Nanda H., Rajaratnam N. (1972) *The Dependability of Behavioral Measurements*. New York, NY: Wiley.
13. Davier von A.A., Mislevy R.J., Hao J. (eds) (2021) *Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment: With Examples in R and Python.* Cham: Springer International. https://doi.org/10.1007/978-3-030-74394-9
14. Engelhardt P.V. (2009) An Introduction to Classical Test Theory as Applied to Conceptual Multiple-Choice Tests. *Getting Started in PER* (eds C. Henderson, K. Harper), College Park, MD: American Association of Physics Teachers, pp. 1–40.
15. Haladyna T.M., Downing S.M., Rodriguez M.C. (2002) A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, vol. 15, no 3, pp. 309–333. https://doi.org/10.1207/S15324818AME1503_5
16. Hild P., Gut C., Brückmann M. (2019) Validating Performance Assessments: Measures That May Help to Evaluate Students' Expertise in 'Doing Science'. *Research in Science & Technological Education*, vol. 37, no 4, pp. 419–445. https://doi.org/10.1080/02635143.2018.1552851
17. Homayounzadeh M., Saadat M., Ahmadi A. (2019) Investigating the Effect of Source Characteristics on Task Comparability in Integrated Writing Tasks. *Assessing Writing*, vol. 41, no 2, pp. 25–46. https://doi.org/10.1016/j.asw.2019.05.003

18. Hooijdonk van M., Mainhard T., Kroesbergen E.H., van Tartwijk J. (2022) Examining the Assessment of Creativity with Generalizability Theory: An Analysis of Creative Problem Solving Assessment Tasks. *Thinking Skills and Creativity*, vol. 43, no 1, Article no 100994. https://doi.org/10.1016/j.tsc.2021.100994
19. Huebner A., Lucht M. (2019) Generalizability theory in R. *Practical Assessment, Research, and Evaluation*, vol. 24, no 5. https://doi.org/10.7275/5065-gc10
20. Jiang Z., Skorupski W. (2018) A Bayesian Approach to Estimating Variance Components Within a Multivariate Generalizability Theory Framework. *Behavior Research Methods*, vol. 50, no 3, pp. 2193–2214. https://doi.org/10.3758/s13428-017-0986-3
21. Jorgensen T.D. (2021) How to Estimate Absolute-Error Components in Structural Equation Models of Generalizability Theory. *Psych*, vol. 3, no 2, pp. 113–133. https://doi.org/10.3390/psych3020011
22. Keller L.A., Clauser B.E., Swanson D.B. (2010) Using Multivariate Generalizability Theory to Assess the Effect of Content Stratification on the Reliability of a Performance Assessment. *Advances in Health Sciences Education*, vol. 15, no 5, pp. 717–733. https://doi.org/10.1007/s10459-010-9233-8
23. Li G., Pan Y., Wang W. (2021) Using Generalizability Theory and Many-Facet Rasch Model to Evaluate In-Basket Tests for Managerial Positions. *Frontiers in Psychology*, vol. 12, July, Article no 660553. https://doi.org/10.3389/fpsyg.2021.660553
24. Liao R.J. (2023) The Use of Generalizability Theory in Investigating the Score Dependability of Classroom-Based L2 Reading Assessment. *Language Testing*, vol. 40, no 1, pp. 86–106. https://doi.org/10.1177/02655322211070840
25. Messick S. (1994) The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*, vol. 23, no 2, pp. 13–23. https://doi.org/10.3102/0013189X023002013
26. Mislevy R.J., Almond R.G., Lukas J.F. (2003) *A Brief Introduction to Evidence-Centered design. ETS Research Report Series no 2003(1)*. Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x
27. Rosen Y. (2017) Assessing Students in Human-to-Agent Settings to Inform Collaborative Problem-Solving Learning. *Journal of Educational Measurement*, vol. 54, no 1, pp. 36–53. https://doi.org/10.1111/jedm.12131
28. Ruiz-Primo M.A., Li M. (2015) The Relationship between Item Context Characteristics and Student Performance: The Case of the 2006 and 2009 PISA Science Items. *Teachers College Record*, vol. 117, no 1, pp. 1–36. https://doi.org/10.1177/016146811511700118
29. Shavelson R.J., Baxter G.P., Gao X. (1993). Sampling Variability of Performance Assessments. *Journal of Educational Measurement*, vol. 30, no 3, pp. 215–232. https://doi.org/10.1111/j.1745-3984.1993.tb00424.x
30. Shavelson R.J., Webb N.M., Rowley G.L. (1992) Generalizability Theory. *Methodological Issues & Strategies in Clinical Research* (ed. A.E. Kazdin), American Psychological Association, pp. 233–256. http://dx.doi.org/10.1037/10109-051
31. Uglanova I., Orel E., Gracheva D., Tarasova K. (2023) Computer-Based Performance Approach for Critical Thinking Assessment in Children. *British Journal of Educational Psychology*, vol. 93, no. 2, pp. 531–544. https://doi.org/10.1111/bjep.12576
32. Uzun N.B., Aktas M., Asiret S., Yormaz S. (2018) Using Generalizability Theory to Assess the Score Reliability of Communication Skills of Dentistry Students. *Asian Journal of Education and Training*, vol. 4, no 2, pp. 85–90. https://doi.org/10.20448/journal.522.2018.42.85.90
33. Wang D., Liu H., Hau K.T. (2022) Automated and Interactive Game-Based Assessment of Critical Thinking. *Education and Information Technologies*, vol. 27, no 4, pp. 4553–4575. https://doi.org/10.1007/s10639-021-10777-9

34. Wu M.Y., Steinkrauss R., Lowie W. (2023) The Reliability of Single Task Assessment in Longitudinal L2 Writing Research. *Journal of Second Language Writing*, vol. 59, no 4, Article no 100950. https://doi.org/10.1016/j.jslw.2022.100950
35. Zhai X., Haudek K.C., Wilson C., Stuhlsatz M. (2021) A Framework of Construct-Irrelevant Variance for Contextualized Constructed Response Assessment. *Frontiers in Education*, vol. 6, October, Article no 751283. https://doi.org/10.3389/feduc.2021.751283

References

Arterberry B.J., Martens M.P., Cadigan J.M., Rohrer D. (2014) Application of Generalizability Theory to the Big Five Inventory. *Personality and Individual Differences*, vol. 69, October, pp. 98–103. https://doi.org/10.1016/j.paid.2014.05.015

Barnett S.M., Ceci S.J. (2002) When and Where Do We Apply What We Learn?: A Taxonomy for Far Transfer. *Psychological Bulletin*, vol. 128, no 4, pp. 612–637. https://doi.org/10.1037/0033-2909.128.4.612

Bouwer R., Béguin A., Sanders T., van den Bergh H. (2015) Effect of Genre on the Generalizability of Writing Scores. *Language Testing*, vol. 32, no 1, pp. 83–100. https://doi.org/10.1177/0265532214542994

Brennan R.L. (1992) Generalizability Theory. *Educational Measurement: Issues and Practice*, vol. 11, no 4, pp. 27–34. https://doi.org/10.1111/j.1745-3992.1992.tb00260.x

Briesch A.M., Swaminathan H., Welsh M., Chafouleas S.M. (2014) Generalizability Theory: A Practical Guide to Study Design, Implementation, and Interpretation. *Journal of School Psychology*, vol. 52, no 1, pp. 13–35. https://doi.org/10.1016/j.jsp.2013.11.008

Brun I.V., Orel E.A., Uglanova I.L. (2020) Izmerenie kreativnosti i kriticheskogo myshleniya v nachal'noy shkole [Measuring Creativity and Critical Thinking in Primary School]. *Psikhologicheskij zhurnal*, vol. 41, no 6, pp. 96–107. https://doi .org/10.31857/S020595920011124-2

Buyukkidik S., Anil D. (2015) Investigation of Reliability in Generalizability Theory with Different Designs on Performance-Based Assessment. *Education and Science*, vol. 40, no 117, pp. 285–296. http://dx.doi.org/10.15390/EB.2015.2454

Cronbach L.J., Gleser G.C., Nanda H., Rajaratnam N. (1972) *The Dependability of Behavioral Measurements.* New York, NY: Wiley.

Davier von A.A., Mislevy R.J., Hao J. (eds) (2021) *Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment: With Examples in R and Python.* Cham: Springer International. https://doi.org/10.1007/978-3-030-74394-9

Engelhardt P.V. (2009) An Introduction to Classical Test Theory as Applied to Conceptual Multiple-Choice Tests. *Getting Started in PER* (eds C. Henderson, K. Harper), College Park, MD: American Association of Physics Teachers, pp. 1–40.

Gracheva D.A. (2022) Analiz sopostavimosti izmereniya metapredmetnykh navykov v tsifrovoy srede [Analysis of Task Comparability in Digital Environment by the Case of Metacognitive Skills]. *Psikhologicheskaya nauka i obrazovanie / Psychological Science and Education*, vol. 27, no 6, pp. 57–67. https://doi.org/10.17759/pse.2022270605

Gracheva D.A, Tarasova K.V. (2022) Podkhody k razrabotke variantov zadaniy scenarnogo tipa v ramkakh metoda dokazatelnoy argumentatsii [Approaches to the Development of Scenario-Based Task Forms within the Framework of Evidence-Centered Design]. *Otechestvennaya i zarubezhnaya pedagogika*, vol. 1, no 3, pp. 83–97. https://doi.org/ 10.24412/2224–0772–2022–84–83–97

Haladyna T.M., Downing S.M., Rodriguez M.C. (2002) A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, vol. 15, no 3, pp. 309–333. https://doi.org/10.1207/S15324818AME1503_5

Hild P., Gut C., Brückmann M. (2019) Validating Performance Assessments: Measures That May Help to Evaluate Students' Expertise in 'Doing Science'. *Research in Science & Technological Education*, vol. 37, no 4, pp. 419–445. https://doi.org/10.1080/02635143.2018.1552851

Homayounzadeh M., Saadat M., Ahmadi A. (2019) Investigating the Effect of Source Characteristics on Task Comparability in Integrated Writing Tasks. *Assessing Writing*, vol. 41, no 2, pp. 25–46. https://doi.org/10.1016/j.asw.2019.05.003

Hooijdonk van M., Mainhard T., Kroesbergen E.H., van Tartwijk J. (2022) Examining the Assessment of Creativity with Generalizability Theory: An Analysis of Creative Problem Solving Assessment Tasks. *Thinking Skills and Creativity*, vol. 43, no 1, Article no 100994. https://doi.org/10.1016/j.tsc.2021.100994

Huebner A., Lucht M. (2019) Generalizability theory in R. *Practical Assessment, Research, and Evaluation*, vol. 24, no 5. https://doi.org/10.7275/5065-gc10

Jiang Z., Skorupski W. (2018) A Bayesian Approach to Estimating Variance Components Within a Multivariate Generalizability Theory Framework. *Behavior Research Methods*, vol. 50, no 3, pp. 2193–2214. https://doi.org/10.3758/s13428-017-0986-3

Jorgensen T.D. (2021) How to Estimate Absolute-Error Components in Structural Equation Models of Generalizability Theory. *Psych*, vol. 3, no 2, pp. 113–133. https://doi.org/10.3390/psych3020011

Keller L.A., Clauser B.E., Swanson D.B. (2010) Using Multivariate Generalizability Theory to Assess the Effect of Content Stratification on the Reliability of a Performance Assessment. *Advances in Health Sciences Education*, vol. 15, no 5, pp. 717–733. https://doi.org/10.1007/s10459-010-9233-8

Li G., Pan Y., Wang W. (2021) Using Generalizability Theory and Many-Facet Rasch Model to Evaluate In-Basket Tests for Managerial Positions. *Frontiers in Psychology*, vol. 12, July, Article no 660553. https://doi.org/10.3389/fpsyg.2021.660553

Liao R.J. (2023) The Use of Generalizability Theory in Investigating the Score Dependability of Classroom-Based L2 Reading Assessment. *Language Testing*, vol. 40, no 1, pp. 86–106. https://doi.org/10.1177/02655322211070840

Messick S. (1994) The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*, vol. 23, no 2, pp. 13–23. https://doi.org/10.3102/0013189X023002013

Mislevy R.J., Almond R.G., Lukas J.F. (2003) *A Brief Introduction to Evidence-Centered design. ETS Research Report Series no 2003(1)*. Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Rosen Y. (2017) Assessing Students in Human-to-Agent Settings to Inform Collaborative Problem-Solving Learning. *Journal of Educational Measurement*, vol. 54, no 1, pp. 36–53. https://doi.org/10.1111/jedm.12131

Ruiz-Primo M.A., Li M. (2015) The Relationship between Item Context Characteristics and Student Performance: The Case of the 2006 and 2009 PISA Science Items. *Teachers College Record*, vol. 117, no 1, pp. 1–36. https://doi.org/10.1177/016146811511700118

Shavelson R.J., Baxter G.P., Gao X. (1993). Sampling Variability of Performance Assessments. *Journal of Educational Measurement*, vol. 30, no 3, pp. 215–232. https://doi.org/10.1111/j.1745-3984.1993.tb00424.x

Shavelson R.J., Webb N.M., Rowley G.L. (1992) Generalizability Theory. *Methodological Issues & Strategies in Clinical Research* (ed. A.E. Kazdin), American Psychological Association, pp. 233–256. http://dx.doi.org/10.1037/10109-051

Uglanova I., Brun I., Vasin G. (2018) Metodologiya Evidence-Centered Design dlya izmereniya kompleksnykh psikhologicheskikh konstruktov [Evidence-Centered Design Method for Measuring Complex Psychological Constructs]. *Journal of Modern Foreign Psychology*, vol. 7, no 3, pp. 18–27. https://doi.org/10.17759/jmfp.2018070302

Uglanova I.L., Zhiltsova L.Y., Lebedeva M.Y. (2021) Izmerenie navykov kommunikatsii i kooperatsii v nachal'noy i sredney shkole: mogut li shkol'niki dogov-

orit'sya s inoplanetyaninom? [Communication and Cooperation Assessment in Primary and Middle School: How Students Negotiate with an Alien?]. Proceedings of the *5th International Conference "Informatization of Education and E-learning Methodology: Digital Technologies in Education" (Krasnoyarsk, 2022, September, 20–23),* Krasnoyarsk: Siberian Federal University, pp. 682–686.

Uglanova I., Orel E., Gracheva D., Tarasova K. (2023) Computer-Based Performance Approach for Critical Thinking Assessment in Children. *British Journal of Educational Psychology*, vol. 93, no. 2, pp. 531–544. https://doi.org/10.1111/bjep.12576

Uzun N.B., Aktas M., Asiret S., Yormaz S. (2018) Using Generalizability Theory to Assess the Score Reliability of Communication Skills of Dentistry Students. *Asian Journal of Education and Training*, vol. 4, no 2, pp. 85–90. https://doi.org/10.20448/journal.522.2018.42.85.90

Wang D., Liu H., Hau K.T. (2022) Automated and Interactive Game-Based Assessment of Critical Thinking. *Education and Information Technologies*, vol. 27, no 4, pp. 4553–4575. https://doi.org/10.1007/s10639-021-10777-9

Wu M.Y., Steinkrauss R., Lowie W. (2023) The Reliability of Single Task Assessment in Longitudinal L2 Writing Research. *Journal of Second Language Writing*, vol. 59, no 4, Article no 100950. https://doi.org/10.1016/j.jslw.2022.100950

Zhai X., Haudek K.C., Wilson C., Stuhlsatz M. (2021) A Framework of Construct-Irrelevant Variance for Contextualized Constructed Response Assessment. *Frontiers in Education*, vol. 6, October, Article no 751283. https://doi.org/10.3389/feduc.2021.751283

Декомпозиция трудности заданий в тесте читательской грамотности

Алина Иванова, Инна Антипкина

Статья поступила в редакцию в марте 2023 г.

Иванова Алина Евгеньевна — старший научный сотрудник Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики». Адрес: 101000 Москва, Потаповский пер., 16, стр. 10. E-mail: aeivanova@hse.ru. ORCID: https://orcid.org/0000-0003-3340-7651 (контактное лицо для переписки)

Антипкина Инна Вениаминовна — научный сотрудник Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики». E-mail: iantipkina@hse.ru. ORCID: https://orcid.org/0000-0003-4865-3433

Аннотация

Исследование посвящено декомпозиции трудности теста в зависимости от характеристик заданий (таких как формат, тип текста, к которому относится задание) и необходимых для ответа читательских действий (поиск информации в тексте, простые выводы, сложные выводы, критическая интерпретация текста). Выборку исследования составили учащиеся 4-х классов школ Красноярска, которые проходили компьютеризированный тест читательской грамотности «Прогресс» весной 2022 г. Исследование выполнено методом психометрического моделирования с использованием модели LLTM+e. Гипотеза исследования: декомпозиция трудностей заданий в тесте позволит показать, что необходимые для выполнения заданий читательские действия будут образовывать иерархию трудности, схожую с традиционными таксономиями, такими как таксономия Б. Блума, т.е. читательские умения, направленные на анализ, синтез, интерпретацию информации, будут придавать заданиям большую трудность, чем простые выводы, а те, в свою очередь, будут делать задания более трудными, чем читательские действия на поиск информации в тексте. Установлено, что принадлежность заданий к той или иной группе читательских умений является значимым фактором степени их трудности. Размеры эффектов не позволяют говорить о строгой иерархии, но при контроле других атрибутов задания на поиск информации в явном виде более просты для учащихся, чем задания на сложные выводы и на критическое осмысление текста.

Ключевые слова

чтение, начальная школа, тестирование, моделирование трудности заданий, LLTM

Для цитирования

Иванова А.Е., Антипкина И.В. (2023) Декомпозиция трудности заданий в тесте читательской грамотности. *Вопросы образования / Educational Studies Moscow*, № 3, сс. 92–112. https://doi.org/10.17323/vo-2023-16925

Decomposing Difficulty of Reading Literacy Test Items

Alina Ivanova, Inna Antipkina

Alina Ye. Ivanova — Senior Research Fellow at the Centre for Psychometrics and Measurement in Education, Institute of Education, National Research University Higher School of Economics. Address: Bld. 10, 16 Potapovsky Ln, 101000 Moscow, Russian Federation. E-mail: aeivanova@hse.ru. ORCID: https://orcid.org/0000-0003-3340-7651 (corresponding author)

Inna V. Antipkina — Research Fellow at the Centre for Psychometrics and Measurement in Education, Institute of Education, National Research University Higher School of Economics. E-mail: iantipkina@hse.ru. ORCID: https://orcid.org/0000-0003-4865-3433

Abstract The current study investigates the question of test difficulty decomposition depending on the characteristics of items (such as: format, belonging to the type of text to which the item belongs) and the reader's actions required to answer it (search for information in the text, simple conclusions, complex conclusions, critical interpretation of the text). The sample of the study consisted of fourth grade elementary school students in Krasnoyarsk, who completed the computerized test of reading literacy "Progress" in the spring of 2022. Research method: psychometric modeling using the LLTM+e model. Research hypothesis: the decomposition of item difficulties will help to prove that the reading actions required to complete the tasks will form a hierarchy of difficulties similar to traditional taxonomies (B. Bloom), that is, reading skills aimed at analyzing, synthesizing, interpreting information will give tasks greater difficulty than simple conclusions, and those, in turn, will make tasks more difficult than the reader's actions to find information in the text. The results show that the assignment of items to the group of reader's actions is a significant factor. The size of the effects does not allow us to speak of a strict hierarchy, but when other attributes are controlled, the tasks for information retrieval in an explicit form are easier for students than the tasks for complex conclusions and for critical understanding of the text.

Keywords reading, elementary school, testing, item difficulty modeling, LLTM

For citing Ivanova A.Ye., Antipkina I.V. (2023) Dekompozitsiya trudnosti zadaniy v teste chitatel´skoy gramotnosti [Decomposing Difficulty of Reading Literacy Test Items]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 3, pp. 92–112. https://doi.org/10.17323/vo-2023-16925

Первыми опубликованными в академических журналах тестовыми заданиями закрытого типа были задания для оценивания чтения, а именно для проверки буквального понимания прочитанного: это был *Kansas Silent Reading Test*, увидевший свет в 1916 г.

В течение XX в. теоретические подходы к оцениванию чтения значительно изменились. Долгое время важнейшим показателем учебных достижений учащихся в этой области была техника чтения, и первые психометрические модели Г. Раш создавал для обработки данных, полученных при оценивании техники чтения учащихся [Mead, 2008]. Во второй половине XX в.

фокус внимания исследователей чтения сместился на изучение понимания прочитанного. Для того чтобы систематизировать разработку заданий по чтению, создавались разнообразные таксономии читательских умений. Например, П. Пирсон и Д. Джонсон [Pearson, Johnson, 1978] делили вопросы в тестах для проверки чтения на две категории: те, ответ на которые эксплицитно представлен в тексте, и те, ответ на которые заложен в тексте имплицитно или основывается на предыдущих (фоновых) знаниях читателя. Позже эта таксономия была до некоторой степени валидизирована [Thompson, Gipe, Pitts, 1985]. У. Уоррен, Д. Николас и Т. Трабассо делят вопросы по тексту на следующие категории: вопросы, ответы на которые следуют из построения логических связей между элементами текстовой информации (ответы на вопросы почему и зачем), на построении информационных связей (ответы на вопросы кто, что, когда, где) и оценочные суждения, которые могут быть связаны с предыдущим опытом учащихся [Warren, Nicholas, Trabasso, 1979].

1. Читательские умения: простые и сложные

Наиболее влиятельные в академических кругах теоретические рамки оценивания чтения используются в международных сравнительных исследованиях PIRLS [Mullis, Martin, 2016] и PISA [OECD, 2019]. На волне интереса к результатам академических исследований читательские умения учащихся чаще всего оцениваются с точки зрения читательской грамотности — способности понимать и использовать разные формы письменной речи, чтобы учиться, участвовать в школьных и внешкольных сообществах и для удовольствия [Mullis, Martin, Sainsbury, 2016]. Масштабность инструмента оценивания навыков чтения у учащихся 4-х классов PIRLS, возможность сравнивать результаты учащихся из разных стран, разнообразие контекстных данных обусловили значительное влияние PIRLS на образовательную политику многих государств [Schwippert, Lenkeit, 2012] и на исследовательскую практику. При этом парадоксальным образом исследований на основании результатов PIRLS, посвященных собственно навыкам чтения, публикуется значительно меньше, чем исследований контекстных факторов — влияния учительских практик, роли родительского участия. Причина — в характере инструмента PIRLS, который создавался для оценки не индивидуальных, а групповых результатов [Lenkeit et al., 2015].

Теоретические рамки инструментов оценивания чтения в PIRLS и PISA близки: в них выделяются похожие группы читательских умений. Приведем группы читательских умений из теоретической рамки чтения PIRLS, актуальные для нашего исследования: 1) умение найти в тексте информацию, изложенную в явном виде; 2) умение делать простые умозаключения на ос-

нове информации, изложенной в тексте в явном виде; 3) умение интегрировать и интерпретировать идеи и информацию текста; 4) умение оценивать содержание и форму текста [Mullis, Martin, Sainsbury, 2016].

Эти группы читательских умений выстроены в логике таксономии образовательных результатов Б. Блума [Anderson et al., 2001], которая широко используется для разработки тестовых заданий и инструментов оценивания. Специфика таксономии Б. Блума состоит в том, что группы оцениваемых образовательных результатов в ней представлены в виде иерархии: предполагается, что «нижнеуровневые» образовательные результаты, например вспоминание информации, должны осваиваться учащимися первыми и при оценивании содержащие их задания должны быть легче, чем задания, направленные на образовательные результаты более высокого уровня, такие как анализ и синтез.

Ответ на вопрос, какие группы читательских умений для учащихся сложнее, а какие проще, не так очевиден, как может показаться. На материалах PIRLS установлено, что умение интегрировать и интерпретировать идеи и информацию из текста развито у российских четвероклассников значимо лучше, чем умение находить в тексте информацию, сформулированную в явном виде [Цукерман, Ковалева, Баранова, 2018]. Этот результат выглядит неожиданным, поскольку поиск информации в тексте считается базовым и более простым читательским умением по сравнению с анализом и синтезом информации. Авторы провели дополнительное исследование, поскольку сочли, что на трудность задания могут влиять не только заложенные в него проверяемые читательские умения, но и другие характеристики задания, например степень знакомства учащегося с содержанием, объем текстового фрагмента, который надо вспомнить для выбора верного ответа, особенности формулировки задания (например, задание сформулировано теми же словами, что и в тексте, или синонимично), наличие у учащегося установки на перепроверку ответа [Там же]. Авторы не называют в числе факторов формат задания, но известно, что он относится к важным предикторам трудности заданий. Например, задания с выбором ответа обычно оказываются легче заданий с конструируемым (открытым) ответом, а те, в свою очередь, проще заданий на поиск и исправление ошибки [Woodcock, Howard, Ehrich, 2020]. Внутри заданий значимыми предикторами степени трудности являются длина предложений, количество легко вычисляемых ошибочных опций в заданиях с выбором ответа, количество заложенных в задание верных опций и уровень абстрактности необходимой информации [Becker, Nekrasova-Beker, 2018].

Цель представленного исследования состоит в проведении декомпозиции трудности заданий в тестах чтения, с тем чтобы статистически оценить вклад потенциальных факторов трудности заданий.

Гипотеза исследования: при контроле «внешних» факторов трудности заданий, таких как формат и принадлежность задания к определенному типу, параметры трудности, связанные с заложенными в задании группами читательских умений, будут образовывать иерархию трудности — от поиска информации, данной в явном виде (наиболее простые задания), до оценивания текста в целом (наиболее трудные задания).

2. Методология

2.1. Инструмент исследования

Русскоязычные инструменты оценивания читательской грамотности как метапредметного результата разрабатывались в связи с введением ФГОС ООО и НОО [Гостева и др., 2019]. Использованный в этом исследовании тест читательской грамотности «Прогресс» создан Центром психометрики и измерений в образовании НИУ ВШЭ как независимый мониторинговый инструмент [Бакай, Юсупова, Антипкина, 2023]. В спецификации к тесту описаны четыре группы читательских навыков. Эта теоретическая рамка схожа с рамкой PIRLS — а следовательно, допустимо сравнение результатов, полученных с помощью теста «Прогресс», с полученными при использовании других инструментов, созданных по рамке PIRLS.

Первую группу читательских умений составляет «поиск информации, представленной в явном виде». Для их тестирования используются задания, в которых учащиеся ищут информацию в тексте или полагаются на память. При этом им не требуется совершать дополнительные когнитивные действия, например, интерпретировать формулировку задания как синонимичную по отношению к формулировке информации в тексте.

Вторая группа читательских умений — «простой вывод». В заданиях на простой вывод учащимся необходимо совершить одно дополнительное когнитивное действие. Например, нужно увидеть, что формулировка задания синонимична по отношению к информации в тексте: в таком случае простой вывод заключается в обработке синонимов. Или «перешагнуть» через знак препинания, объединив информацию из двух разных предложений, не связанных союзом «потому что»: простой вывод состоит в интерпретации их как причины и следствия.

Третья группа умений — «сложные выводы». Для их осуществления требуются более сложные действия по обработке информации в два и более когнитивных действия с использованием навыков обобщения, сопоставления, интерпретации.

Четвертая группа — «умение оценивать содержание и форму текста» — требует владения метанавыком оценки текста как произведения, включая анализ использованных автором художественных средств и приемов, обобщение замысла автора, т.е. учащийся должен критически осмыслить текст.

Тест читательской грамотности «Прогресс» состоит из двух частей: одна с информационным стимульным текстом, другая — с художественным. Две части инструмента различаются не только типами текстов, но и структурой. В той части теста, где стимульным материалом служит художественный текст, а именно фантастический рассказ, все 23 тестовых вопроса расположены после текста. Информационных текстов в тесте три, они предъявляются по очереди, и после каждого фрагмента учащийся отвечает на вопросы по данному тексту, а после всех текстов он получает несколько вопросов, относящихся к ним всем (эти группы заданий по фрагментам в дальнейшем анализе фигурируют как кластеры заданий). Всего в тесте с информационными текстами 17 вопросов. Сложная структура информационных текстов имитирует гипертекст, с которым неизбежно сталкиваются школьники в интернете [Мелентьева, 2015]. И к художественному, и к информационному стимульному материалу предлагаются проверочные задания пяти разных форматов: 1) выбор одного верного ответа из четырех предложенных; 2) выбор нескольких верных ответов из нескольких предложенных; 3) ряд утверждений, на каждое из которых нужно дать ответ «верно» или «неверно»; 4) задания на поиск пары с «перетаскиванием» ответа; 5) открытые задания со свободно конструируемым ответом. Все типы заданий, кроме заданий с выбором одного ответа из четырех, создавались как политомические. Однако в интересах простоты применения и интерпретации результатов для дальнейшего анализа мы выбрали линейную логистическую тестовую модель LLTM+e, которая требует дихотомических данных. Поэтому политомические вопросы преобразованы в дихотомические следующим образом: каждая опция политомического задания оценивалась как 1 или как 0 в зависимости от того, была она верно выбрана или верно проигнорирована (в случае с дистракторами). Такой дизайн позволил очень точно соотнести каждую опцию заданий с группами читательских умений. Например, в задании: «Верны ли приведенные ниже утверждения? Отметь “верно” или “НЕверно” в каждой строчке: А) Женька забыл номер квартиры главного героя; Б) Петька Грозный не был храбрым; В) Мама не увидела ничего плохого в стрижке Бобрика; Д) На рыбалке дети поймали большую щуку; Е) Павлик решил заниматься спортом, посмотрев передачу по телевизору». Верные опции Б, Д, Е. В политомическом варианте используется способ подсчета баллов

со штрафами по формуле: количество выбранных верных опций минус количество отмеченных неверных опций, в случае отрицательного числа балл обнуляется. В этом задании созданы пять дихотомических опций: за выбор опций Б, Д, Е дети получали 1 балл, за невыбор — ноль; за невыбор опций А и В учащиеся получали 1 балл, за отмечание — ноль. В вопросах с выбором одного верного варианта ответа из четырех дистракторы (ошибочные варианты) не выделялись в отдельные псевдозадания, как это было сделано в политомических вопросах. Поэтому мы ожидаем увидеть искусственно завышенную трудность дихотомических вопросов по сравнению с политомическими.

2.2. Выборка

Выборку составили 2188 учащихся 4-х классов школ сибирского города-миллионника. Оценивание состоялась весной 2022 г. Учащиеся выполняли компьютеризированный тест чтения частями, в разные дни, в течение одного урока, проходившего в компьютерном классе. Выборка является репрезентативной по району города и типу школ.

2.3. Метод анализа

В анализе данных, полученных в ходе тестирования, использован экспланаторный подход современной теории тестирования, в частности линейная логистическая тестовая модель (*linear logistic test model*, LLTM) [Fischer, 1973]. Модели этой категории позволяют моделировать и параметризовать различные процессы и характеристики заданий, в том числе относящиеся к коллатеральной информации. Коллатеральной называют побочную информацию об измерениях, использование которой в психометрическом моделировании не меняет интерпретацию оценок, но уменьшает ошибку измерения [Whitely, 1983]. К числу таких характеристик относятся, например, когнитивные операции или другие атрибуты, которые лежат в основе выполнения тестового задания, уровни таксономии, форматы заданий.

В экспланаторной парадигме генерализованных линейных смешанных моделей (*generalized linear mixed models*, GLMM) модель Раша считается описательной, поскольку в ней каждое задание описывается с помощью одного параметра трудности и одного параметра латентной способности испытуемого [De Boeck, Wilson, 2004]. А когда предикторы, объясняющие индивидуальные эффекты, например эффекты заданий, инкорпорируются в модель Раша, новая модель (в данном случае LLTM) становится экспланаторной моделью современной теории тестирования.

При этом модель Раша тоже может быть определена как генерализованная линейная смешанная модель, где зависимая

переменная (дихотомические ответы на задания) предсказывается некоторыми фиксированными эффектами (заданиями) наряду со случайными эффектами (латентной способностью испытуемого).

В литературе, посвященной измерениям, дихотомическая модель Раша часто представлена следующим образом:

$$P_{ni} = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}, \quad (1)$$

где P_{ni} — вероятность успешного выполнения учеником n задания i; θ_n — латентная способность ученика n; δ_i — трудность задания i (i = 1, 2, ..., k).

Модели Раша и LLTM могут рассматриваться как вложенные. Важной характеристикой LLTM является ее способность декомпозировать трудность задания и таким образом предоставить исследователям дополнительную информацию об эффекте компонентов (атрибутов), формирующих эту трудность:

$$\delta i = \sum_{j=1}^{m} q_{ij} \eta_j, \quad (2)$$

где δ_i — параметр трудности задания i; q_{ij} — вес атрибута в задании i; n_j — оцениваемая трудность атрибута. Чаще всего вес атрибута фиксируется дихотомически: присутствует этот атрибут в задании или нет.

Линейная логистическая тестовая модель использует заданные атрибуты задания как предикторы для объяснения вариации между заданиями относительно их влияния на вероятность верного решения этого задания. В LLTM меньше параметров трудности, чем в модели Раша, которая оценивает для каждого задания множество уникальных параметров трудности. Дополнительно в модель может быть включен ошибочный компонент. Такую модель часто называют LLTM с ошибкой. В отличие от обычной LLTM модель с ошибкой (LLTM+e) учитывает возможность неточности в предсказании, а потому является более гибкой в практическом применении [Desjardins, Bulut, 2018]. Математическая формулировка LLTM с ошибкой такова:

$$P_{ni} = \frac{\exp(\theta_n - (\sum_{j=1}^{m} q_{ij} n_j + \varepsilon_i))}{1 + \exp(\theta_n - (\sum_{j=1}^{m} q_{ij} n_j + \varepsilon_i))}, \quad (3)$$

где P_{ni} — вероятность успешного выполнения учеником n задания i; θ_n — латентная способность ученика n; q_{ij} — вес атрибута в задании i; n_j — оцениваемая трудность атрибута; ε_i — ошибочный компонент, не объясняемый атрибутами задания, распределенный нормально со средним, равным нулю и оцениваемой дисперсией. Именно эту модель мы используем в дальнейшем анализе.

2.4. *Q*-матрица

Центральный компонент модели LLTM+e — *Q*-матрица, фиксирующая операции (атрибуты), заложенные в задания. Каждая строка в ней соответствует заданию, а каждый столбец — операции (атрибуту). Именно *Q*-матрица определяет, какой атрибут вносит вклад в каждое задание [Effatpanah, Baghaei, 2021]. В тесте, в который заложено *M* атрибутов, каждое из I заданий требует реализации некоторого набора этих атрибутов, чтобы тестируемый смог ответить на него верно. Задания и атрибуты складываются в матрицу $Q = \{q_{mi}\}$, где $m = 1, ..., M$ и $i = 1, ..., I$. Матрица показывает, требует ли *i*-е задание реализации *m*-го атрибута (выполнения *m*-й операции). Цифра 1 в *Q*-матрице означает, что конкретный атрибут необходим для выполнения соответствующего задания, в то время как 0 означает, что атрибут не требуется.

Корректная идентификация операций или атрибутов и их связи с заданиями теста улучшает качество информации, получаемой при моделировании. В данной статье в качестве атрибутов использованы группы читательских умений, форматы заданий и принадлежность текста к информационному или художественному типу. Для определения соответствия определенного задания той или иной группе читательских умений привлекались три эксперта: учитель начальных классов, эксперт в области читательской грамотности, филолог с опытом разработки тестов читательской грамотности. Перед выполнением кодирования и формирования *Q*-матрицы эксперты обсудили способы интерпретации и кодирования атрибутов для теста. Другие атрибуты (принадлежность задания к конкретному тексту, к конкретному формату) определялись уже без участия экспертов, поскольку являются внешними, объективными признаками заданий. В *дополнительных материалах* к статье представлена *Q*-матрица, которая содержит 14 атрибутов и 93 задания.

3. Результаты анализа

Пакет *eRm* [Mair et al., 2020] в статистическом программном обеспечении *R* (версия 4.2.1) использовался для предварительного анализа данных с помощью дихотомической модели Раша; пакет *lme4* [De Boeck et al., 2011] применялся для оценки параметров моделей Раша и LLTM в парадигме генерализованных линейных смешанных моделей.

При оценке LLTM+e необходимо, чтобы имеющиеся данные подходили дихотомической модели Раша. Если данные не согласуются с моделью Раша, не имеет смысла декомпозировать трудность задания, поскольку сам параметр трудности каждого задания и оценка тестируемого не будут иметь практически значимой интерпретации [Fischer, 2005]. Чтобы проверить со-

гласие данных с базовой моделью Раша, исследованы статистики согласия с моделью и проведен общий тест Мартин-Лёфа [Verguts, De Boeck, 2000].

Статистики согласия представляют собой среднеквадратичные отклонения эмпирических значений (наблюдаемого балла за задание) от ожидаемых моделью для каждого задания, взвешенные (*infit MNSQ*) и невзвешенные (*outfit MNSQ*). Значения статистик согласия должны находиться в пределах рекомендуемых специалистами значений (0,7; 1,3) [Linacre, 2004]. Аналогичным образом оценены параметры выборки учащихся (для оценки согласия параметров испытуемых использованы более мягкие критерии — 0,5; 1,5). Первичный анализ выявил, что 8 заданий не согласуются с моделью. Они были удалены, а данные рекалиброваны. В дальнейшем анализе использованы ответы 2136 тестируемых (~98% выборки) на 93 задания.

В табл. 1 приведены показатели согласия итогового набора данных модели Раша. Общая надежность теста (*separation reliability*), показывающая воспроизводимость иерархии оценок испытуемых [Wright, 1996], составила 0,9. Средняя стандартная ошибка измерения — 0,27 логита.

Таблица 1. **Статистики согласия для 93 заданий теста**

Задание	Outfit MNSQ	Infit MNSQ	Задание	Outfit MNSQ	Infit MNSQ	Задание	Outfit MNSQ	Infit MNSQ
df_1	0,97	0,97	df_13	0,95	0,97	p_6_5	1,20	1,16
df_2_3	0,69	0,92	df_14	1,06	0,99	p_7	0,94	0,95
df_2_4	0,61	0,92	df_16	0,94	0,95	p_8_1	1,18	1,11
df_2_5	0,97	0,96	df_17_1	0,85	0,98	p_8_2	1,16	1,13
df_2_6	0,76	0,92	df_17_3	0,93	1,01	p_8_3	0,88	0,95
df_3_2	1,03	1,05	df_07_4	0,82	1,00	p_8_4	0,98	1,01
df_3_3	0,87	0,94	df_17_5	0,81	0,86	p_8_5	1,19	1,10
df_3_4	1,17	1,14	df_17_6	0,90	0,92	p_8_6	1,32	1,24
df_5	0,94	0,95	df_18	1,01	0,98	p_9	0,89	0,91
df_6_1	0,56	0,91	df_19	1,08	1,05	p_10_11	0,78	0,87
df_6_2	0,86	0,97	df_21	1,04	1,03	p_10_12	0,92	0,94
df_6_3	0,99	1,00	df_22	0,75	0,89	p_10_21	0,91	0,91
df_6_4	0,75	0,84	p_1_1	0,78	0,89	p_10_22	0,88	0,90
df_6_5	1,22	1,06	p_1_2	1,31	1,08	p_10_31	0,77	0,82
df_6_6	1,08	1,05	p_1_3	1,23	1,09	p_10_32	0,95	0,96
df_7	0,93	0,97	p_1_4	1,29	1,05	p_10_41	0,79	0,84
df_8_1	0,93	0,96	p_1_5	1,03	1,05	p_10_42	0,90	0,92

Задание	Outfit MNSQ	Infit MNSQ	Задание	Outfit MNSQ	Infit MNSQ	Задание	Outfit MNSQ	Infit MNSQ
df_8_2	1,04	1,04	p_1_6	1,15	1,10	p_11_2	1,01	0,99
df_8_3	1,04	1,03	p_2	0,93	0,95	p_11_3	1,02	0,98
df_8_4	0,58	0,89	p_3_1	0,96	0,98	p_11_6	0,89	0,92
df_8_5	0,85	0,94	p_3_3	1,24	1,11	p_12	1,06	1,03
df_9	1,02	1,01	p_3_4	0,87	0,93	p_13	1,31	1,25
df_10_1	0,56	0,92	p_4	1,06	1,05	p_14	1,20	1,08
df_10_2	1,21	1,17	p_5_1	1,05	1,06	p_15_1	0,99	1,02
df_10_4	1,00	1,01	p_5_2	0,83	0,89	p_15_2	1,03	1,00
df_11	0,71	0,94	p_5_3	1,00	1,01	p_15_3	1,32	1,22
df_12_1	0,87	0,89	p_5_4	0,87	0,98	p_15_4	1,18	1,13
df_12_2	0,78	0,95	p_5_5	0,72	0,82	p_15_5	0,76	0,90
df_12_3	1,26	1,10	p_6_1	0,66	0,94	p_15_7	1,14	1,06
df_12_4	0,94	0,99	p_6_2	1,01	1,00	p_16	0,85	0,91
df_12_5	1,16	1,13	p_6_4	0,96	0,97	p_17	0,99	1,00

Как видно из табл. 1, задания теста в целом хорошо согласуются с выбранной моделью измерения. Невзвешенные статистики согласия лежат в диапазоне от 0,56 до 1,32, взвешенные — в диапазоне 0,82–1,25.

Для оценки глобального соответствия данных теста модели Раша выполнен статистический тест Мартин-Лёфа [Douglas, 1982; Verguts, De Boeck, 2000]. В этом тесте среднее и, дополнительно, медиана сырых баллов позволяют разделить задания на две группы и проверить допущение, что обе группы формируют одну Раш-размерность. Результаты теста свидетельствуют о глобальном соответствии данных модели Раша (MLöef = 915,889, *df* = 2155, *p* = 0,99 (среднее), MLöef = 900,597, *df* = 2161, *p* = 0,99 (медиана)).

На следующем этапе проведен анализ данных на базе модели LLTM+e с использованием *Q*-матрицы, включающей 93 задания и заложенные в них 14 атрибутов. В табл. 2 приведены результаты анализа с помощью серии моделей LLTM+e, которые содержат показатели трудностей параметров используемых нами атрибутов, их стандартные ошибки и их статистическую значимость.

В табл. 3 приведены результаты анализа качества и сравнение моделей. В дополнительных материалах также приведены параметры трудности заданий и ошибка измерения для каждого задания на базе дихотомической модели Раша и итоговой модели LLTM+e. Атрибуты с положительными параметрами трудности делают задание легче, атрибуты с отрицательными параметрами — сложнее.

Таблица 2. **Серия LLTM-моделей. Оценка эффектов параметров трудности**

Параметры	Модель Раша			Модель LLTM 1			Модель LLTM 2			Модель LLTM 3			Модель LLTM 4		
	Оценка	Стандартная ошибка	Значимость*	Оценка	Стандартная ошибка	Значимость	Оценка	Стандартная ошибка	Значимость	Оценка	Стандартная ошибка	Значимость	Оценка	Стандартная ошибка	Значимость
Фиксированные эффекты															
Формат: задание с выбором нескольких верных ответов				1,51	0,16	0,00				0,76	0,22	0,00	0,36	0,24	0,13
Формат: задание из нескольких утверждений с ответом «да/нет»				1,43	0,18	0,00				0,61	0,25	0,01	0,21	0,24	0,38
Формат: задание с перетаскиванием пар				0,10	0,28	0,73				–0,64	0,32	0,05	–0,42	0,37	0,26
Формат: открытое задание				0,53	0,41	0,20				–0,24	0,36	0,50	–0,43	0,45	0,34
Формат: задание с выбором одного варианта ответа				0,37	0,21	0,08				–0,39	0,26	0,14	–0,71	0,26	0,01
Группа умений 1: поиск информации в явном виде							1,26	0,27	0,00	0,96	0,28	0,00	1,45	0,30	0,00
Группа умений 2: простые выводы							0,93	0,16	0,00	0,74	0,21	0,00	1,23	0,23	0,00
Группа умений 3: сложные выводы							1,16	0,15	0,00	0,78	0,21	0,00	1,36	0,24	0,00
Группа умений 4: оценка всего текста							0,47	0,33	0,15	0,60	0,36	0,09	0,96	0,36	0,01
Информационный текст 1													–0,25	0,22	0,26
Информационный текст 2													–0,69	0,30	0,02
Информационный текст 3													–0,70	0,29	0,02
Задание относится ко всем трем информационным текстам													–0,19	0,36	0,59
Случайные эффекты															
Ученики (вариация в оценках)	0,73			0,73			0,73			0,73			0,73		
Задания (вариация в оценках)				0,82			1,01			0,75			0,71		

* Значимость в колонках таблицы для каждой модели отражает статистическую значимость эффекта каждого атрибута.

Таблица 3. **Критерии сравнения моделей**

Модель	Количество параметров	AIC	BIC	logLik	deviance	LR Тест
Модель Раша	94	206328,7	207289,7	–103070,3	206140,7	—
Модель LLTM 1	6	206790,1	206851,4	–103389,1	206778,1	$p < 0,01$
Модель LLTM 2	7	206772,7	206844,3	–103379,4	206758,7	$p < 0,01$
Модель LLTM 3	11	206772,5	206885,0	–103375,3	206750,5	$p < 0,01$
Модель LLTM 4	15	206775,2	206928,6	–103372,6	206745,2	$p < 0,01$

Нулевой моделью в нашем анализе является модель Раша.

В первой модели в качестве влияющих на трудность оценены эффекты форматов заданий теста. Статистически значимыми оказались эффекты двух форматов из пяти. В целом искусственно дихотомизированные задания и задания с выбором нескольких верных ответов оказались для учащихся легче. Это легкость искусственного происхождения, поскольку при дихотомизации политомических заданий каждая ответная опция могла быть либо выбрана, либо не выбрана, т.е. вероятность верного ответа случайным образом составляла 0,5, в то время как в изначально дихотомических заданиях выбор производился не из двух, а из четырех ответных опций.

Во второй модели мы отдельно оценили вклад в трудность задания параметров читательских действий согласно теоретической рамке инструмента. Задания первой группы читательских умений наиболее легкие, задания четвертой группы наиболее трудные. При этом четкой иерархии между уровнями нет, поскольку четвертая группа читательских умений не является значимо более трудной, а трудности параметров для второго и третьего уровня не упорядочены.

В третьей модели мы объединили параметры уровней и форматов. Интерпретация всех параметров не изменилась, как и значимость их эффектов. При этом, как можно заметить в табл. 2, разброс в оцениваемых эффектах между группами умений стал меньше.

Наконец, в четвертой модели мы дополнительно проанализировали эффекты художественной (референтная категория) и информационной частей теста и четырех кластеров заданий. Объединенная модель показала, что при учете всех параметров значимо труднее задание делает формат с выбором одного верного ответа из нескольких предложенных (как мы отмечали выше, эта трудность искусственная, вызванная тем, что остальные задания были дихотомизированы с вероятностью угадать верный ответ, равной 0,5). Все группы читательских действий вносят значимый эффект: задания первого уровня по сравне-

нию с остальными уровнями значимо легче, четвертого — значимо труднее. Однако сами по себе группы читательских умений не делают задания сложнее для учащихся, в отличие от совокупности других параметров: формата выбора ответа и принадлежности задания ко второму и третьему текстам в субтесте на информационное чтение. Другими словами, трудность заданий для учащихся связана прежде всего с характеристиками текста, к которому они относятся, и с форматами предъявления, в то время как заложенные в задание читательские умения, хотя и отличаются друг от друга, не делают задания именно трудными.

Как правило, в анализе модель Раша и LLTM сравнивают с помощью теста отношения правдоподобия (*LR test*) [Effatpanah, Baghaei, 2021]. Эти модели считаются вложенными, и это позволяет нам оценить возможность того, что модель с меньшим числом параметров подходит не хуже, чем более параметризированная модель; разница показателей –2*log-likelihood* (*deviance*) для моделей Раш и LLTM+e имеет приближенное распределение хи-квадрат с количеством степеней свободы, равным разнице между количеством параметров в двух моделях [Fischer, 1973]. Для сравнения использовались также индексы согласия — информационный критерий Акайке (AIC) и байесовский информационный критерий (BIC) [Burnham, Anderson, 2002]. Предполагается, что хорошо специфицированная модель LLTM может подходить данным не хуже, чем модель Раша. Однако нам неизвестны работы, в которых исследователям или разработчикам теста удалось бы этого достичь. Обычно модель LLTM не подходит данным так же хорошо, как модель Раша, — а значит, указанные исследователями когнитивные операции или атрибуты недостаточно хорошо объясняют параметры трудности заданий. И в нашем случае также модель Раша лучше объясняет трудность задания при сравнении показателей AIC, а также по результатам теста правдоподобия. В табл. 3 представлено сравнение всех использованных нами моделей, в последней колонке приведены результаты теста отношения правдоподобия для всех моделей LLTM в сравнении с моделью Раша: ни одна из моделей с меньшим числом параметров не является статистически значимо лучшей по сравнению с моделью Раша. Из моделей LLTM+e наименее подходящей оказалась модель 1, наиболее подходящей — модель 4.

Кроме того, мы оценили связь показателей заданий и испытуемых в модели Раша и наиболее подходящей модели LLTM+e. Корреляция между оцениваемыми трудностями заданий в модели Раша (см. табл. 2 в дополнительных материалах) и трудностями заданий в модели LLTM+e 4 составила 0,82 ($p < 0{,}001$). То есть параметры четвертой модели LLTM+e объясняют 67,2% предполагаемых трудностей заданий, объясняемых

моделью Раша. Корреляция оценок учащихся по тесту, полученных с помощью двух моделей, — 0,99 ($p < 0,001$).

4. Обсуждение результатов

Исследование посвящено декомпозиции трудности заданий в связи с различными их характеристиками (атрибутами). Наиболее важный результат состоит в том, что принадлежность задания к одной из четырех групп читательских умений стабильно остается значимым параметром при учете других атрибутов. Размеры эффектов не дают оснований утверждать, что четыре группы читательских умений образуют иерархию трудности, хотя задания на осмысление всего текста являются относительно более трудными, чем задания трех остальных групп. Однако атрибут принадлежности задания к любой группе читательских умений не делает задания труднее для учащихся — за собственно сложность задания отвечают скорее такие атрибуты, как текст, к которому относится задание (задания к информационному тексту сложнее относящихся к художественному) и формат задания. Полученные нами результаты, согласно которым художественные тексты для детей легче, чем информационные, согласуются с данными исследований, проводившихся в рамках той же методологии. Так, на выборке 8-классников показано, что читатели лучше выполняют задания, когда они знакомы с темой текста и заинтересованы в ней, когда тексты нарративные и когда задания основаны на тексте и время на их выполнение не ограничено [Rahman, Alexander, Chae, 2022].

Продолжая рассмотренное выше исследование [Цукерман, Ковалева, Баранова, 2018], мы на материале тестов, разработанных на теоретической основе, схожей с PIRLS, показали, что сами по себе группы читательских умений нельзя рассматривать как надежные факторы трудности заданий. Нецелесообразно также выстраивать список читательских умений в иерархию трудности и ожидать, что задания на поиск в тексте информации в явном виде обязательно будут значимо легче, чем задания на анализ и синтез информации из текста, в которых для получения верного ответа необходимо совершить несколько когнитивных действий. Отнесенность задания к той или иной группе читательских умений действительно является значимым фактором при выполнении заданий, но она не позволяет уверенно устанавливать желаемую трудность заданий. С точки зрения практики, например, при разработке заданий на чтение опора на четыре группы читательских умений позволяет только сбалансировать содержание инструмента оценивания, но, чтобы более надежно достичь желаемой трудности заданий, нужно работать с собственно стимульными текстами и форматами заданий.

Мы использовали компьютеризированные инструменты оценки, а в исследовании Г.А. Цукерман, Г.С. Ковалевой и В.Ю. Барановой анализировались результаты PIRLS, полученные в бланковом тестировании. Бланковая и цифровая формы чтения существенно различаются [Støle, Mangen, Schwippert, 2020; Delgado et al., 2018]. Учащиеся также могут по-разному реализовывать свои читательские умения в разных формах оценивания. Поэтому наши результаты требуют дальнейшего уточнения в исследованиях декомпозиции трудности тестовых заданий в компьютеризированной и бланковой форме оценивания.

В данном исследовании мы использовали экспланаторный подход к моделированию на базе модели LLTM+e. Базовая модель LLTM имела существенные ограничения, в частности, она не позволяла учесть необъяснимую дисперсию в оценке параметров трудности заданий. Однако в начале 2000-х годов группа исследователей показала, как методы и технологии, разработанные для оценки многоуровневых моделей, могут быть использованы для оценки LLTM и ее расширений [De Boeck, Wilson, 2004; Lang, Tay, 2021]. Сегодня психометрики применяют объяснительный подход к моделированию результатов оценки в современной теории тестирования, который позволяет анализировать успешность выполнения тестовых заданий с учетом параметров заданий или учеников, а также с учетом дополнительного ошибочного компонента и других нюансов. В частности, использованная нами модель LLTM+e учитывает оставшуюся вариацию в трудности заданий после добавления в модель отдельных заложенных при разработке теста характеристик заданий.

Ограничением данного исследования является отсутствие учета в его дизайне таких важных для результатов оценивания факторов, как мотивация учащихся и их обученность стратегиям чтения и стратегиям выполнения тестов (включая установку на перепроверку ответов). Мы считаем перспективными дальнейшие исследования, в которых помимо атрибутов заданий будут учитываться атрибуты респондентов.

5. Заключение

Данное исследование проведено на достаточно большой выборке учащихся 4-х классов, которые выполняли компьютеризированный тест читательской грамотности «Прогресс» весной 2022 г. Целью исследования было применить экспланаторный подход к анализу тестовых данных, для того чтобы учесть отдельные заложенные на этапе разработки теста характеристики заданий и определить, какие из них оказывают значимый эффект на трудность заданий. Мы предполагали, что необходимые для выполнения заданий читательские действия будут

образовывать иерархию трудности, схожую с традиционными таксономиями, такими как таксономия Б. Блума, т.е. задания, требующие читательских умений анализировать, синтезировать, интерпретировать информацию, будут для учащихся труднее, чем задания, предполагающие простые выводы, а те, в свою очередь, будут труднее, чем задания, требующие поиска информации в тексте.

Проведенный анализ показал, что заложенная нами в аналитическом подходе *Q*-матрица теста не позволяет полностью объяснить трудность заданий теста. Тем не менее использование LLTM+e может дать полезную информацию разработчикам теста, а также учителям и исследователям. Интерпретация значимости и эффекта трудности или легкости параметра позволяет объяснить, почему дети реагируют на задания теста определенным образом.

Так, принадлежность заданий к той или иной группе читательских умений в общем и целом является значимым фактором степени трудности этих заданий. Размеры и направленность эффектов не позволяют говорить о строгой иерархии читательских умений, но при контроле других атрибутов задания на поиск информации в явном виде более просты для учащихся, чем задания на сложные выводы и на критическое осмысление текста.

Благодарности

Статья подготовлена в рамках гранта, предоставленного Министерством науки и высшего образования Российской Федерации (соглашение о предоставлении гранта № 075-15-2022-325 от 25.04.2022).

Дополнительные материалы к статье можно найти по ссылке: https://vo.hse.ru/article/view/16925/16280.

Литература

1. Бакай Е.А., Юсупова Э.М., Антипкина И.В. (2023) Читают или делают вид? Анализ поведения учащихся начальных классов при выполнении заданий теста читательской грамотности. *Вопросы образования / Educational Studies Moscow*, № 1, сс. 8–28. https://doi.org/10.17323/1814-9545-2023-1-8-28
2. Гостева Ю.Н., Кузнецова М.И., Рябинина Л.А., Сидорова Г.А., Чабан Т.Ю. (2019) Теория и практика оценивания читательской грамотности как компонента функциональной грамотности. *Отечественная и зарубежная педагогика*, т. 1, № 4 (61), сс. 34–57.
3. Мелентьева Ю.П. (2015) *Общая теория чтения*. М.: Наука.
4. Цукерман Г.А., Ковалева Г.С., Баранова В.Ю. (2018) Читательские умения российских четвероклассников: уроки PIRLS-2016. *Вопросы образования / Educational Studies Moscow*, № 1, сс. 58–78. https://doi.org/10.17323/1814-9545-2018-1-58-78

5. Anderson L.W., Krathwohl D.R., Airasian P.W., Cruikshank K.A., Mayer R., Pintrich P.R., Raths J., Wittrock M.C. (eds) (2001) *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York, NY: Longman.
6. Becker A., Nekrasova-Beker T. (2018) Investigating the Effect of Different Selected-Response Item Formats for Reading Comprehension. *Educational Assessment*, vol. 23, no 4, pp. 296–317. http://dx.doi.org/10.1080/10627197.2018.1517023
7. Burnham K., Anderson D. (eds) (2002) *Model Selection and Multi-Model Inference. A Practical Information-Theoretic Approach*. New York; Berlin; Heidelberg: Springer.
8. De Boeck P., Bakker M., Zwitser R., Nivard M., Hofman A., Tuerlinckx F., Partchev I. (2011) The Estimation of Item Response Models with the lmer Function from the lme4 Package in R. *Journal of Statistical Software*, vol. 39, iss. 12. http://dx.doi.org/10.18637/jss.v039.i12
9. De Boeck P., Wilson M. (eds) (2004) *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach.* New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4757-3990-9
10. Delgado P., Vargas C., Ackerman R., Salmerón L. (2018) Don't Throw Away Your Printed Books: A Meta-Analysis on the Effects of Reading Media on Reading Comprehension. *Educational Research Review*, vol. 25, November, pp. 23–38. http://dx.doi.org/10.1016/j.edurev.2018.09.003
11. Desjardins C.D., Bulut O. (2018) *Handbook of Educational Measurement and Psychometrics Using R.* Boca Raton, FL: CRC. http://dx.doi.org/10.1201/b20498
12. Douglas G. (1982) Issues in the Fit of Data to Psychometric Models. *Education Research and Perspectives*, vol. 9, no 1, pp. 32–43.
13. Effatpanah F., Baghaei P. (2021) Cognitive Components of Writing in a Second Language: An Analysis with the Linear Logistic Test Model. *Psychological Test and Assessment Modeling*, vol. 63, no 1, pp. 13–44.
14. Fischer G.H. (2005) Linear Logistic Test Models. *Encyclopedia of Social Measurement* (ed. K. Kempf-Leonard), Boston; London: Elsevier, pp. 505–514.
15. Fischer G.H. (1973) The Linear Logistic Test Model as an Instrument in Educational Research. *Acta Psychologica*, vol. 37, no 6, pp. 359–374. https://doi.org/10.1016/0001-6918(73)90003-6
16. Lang J.W., Tay L. (2021) The Science and Practice of Item Response Theory in Organizations. *Annual Review of Organizational Psychology and Organizational Behavior*, vol. 8, pp. 311–338. http://dx.doi.org/10.1146/annurev-orgpsych-012420-061705
17. Lenkeit J., Chan J., Hopfenbeck T.N., Baird J.A. (2015) A Review of the Representation of PIRLS Related Research in Scientific Journals. *Educational Research Review*, vol. 16, October, pp. 102–115. http://dx.doi.org/10.1016/j.edurev.2015.10.002
18. Linacre J.M. (2004) Rasch Model Estimation: Further Topics. *Journal of Applied Measurement*, vol. 5, no 1, pp. 95–110.
19. Mair P., Hatzinger R., Maier M.J., Rusch T., Mair M.P. (2020) *ERm: Extended Rasch Modeling. 1.0-2*. Available at: https://cran.r-project.org/package=eRm (accessed 17 September 2023).
20. Mead R. (2008) *A Rasch Primer: The Measurement Theory of Georg Rasch. Psychometrics Services Research Memorandum 2008–001*. Maple Grove, MN: Data Recognition Corporation. Available at: http://www.edmeasurement.net/8226/Mead-2008-Rasch-primer.pdf (accessed 13 September 2023).
21. Mullis I.V., Martin M.O., Sainsbury M. (2016) *PIRLS 2016 Reading Framework*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, pp. 11–29.
22. OECD (2019) PISA 2018 Assessment and Analytical Framework. Paris: OECD. https://doi.org/10.1787/b25efab8-en

23. Pearson P.D., Johnson D.D. (1978) *Teaching Reading Comprehension*. New York, NY: Holt, Winehart and Winston.
24. Rahman T., Alexander P.A., Chae S.E. (2022) Reader Attributes, Task Attributes, and Reading Comprehension Proficiency: The Relation Revealed by Two Analytic Approaches. *Reading Psychology*, vol. 43, no 7, pp. 495–522. http://dx.doi.org/10.1080/02702711.2022.2126044
25. Støle H., Mangen A., Schwippert K. (2020) Assessing Children's Reading Comprehension on Paper and Screen: A Mode-Effect Study. *Computers & Education*, vol. 151, March, Article no 103861. http://dx.doi.org/10.1016/j.compedu.2020.103861
26. Schwippert K., Lenkeit J. (eds) (2012) *Progress in Reading Literacy in National and International Context. The Impact of PIRLS 2006 in 12 Countries*. Münster: Waxmann Verlag.
27. Thompson B., Gipe J.P., Pitts M.M. (1985) Validity of the Pearson-Johnson Taxonomy of Comprehension Questions. *Reading Psychology: An International Quarterly*, vol. 6, no 1–2, pp. 43–49. https://doi.org/10.1080/0270271850060105
28. Verguts T., De Boeck P. (2000) A Note on the Martin-Löf Test for Unidimensionality. *Methods of Psychological Research*, vol. 5, no 1, pp. 77–82.
29. Warren W., Nicholas D., Trabasso T. (1979) Event Chaince and Inferences in Understanding Narratives. *New Directions in Discourse Processing, vol. 2. Advances in Discourse Processing* (ed. R.O. Freedle), Norwood, NJ: Ablex Publication Corporation. https://doi.org/10.1017/S002222670000685X
30. Whitely S.E. (1983) Construct Validity: Construct Representation Versus Nomothetic Span. *Psychological Bulletin*, vol. 93, no 1, pp. 179–197. https://doi.org/10.1037/0033-2909.93.1.179
31. Woodcock S., Howard S.J., Ehrich J. (2020) A Within-Subject Experiment of Item Format Effects on Early Primary Students' Language, Reading, and Numeracy Assessment Results. *School Psychology*, vol. 35, no 1, pp. 80–87. http://dx.doi.org/10.1037/spq0000340
32. Wright B.D. (1996) Reliability and Separation. *Rasch Measurement Transactions*, vol. 9, no 4, p. 472.

References

Anderson L.W., Krathwohl D.R., Airasian P.W., Cruikshank K.A., Mayer R., Pintrich P.R., Raths J., Wittrock M.C. (eds) (2001) *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York, NY: Longman.

Bakai E.A., Yusupova E.M., Antipkina I.V. (2023) Chitayut ili delayut vid? Analiz povedeniya uchashchikhsya nachal'nykh klassov pri vypolnenii zadaniy testa chitatel'skoy gramotnosti [Reading or Pretending to Read? Analysis of the Behavior of Primary School Students during a Reading Comprehension Test]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 1, pp. 8–28. https://doi.org/10.17323/1814-9545-2023-1-8-28

Becker A., Nekrasova-Beker T. (2018) Investigating the Effect of Different Selected-Response Item Formats for Reading Comprehension. *Educational Assessment*, vol. 23, no 4, pp. 296–317. http://dx.doi.org/10.1080/10627197.2018.1517023

Burnham K., Anderson D. (eds) (2002) *Model Selection and Multi-Model Inference. A Practical Information-Theoretic Approach*. New York; Berlin; Heidelberg: Springer.

De Boeck P., Bakker M., Zwitser R., Nivard M., Hofman A., Tuerlinckx F., Partchev I. (2011) The Estimation of Item Response Models with the lmer Function from the lme4 Package in R. *Journal of Statistical Software*, vol. 39, iss. 12. http://dx.doi.org/10.18637/jss.v039.i12

De Boeck P., Wilson M. (eds) (2004) *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach.* New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4757-3990-9

Delgado P., Vargas C., Ackerman R., Salmerón L. (2018) Don't Throw Away Your Printed Books: A Meta-Analysis on the Effects of Reading Media on Reading Comprehension. *Educational Research Review*, vol. 25, November, pp. 23–38. http://dx.doi.org/10.1016/j.edurev.2018.09.003

Desjardins C.D., Bulut O. (2018) *Handbook of Educational Measurement and Psychometrics Using R.* Boca Raton, FL: CRC. http://dx.doi.org/10.1201/b20498

Douglas G. (1982) Issues in the Fit of Data to Psychometric Models. *Education Research and Perspectives*, vol. 9, no 1, pp. 32–43.

Effatpanah F., Baghaei P. (2021) Cognitive Components of Writing in a Second Language: An Analysis with the Linear Logistic Test Model. *Psychological Test and Assessment Modeling*, vol. 63, no 1, pp. 13–44.

Fischer G.H. (2005) Linear Logistic Test Models. *Encyclopedia of Social Measurement* (ed. K. Kempf-Leonard), Boston; London: Elsevier, pp. 505–514.

Fischer G.H. (1973) The Linear Logistic Test Model as an Instrument in Educational Research. *Acta Psychologica*, vol. 37, no 6, pp. 359–374. https://doi.org/10.1016/0001-6918(73)90003-6

Gosteva Yu.N., Kuznetsova M.I., Ryabinina L.A., Sidorova G.A., Chaban T. Yu. (2019) Teoriya i praktika otsenivaniya chitatel´skoy gramotnosti kak komponenta funktsional´noy gramotnosti [Theory and Practice of Reading Literacy as a Component of Functional Literacy]. *Otechestvennaya i zarubezhnaya pedagogika*, vol. 1, no 4 (61), pp. 34–57.

Lang J.W., Tay L. (2021) The Science and Practice of Item Response Theory in Organizations. *Annual Review of Organizational Psychology and Organizational Behavior*, vol. 8, pp. 311–338. http://dx.doi.org/10.1146/annurev-orgpsych-012420-061705

Lenkeit J., Chan J., Hopfenbeck T.N., Baird J.A. (2015) A Review of the Representation of PIRLS Related Research in Scientific Journals. *Educational Research Review*, vol. 16, October, pp. 102–115. http://dx.doi.org/10.1016/j.edurev.2015.10.002

Linacre J.M. (2004) Rasch Model Estimation: Further Topics. *Journal of Applied Measurement*, vol. 5, no 1, pp. 95–110.

Mair P., Hatzinger R., Maier M.J., Rusch T., Mair M.P. (2020) *ERm: Extended Rasch Modeling. 1.0-2*. Available at: https://cran.r-project.org/package=eRm (accessed 17 September 2023).

Mead R. (2008) *A Rasch Primer: The Measurement Theory of Georg Rasch. Psychometrics Services Research Memorandum 2008–001*. Maple Grove, MN: Data Recognition Corporation. Available at: http://www.edmeasurement.net/8226/Mead-2008-Rasch-primer.pdf (accessed 13 September 2023).

Melentieva Yu.P. (2015) *Obshchaya teoriya chteniya* [Theory of Reading]. Moscow: Nauka.

Mullis I.V., Martin M.O., Sainsbury M. (2016) *PIRLS 2016 Reading Framework*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, pp. 11–29.

OECD (2019) PISA 2018 Assessment and Analytical Framework. Paris: OECD. https://doi.org/10.1787/b25efab8-en

Pearson P.D., Johnson D.D. (1978) *Teaching Reading Comprehension*. New York, NY: Holt, Winehart and Winston.

Rahman T., Alexander P.A., Chae S.E. (2022) Reader Attributes, Task Attributes, and Reading Comprehension Proficiency: The Relation Revealed by Two Analytic Approaches. *Reading Psychology*, vol. 43, no 7, pp. 495–522. http://dx.doi.org/10.1080/02702711.2022.2126044

Støle H., Mangen A., Schwippert K. (2020) Assessing Children's Reading Comprehension on Paper and Screen: A Mode-Effect Study. *Computers & Edu-*

cation, vol. 151, March, Article no 103861. http://dx.doi.org/10.1016/j.compedu.2020.103861

Schwippert K., Lenkeit J. (eds) (2012) *Progress in Reading Literacy in National and International Context. The Impact of PIRLS 2006 in 12 Countries*. Münster: Waxmann Verlag.

Thompson B., Gipe J.P., Pitts M.M. (1985) Validity of the Pearson-Johnson Taxonomy of Comprehension Questions. *Reading Psychology: An International Quarterly*, vol. 6, no 1–2, pp. 43–49. https://doi.org/10.1080/0270271850060105

Verguts T., De Boeck P. (2000) A Note on the Martin-Löf Test for Unidimensionality. *Methods of Psychological Research*, vol. 5, no 1, pp. 77–82.

Warren W., Nicholas D., Trabasso T. (1979) Event Chaince and Inferences in Understanding Narratives. *New Directions in Discourse Processing, Vol. 2. Advances in Discourse Processing* (ed. R.O. Freedle), Norwood, NJ: Ablex Publication Corporation. https://doi.org/10.1017/S002222670000685X

Whitely S.E. (1983) Construct Validity: Construct Representation Versus Nomothetic Span. *Psychological Bulletin*, vol. 93, no 1, pp. 179–197. https://doi.org/10.1037/0033-2909.93.1.179

Woodcock S., Howard S.J., Ehrich J. (2020) A Within-Subject Experiment of Item Format Effects on Early Primary Students' Language, Reading, and Numeracy Assessment Results. *School Psychology*, vol. 35, no 1, pp. 80–87. http://dx.doi.org/10.1037/spq0000340

Wright B.D. (1996) Reliability and Separation. *Rasch Measurement Transactions*, vol. 9, no 4, p. 472.

Zuckerman G.A., Kovaleva G.S., Baranova V.Yu. (2018) Chitatel´skie umeniya rossijskikh chetveroklassnikov: uroki PIRLS-2016 [Reading Literacy of Russian Fourth-Graders: Lessons from PIRLS-2016]. *Voprosy obrazovaniya / Educational Studies* Moscow, no 1, pp. 58–78. https://doi.org/10.17323/1814-9545-2018-1-58-78

Психометрика и когнитивные исследования: противоречия и возможности кооперации

Юлия Кузьмина

Статья поступила в редакцию в марте 2023 г.

Кузьмина Юлия Владимировна — кандидат психологических наук, старший научный сотрудник Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики». Адрес: 101000, Москва, Потаповский пер., 16, стр. 10. E-mail: papushka7@gmail.com. ORCID: https://orcid.org/0000-0002-4243-8313

Аннотация

Когнитивная психология в основном развивалась в рамках экспериментальной парадигмы, в отличие от психометрики, занимающейся оценкой индивидуальных различий и корреляционными исследованиями. С целью обнаружить барьеры, стоящие на пути сотрудничества когнитивной психологии с психометрикой, в статье рассмотрена краткая история взаимоотношений между экспериментальными исследованиями и психометрикой с конца XIX в. до настоящего времени. Обсуждаются основные проблемы, возникающие в когнитивных исследованиях из-за недостаточного использования психометрических моделей и применения устаревших методов анализа результатов тестирования. По итогам предлагается ряд рекомендаций с точки зрения психометрики для повышения точности измерений индивидуальных различий в когнитивных процессах и способностях.

Ключевые слова

психометрика, когнитивная психология, экспериментальная психология, надежность, индивидуальные различия, анализ времени ответа, ингибиторная функция

Для цитирования

Кузьмина Ю.В. (2023) Психометрика и когнитивные исследования: противоречия и возможности кооперации. *Вопросы образования / Educational Studies Moscow*, № 3, сс. 113–144. https://doi.org/10.17323/vo-2023-16875

Psychometrics and Cognitive Research: Contradictions and Possibility for Cooperation

Yulia Kuzmina

Yulia V. Kuzmina — PhD in Psychology, Researcher at the Center for Psychometrics and Measurement in Education, National Research University Higher School of Economics. Address: Bld. 10, 16 Potapovsky Ln, 101000 Moscow, Russian Federation. E-mail: papushka7@gmail.com. ORCID: https://orcid.org/0000-0002-4243-8313

Abstract The article considered several issues of the relationships between cognitive psychology and psychometrics. Cognitive psychology has mainly developed within experimental paradigm in psychology, whereas psychometrics has developed within a different paradigm — assessment of individual differences and correlational studies. In the article it has been considered a brief history of the development of relationships between experimental studies and psychometrics, from the end of 19th century to the present. The historical view allows understanding problems in the use of experimental tasks for assessing individual differences and obstacles to the widespread of use psychometric models in experimental studies. Several recommendations are proposed to improve the accuracy of measurements of individual differences in cognitive abilities and processes, from psychometric perspectives.

Keywords psychometrics, cognitive psychology, experimental psychology, reliability, individual differences, analysis of reaction time, inhibitory function

For citing Kuzmina Yu.V. (2023) Psikhometrika i kognitivnye issledovaniya: protivorechiya i vozmozhnosti kooperatsii [Psychometrics and Cognitive Research: Contradictions and Possibility for Cooperation]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 3, pp. 113–144. https://doi.org/10.17323/vo-2023-16875

Психометрика как наука и область деятельности, связанная с измерением способностей, психологических конструктов и черт, с момента своего зарождения была тесно связана с психологией [Galton, 1879; Cattell, Galton, 1890; Borsboom, 2006], но это не значит, что психометрические подходы, в особенности то, что называется современной теорией тестирования (*Item Response Theory*, IRT), легко усваивались различными направлениями психологии [Borsboom, 2006]. Меньше других, в частности меньше социальной психологии и психологии личности, использует психометрические подходы и модели современной теории тестирования когнитивная психология.

Рассматривая специфику взаимоотношений между когнитивной психологией и психометрикой, следует иметь в виду изменения в содержании понятия «психометрика». На первых этапах развития психометрика представляла собой деятельность по измерению способностей — в первую очередь интеллекта, а также образовательных достижений и индивидуальных различий в способностях, что подразумевало прежде всего разработку и применение стандартизированных тестов. В настоящее время под психометрикой, скорее, подразумевается деятельность по разработке и применению статистических моделей, которые могут использоваться в образовательных или психологических исследованиях для моделирования взаимосвязи между латентными переменными и наблюдаемым поведением. Чаще других применяются модели современной теории тестирования или моделирование структурными уравнениями (*Structural Equation Modeling*, SEM) [Wijsen, Borsboom, Alexandrova, 2022].

Двум разным этапам развития психометрики соответствуют специфические проблемы в ее взаимоотношениях с когнитивной психологией. На первых этапах развития имели место трудности во взаимодействии исследований индивидуальных различий с экспериментальными исследованиями. В настоящее время проблемы возникают из-за использования неподходящих статистических моделей при измерении когнитивных конструктов, например из-за применения в когнитивных исследованиях устаревших подходов классической теории тестирования вместо современной теории тестирования.

Основным препятствием к тесному сотрудничеству между психометрикой и когнитивной психологией является их принадлежность к разным исследовательским традициям, или исследовательским парадигмам, в психологии: когнитивные исследования чаще проводятся в русле экспериментальной парадигмы, а психометрика как средство для оценки индивидуальных различий в интеллекте, в образовательных достижениях и т.п. развивалась в парадигме корреляционных исследований [Borsboom et al., 2009; Cronbach, 1957].

Цель исследований, проводимых в рамках экспериментальной парадигмы, обычно состоит в выделении общих закономерностей или эффектов, для чего могут сравниваться те или иные показатели в контрольной и экспериментальной группах (например, время, потраченное на выполнение заданий, т.е. время ответа) или в двух или нескольких экспериментальных условиях [Smedt de, Gilmore, 2010; Corneille, Mierop, Unkelbach, 2020]. При этом индивидуальные различия между участниками внутри групп считаются «шумом», который по возможности надо свести к минимуму [Borsboom et al., 2009]. Надежность эксперимента определяется тем, насколько он поддается воспроизведению. Чем меньше различий между участниками, чем больше среди участников тех, кто демонстрирует искомый эффект, тем более надежен полученный эффект и, соответственно, эксперимент. В этой парадигме люди взаимозаменяемы, поскольку предполагается, что выделяемые закономерности и процессы одинаковы для всех людей — по крайней мере для всех людей, не имеющих отклонений.

В рамках парадигмы корреляционных исследований и исследований индивидуальных различий, наоборот, гомогенность исследуемой выборки и низкий уровень межиндивидуальной дисперсии являются показателем неуспеха. Чем сильнее люди различаются между собой, чем выше уровень межиндивидуальной дисперсии, тем лучше. Задача состоит в том, чтобы как можно надежнее оценить уровень межиндивидуальных различий. Надежность теста или инструмента в таком случае понимается как способность одинаковым образом

ранжировать участников или как способность измерить оцениваемую способность с минимальной ошибкой [Dunn, Baguley, Brunsden, 2014; Cronbach, Shavelson, 2004]. Таким образом, специфика экспериментальных и корреляционных исследований определяет ограничения в кооперации между ними.

Взаимоотношения исследований индивидуальных различий и экспериментальных исследований станут понятнее, если рассмотреть их в исторической перспективе. Такое рассмотрение покажет, во-первых, что разрыв между когнитивными исследованиями и психометрикой возник не сразу, но связан с разной исследовательской логикой в двух подходах, и во-вторых, что этот разрыв не предопределен — а следовательно, может быть преодолен. На всем протяжении развития психологии многие исследователи подчеркивали возможность сближения рассматриваемых подходов, которое может обогатить психологию в целом.

1. Всегда ли экспериментальные исследования и исследования индивидуальных различий были разделены

Родоначальником исследований индивидуальных различий и первым психометриком считается Ф. Гальтон [Goldstein, 2012; Ludlow, 1998]. Он, в частности, предположил, что индивидуальные различия в сенсомоторных реакциях являются проявлением индивидуальных различий во врожденных способностях [Galton, 1883]. Он разработал систему тестирования некоторых сенсомоторных данных и организовал работу антропометрических лабораторий, которые за несколько лет собрали данные измерений около 17 тыс. человек. Уже в XX в. исследователи проанализировали данные, собранные Ф. Гальтоном, и нашли некоторые из них достаточно надежными для использования [Johnson et al., 1985].

Идеями измерения индивидуальных различий вслед за Ф. Гальтоном воодушевился известный американский психолог, одно время обучавшийся у В. Вундта, Джеймс Маккин Кеттэлл. Он считал, что психология должна развиваться и как наука экспериментальная (к этой традиции он относил исследования Вундта по измерению времени реакции и психофизиков с их исследованиями связи силы стимулов и ощущений), и как наука, использующая тесты и измерения: «Психология не может достичь достоверности и точности физических наук, если она не опирается на эксперимент и измерения. Шаг в этом направлении можно было бы сделать, применив ряд ментальных тестов и измерений к большому числу людей. Результаты будут иметь значительную научную ценность для открытия постоянства психических процессов, их взаимозависимости и их изменчивости при разных обстоятельствах»[1].

[1] "Psychology cannot attain the certainty and exactness of the physical sciences, unless it rests on a foundation of experiment and measurement. A step in this direction could be made by applying a series of mental tests and mea-

Дж. Кеттэлл разработал широкомасштабную программу тестирования способностей, которая должна была быть внедрена в американских колледжах. Программа предполагала сбор результатов о выполнении студентами многочисленных заданий, например определения скорости движений, минимально различаемой разницы в весе, выполнения задания на называние цветов и др. [Cattell, Galton, 1890]. Дж. Кеттэлл подчеркивал, что каждый участник должен пройти достаточно много заданий каждого типа, после чего должны быть рассчитаны средние показатели, дисперсия, максимум и минимум, все эти параметры должны описывать способности каждого тестируемого и предсказывать его академические успехи. Результаты осуществления этой программы не оправдали ожиданий: показатели выполнения студентами предложенных тестов не коррелировали ни друг с другом, ни с академическими успехами испытуемых [Wissler, 1901].

Тем не менее в начале становления и развития психологии экспериментальные исследования и исследования индивидуальных различий скорее дополняли друг друга, чем противопоставлялись. Например, перед экспериментатором выдвигалась задача выявить источник обнаруженных в эксперименте индивидуальных различий в результативности испытуемых: являются ли они следствием различий во врожденных способностях или эффектом обучения и практики [Wells, 1912]. Э. Боринг считал тесты сокращенной версией психологических экспериментов, а Г. Годдард подчеркивал, что разницу между тестами и экспериментами определяет в основном способ использования их результатов [Terman, 1924].

Однако по мере развития теорий и методологических подходов и накопления эмпирических данных различия между экспериментальными исследованиями и исследованиями индивидуальных различий углублялись, и в 1920-х годах научное сообщество уже ясно осознавало специфику подходов и результатов в экспериментальных исследованиях и в тестах. Л. Термен описывал следующие различия: тесты имеют дело с оценкой индивидуальных различий, а не с выявлением общих законов; тесты применяются к большому числу субъектов и нацелены на быстрое определение состояния и поведения, а не внутреннего психологического содержания; результаты тестов, хотя и имеют научную ценность, как правило, менее точны, чем результаты психологических экспериментов [Ibid.].

surements to a large number of individuals. The results would be of considerable scientific value in discovering the constancy of mental processes, their interdependence, and their variation under different circumstances" [Cattell, Galton, 1890. P. 373].

Вряд ли можно выделить какую-то одну причину появления, а затем и углубления разрыва между экспериментальными исследованиями и применением тестов. Показательна история их отношений в США: там с середины 1910-х годов тесты активно внедрялись во все сферы деятельности, их широко применяли в образовании, армии, бизнесе, рекламе. Популярность тестов была обусловлена многими политическими и экономическими обстоятельствами [Sokal, 1987]. С одной стороны, распространение тестирования позволило большому числу психологов найти себе достойную работу и зарплату, а психологии — доказать свою общественную значимость [Шульц, Шульц, 1998]. Некоторые исследователи считают, что тестирование в США стало технологией, так что американские психологи могли позиционировать свою деятельность в качестве такой же полезной, как инженерное дело, и такой же уважаемой, как медицина [Brown, 1992]. С другой стороны, с массовым внедрением тестирования деятельность психометриков приобретала характер, существенно отличающийся от чисто академических исследований.

Распространившийся в те же годы в США бихевиоризм провозгласил, что психология должна стать одной из ветвей естественных наук, а для этого она должна заниматься изучением только наблюдаемого поведения, используя экспериментальные методы [Watson, 1913]. Психология должна была строиться по образцу естественных наук, поэтому основным исследовательским методом должен был стать эксперимент. Бихевиоризм не интересуется индивидуальными различиями, для него есть только общие законы поведения, универсальные для человека и для животных [Braat et al., 2020]. Экспериментальная психология и психометрика продолжали активно развиваться, но отдельно друг от друга.

В середине 1950-х годов в психологии окончательно оформилось представление о психологах-экспериментаторах и исследователях индивидуальных различий (психометриках) как о двух разных классах исследователей, отношения между которыми характеризуются соперничеством и непониманием: «Если психометрик и экспериментатор в чем-то и согласны — а в этом есть определенное сомнение, — то только в одном: каждый из них считает, что другой играет в низшей лиге»[2].

Чуть позже, в 1957 г., вышла известная статья Л. Кронбаха, в которой он констатировал разделение психологии на две ветви: корреляционную психологию и экспериментальную психологию [Cronbach, 1957]. Л. Кронбах подчеркивал, что корреляционные психологи, к которым он относит, например,

[2] "If the psychometric researcher and the experimentalist agree on anything, and there is some doubt about this, it is that the other kind of psychologist plays in the other league (class B)" [Bindra, Scheier, 1954. P. 69].

исследователей психологии развития, индивидуальных различий, личности, хотя и занимаются исследованиями в разных областях психологии, получают схожую подготовку и образование. А экспериментальные психологи получают иную подготовку и могут заниматься экспериментальными исследованиями, не имея знаний в теории тестирования или дифференциальной психологии. О взаимоотношениях представителей двух ветвей психологии Л. Кронбах писал: «Психологи, занимающиеся исследованиями личности, детской и социальной психологии, пошли одним путем, исследователи восприятия и обучения пошли другим, и страна между ними превратилась в пустыню»[3]. Л. Кронбах полагал, что будущее психологии связано с объединением этих двух ветвей, поскольку они могут обогатить друг друга. В частности, он писал, что объединенная психология должна исследовать и различия между субъектами, и различия между условиями, и взаимодействие между субъектом и ситуацией. Исследователи, работающие в корреляционной парадигме, могут помочь исследователям-экспериментаторам выработать новое понимание межиндивидуальных различий, а также предложить новые методы, такие как факторный анализ.

Оживлению дискуссии о различиях в двух исследовательских парадигмах и о необходимости их объединения в 1950-е годы могли способствовать несколько факторов. Первый фактор — это, несомненно, изменения, произошедшие в психологии. В начале 1950-х годов в психологии началось движение, названное в исследовательской литературе когнитивной революцией, что означало возвращение в психологию исследований познавательных процессов [Miller, 2003]. Термин «когнитивная революция» получил широкое распространение, но некоторые авторы считают, что речь шла, скорее, не о революции, а о развитии идей необихевиоризма и отказе от радикального бихевиоризма [Watrin, Darwich, 2012; Moore, 1999]. Из всех направлений психологии именно когнитивная психология (этот термин закрепился уже в 1960-е годы), по мнению некоторых исследователей, теснее всего связана с идеями бихевиоризма и, по сути, является эволюцией необихевиоризма, по крайней мере с точки зрения методологии [Moore, 1996; Watrin, Darwich, 2012]. Отказ от радикального бихевиоризма означал оживление интереса к исследованиям способностей и индивидуальных различий, что отразилось на всех областях психологии [Lamiell, 1992; Royer, 2006].

[3] "The personality, social, and child psychologists went one way; the perception and learning psychologists went the other; and the country between them turned into desert" [Cronbach, 1957. P. 673].

С другой стороны, большие изменения происходят и в психометрике. Когнитивная революция в психологии происходила практически одновременно с рождением современной теории тестирования. Ее основу составили работы, которые в 40–50-х годах XX в. вели параллельно разные исследователи: Ф. Лорд и А. Бирнбаум в США, Г. Раш в Дании, П. Лазерфельд в Австрии [Lord, 1953; Rasch, 1960; Lazerfeld, 1950; Birnbaum, 1958]. В некотором роде разработка моделей современной теории тестирования — это поиск объективности в измерении, попытка преодолеть ограничения классической теории тестирования и создать статистические модели, которые бы удовлетворяли условиям объективного измерения, где баллы оценки латентных способностей не зависят от набора заданий и выборки [Rasch, 1968]. Казалось бы, возвращение когнитивных процессов в психологию и развитие теории объективных измерений в психометрике могут способствовать объединению двух подходов, но оно не произошло.

Нельзя сказать, что таких попыток не было. Необходимость и перспективы сотрудничества когнитивной психологии и психометрики периодически обсуждались и продолжают обсуждаться [Sternberg, 1981; Glaser, 1981; Embretson, 1994; Embretson, Gorin, 2001]. Но складывается впечатление, что в таком объединении долгое время были больше заинтересованы психометрики, чем когнитивные психологи: прежде всего обсуждалось, как когнитивная психология может помочь разработчикам тестов и усилить сферу тестирования [Embretson, 1994]. Для этого разработка заданий тестов должна происходить с опорой на теории когнитивных процессов [Embretson, Gorin, 2001].

Позже в психометрии усилилось влияние идей современной теории тестирования и применения более продвинутых статистических моделей для анализа результатов когнитивных тестов [Rouder, Haaf, 2019; Maas van der et al., 2011; Rouder, Kumar, Haaf, 2019]. Но, по мнению некоторых психометриков, проникновению психометрики в психологию препятствовало отсутствие в психологии сильных теорий [Borsboom, 2006]. Имеются в виду теории, требующие применения какой-то конкретной психометрической модели. Такое предположение касалось прежде всего исследований в психологии личности, но его можно отнести и к когнитивным исследованиям.

Интерес когнитивных психологов к исследованию индивидуальных различий в когнитивных процессах оживился с появлением возможности анализировать нейрофизиологические корреляты тех или иных процессов [Gevins, Smith, 2000; Drew, Vogel, 2008]. Кроме того, в когнитивных исследованиях стали активно использоваться техники конфирматорного факторного анализа и моделирования структурными уравнениями для

исследования факторной структуры сложных когнитивных конструктов [Kane et al., 2004; Friedman, Miyake, 2017].

Таким образом, мы можем наблюдать сближение когнитивной психологии — в той ее части, которая отходит от методологии необихевиоризма, — и психометрики. При этом существующие подходы к оценке индивидуальных различий в когнитивных исследованиях подвергаются критике с учетом уже сложившихся традиций и практик [Hedge, Powell, Sumner, 2018; Goodhew, Edwards, 2019].

2. Сложившиеся традиции и практики анализа результатов исследования

Помимо уже указанных выше особенностей понимания надежности и отношения к межиндивидуальной дисперсии, между когнитивными исследованиями, проводимыми в экспериментальной парадигме, с одной стороны, и исследованиями индивидуальных различий и психометрикой — с другой, существуют и другие различия.

2.1. Трудность заданий: фиксированная или вариативная

Когнитивные исследования и психометрика по-разному подходят к конструированию заданий. В частности, в когнитивных исследованиях часто используются так называемые элементарные когнитивные задания — задачи с низкой трудностью, разработанные в рамках информационно-процессуального подхода [Simon, 1979]. Предполагается, что любой человек при отсутствии ограничений по времени может выполнить такое задание, поэтому, как правило, точность их выполнения в среднем очень высокая и приближается к 100% [Jensen, 2006; Rouder, Haaf, 2019]. Ошибки, совершаемые в таких заданиях, ничего не говорят ни об особенностях процесса решения, ни о способности респондента, а могут интерпретироваться как «шум». В ситуации, когда точность выполнения задания приближается к 100%, информация о правильности выполнения задания не может быть использована для оценки способности респондента или различий между группами или условиями. Поэтому в качестве основного индикатора выполнения заданий чаще всего используется время ответа (или время реакции) [Whelan, 2008; Baayen, Milin, 2010].

В психометрических исследованиях, наоборот, в тесты включаются задания разной трудности и способность респондента анализируется с учетом точности (правильности) выполнения задания (например, при оценке образовательных достижений [Ackerman, Gierl, Walker, 2003]) или ответов в тестах психологических черт с использованием шкал Ликерта [Spens, Owens, Goodyer, 2012]. Одно из преимуществ психометрических моделей в рамках современной теории тестирования состоит

в моделировании вероятности правильного выполнения задания с учетом и способности респондента, и трудности задания [Hambleton, 1989].

Описанные различия в конструировании тестов, безусловно, не являются абсолютными и непреодолимыми. В когнитивных исследованиях, как и в психометрике, могут использоваться задания с разным уровнем трудности, в которых возможна градация точности ответа [Dietrich, Huber, Nuerk, 2015]. В то же время в психометрических исследованиях для более точной оценки способности респондентов все чаще используют информацию о времени ответа, хотя, как правило, совместно с информацией о точности [Linden van der, 2009; Molenaar et al., 2015].

2.2. Анализ времени ответов: процессные модели или моделирование латентных конструктов

Время выполнения задания часто используется в когнитивных исследованиях для анализа результатов выполнения заданий и описания характеристик когнитивных процессов. При этом учитывается, что общее время ответа не равно скорости оцениваемых процессов. Еще в 1890 г. Дж. Кеттэлл указывал на трудности при анализе времени ответа, поскольку общее время ответа не дает информации о скорости отдельных процессов, протекающих во время выполнения задания. Для того чтобы отделить друг от друга разные процессы и оценить скорость протекания отдельного процесса, используются различные математические модели, которые условно можно назвать процессными, поскольку их задача — отделить целевой процесс от сопутствующих. К этому классу можно отнести диффузионную модель [Ratcliff, Smith, McKoon, 2015; Воронин и др., 2020].

Несмотря на существование процессных моделей, результаты тестирования в когнитивных исследованиях до сих пор чаще всего анализируются с применением оценки среднего времени ответа (или медианы) для определенных условий и групп и сравнения этих показателей. Иногда используется время ответа только для правильно выполненных заданий, может применяться трансформация времени ответа, например логарифмирование [Whelan, 2008; Baayen, Milin, 2010; Lo, Andrews, 2015]. При таком подходе время ответа отождествляется со скоростью процессов и становится возможным индикатором способности респондента — при условии одинаковой трудности заданий [Dodonova, Dodonov, 2013].

В рамках современной теории тестирования анализ времени ответа не получил широкого распространения. Однако в последние годы психометрики также стали предлагать модели для анализа времени ответа. При этом, в отличие от процессных моделей, используемые в психометрике модели со време-

нем ответа используются для более точной оценки способности респондента. Так как для IRT критически важна отдельная оценка параметров респондента и задания, рассматриваемые модели выделяют отдельные параметры для оценки трудности задания и его временной нагрузки, способности респондента и его быстроты [Goldhammer et al., 2014; Molenaar et al., 2015; Bolsinova, Tijmstra, 2018; Maas van der et al., 2011].

В настоящее время существует много разновидностей психометрических моделей с использованием времени ответа и точности, в некоторых из них эти параметры могут моделироваться вместе, но могут быть не связаны друг с другом [De Boeck, Jeon, 2019]. Таковы, например, обобщенные линейные иерархические модели [Linden van der, 2007]. В них применяются три иерархических уровня моделирования: внутрииндивидуальный, индивидуальный и уровень популяции (межиндивидуальный). Для точности и времени ответа создаются две разные модели, на каждом из уровней выделяются параметры заданий и параметры респондента.

В других моделях моделируется связь между двумя латентными переменными для точности и скорости: например, в *Bivariate Generalized IRT model* (B-GLIRT-модель) [Molenaar, Tuerlinckx, van der Maas, 2015] или в модели локальной зависимости [Bolsinova, Tijmstra, 2018]. В B-GLIRT-модели идентифицируются два латентных фактора для точности и скорости, а также моделируется функция связи между ними [Molenaar, Tuerlinckx, van der Maas, 2015].

Наконец, есть психометрические модели, в которых время ответа является независимой переменной для точности [De Boeck, Jeon, 2019]. Таковы, в частности, модели со смешанными эффектами для точности как зависимой переменной. С их помощью было показано, что связь между временем ответа и вероятностью дать правильный ответ зависит от типа задач и уровня подготовленности респондента [Goldhammer et al., 2014].

Некоторые из предлагаемых психометриками моделей представляют собой доработанные варианты диффузионной процессной модели (например, *positive ability model* [Maas van der et al., 2011]). Диффузионная модель является разновидностью двухпараметрической IRT-модели и при определенных условиях может быть использована для оценки способностей [Tuerlinckx, De Boeck, 2005].

Итак, в психометрике время ответа позволяет уточнить или выделить дополнительные параметры заданий и используется для оценки способности респондента, а в когнитивной психологии время ответа служит для описания когнитивного процесса, стоящего за выполнением заданий. В некоторых случаях в когнитивной психологии время ответа также может быть ис-

пользовано для описания способности респондента, но практически всегда это происходит без учета трудности задания.

2.3. Количество заданий: фиксированное или меняющееся

Определяя количество заданий, достаточное для тестирования той или иной способности или черты или для оценки определенного процесса, психометрики и исследователи когнитивных процессов исходят из разных целей. Стандартизированные тесты и психологические опросники, как правило, содержат фиксированное количество заданий, которое не меняется от исследования к исследованию. При этом для идентификации латентного конструкта, т.е. измеряемой способности, часто используется небольшое количество заданий. Например, для идентификации латентного конструкта с помощью конфирматорного факторного анализа достаточно трех заданий. Чем меньше заданий, тем быстрее испытуемый пройдет тест. Время, необходимое для тестирования, оказывается принципиальным фактором при сборе данных на больших выборках — а в корреляционных исследованиях размер выборки важен. Для экзаменов же с высокими ставками, конечно, необходимо больше заданий, чтобы свести к минимуму ошибку при оценке способности экзаменуемого. В любом случае в стандартизированных тестах количество заданий не меняется. Более того, психометрики настаивают на том, что изменение количества заданий требует новой идентификации модели и оценки психометрических свойств инструмента [Rouder, Haaf, 2019; Kleka, Soroko, 2018].

Так как в когнитивных исследованиях стоит задача элиминировать «шум», а размер выборки долгое время был не очень важен, здесь, как правило, для оценки того или иного процесса или функции используется много заданий. При этом количество заданий может существенно варьировать от исследования к исследованию. Например, одно из самых популярных заданий для исследования процесса подавления действия нежелательных стимулов (ингибиторной функции) и когнитивного контроля — вербально-цветовой тест Струпа [Stroop, 1935]. В классическом исследовании Д. Струпа предлагалось по 100 заданий для каждого условия: в первом условии слово, обозначающее цвет, не совпадает с цветом шрифта, во втором — слово, обозначающее цвет, напечатано черным шрифтом. Затем тест стали использовать другие исследователи, они меняли число заданий и другие параметры эксперимента, например вводили условие, при котором значение слова, обозначающего цвет, и цвет шрифта совпадают. В разных исследованиях может быть от 10 до 100 заданий [Scarpina, Tagini, 2017].

2.4. Измерительная инвариантность: проверять или нет

Измерительная инвариантность, т.е. одинаковая работа заданий на разных выборках, — очень важный принцип психометрики [Putnick, Bornstein, 2016; Leitgöb et al., 2023]. Чтобы сравнивать группы по уровню способности или степени выраженности какого-то латентного конструкта, необходимо убедиться, что используемый инструмент (задания) работает одинаково во всех рассматриваемых группах. Измерительная инвариантность может быть проверена с помощью конфирматорного факторного анализа или в рамках современной теории тестирования [Kim, Yoon, 2011; Meade, Lautenschlager, 2004].

Чтобы убедиться в измерительной инвариантности используемого инструмента, необходимо проверить три предположения: о сохранении факторной структуры в группах (конфигуральная инвариантность), об одинаковых факторных нагрузках для сравниваемых групп (метрическая инвариантность) и об одинаковых интерцептах для метрических индикаторов или пороговых значений для категориальных индикаторов (скалярная инвариантность).

В рамках современной теории тестирования контроль измерительной инвариантности равнозначен проверке DIF (*Different Item Functioning*). Проверить DIF — значит оценить инвариантность функции ответа на задания: будет ли одна и та же модель подходить всем группам респондентов с одними и теми же параметрами заданий [Kim, Yoon, 2011]. Проверка измерительной инвариантности, или оценка DIF, — распространенная практика оценки психометрических свойств, без доказанной инвариантности сравнивать группы нельзя [Putnick, Bornstein, 2016].

Сравнение групп участников по успешности выполнения заданий нередко выполняется и в когнитивных исследованиях [Passolunghi, Siegel, 2001], но измерительная инвариантность инструментария контролируется обычно только на клинических выборках. Д. Борсбум, рассматривая вопросы взаимодействия психометрики и психологии, указывал на проблему измерительной инвариантности тестов интеллекта [Borsboom, 2006]. Впрочем, с тех пор появилось немало исследований измерительной инвариантности когнитивных тестов: тестов интеллекта, тестов для оценки рабочей памяти [Wicherts, 2016; Willoughby et al., 2012]. При этом проверка измерительной инвариантности производится преимущественно в исследованиях интеллекта — конструкта, который относится скорее к психометрике, чем собственно к когнитивной психологии.

Итак, когнитивная психология начинает более активно заниматься исследованиями индивидуальных различий, но между ней и психометрикой сохраняется определенный разрыв, обусловленный различиями в исследовательских подходах и

сложившимися традициями. Вместе с ростом интереса исследователей когнитивных процессов к оценке индивидуальных различий в тестируемых способностях увеличивается и число публикаций, в которых обсуждаются проблемы, связанные с использованием для этой цели привычных для когнитивных психологов инструментов, например теста Струпа или теста флангов.

3. Проблемы измерений индивидуальных различий в когнитивных исследованиях с точки зрения психометрики

3.1. Ограничение трудности заданий

Ограничение трудности заданий может иметь несколько неблагоприятных последствий, связанных друг с другом. Во-первых, в результате отбора заданий с низким уровнем трудности сокращается дисперсия точности выполнения заданий на межиндивидуальном уровне. Из-за этого может снижаться надежность используемых тестов и уменьшаться возможная корреляция между измеряемым конструктом и другими переменными. Во-вторых, вследствие ограничения трудности заданий возникает необходимость анализировать время ответа, что создает дополнительные методологические трудности. Далее мы рассмотрим кратко обе проблемы.

3.2. Низкая надежность используемых инструментов

Многие инструменты, применяемые в когнитивных исследованиях, не подходят для оценки индивидуальных различий, поскольку имеют низкую надежность как на гомогенной выборке, так и на гетерогенной [Hedge, Powell, Sumner, 2018; Pronk et al., 2023]. Особенно много критики звучит в адрес заданий, используемых для оценки функции подавления [Rey-Mermet, Gade, Oberauer, 2018; Rouder, Kumar, Haaf, 2019]. Некоторые исследователи считают, что низкая надежность инструментов может быть связана с особенностями расчета индивидуальных баллов во многих заданиях на измерение функции подавления [Hedge, Powell, Sumner, 2018].

Обычно при расчете индивидуальных показателей по тестам для оценки функции подавления используют разницу в средних показателях времени ответа между конгруэнтными и неконгруэнтными заданиями или между нейтральными и неконгруэнтными. Чем больше разница между конгруэнтными и неконгруэнтными заданиями, тем ниже уровень ингибиторной функции. Отмечалось, что в целом надежность баллов, рассчитанных как разница времени ответа между двумя условиями, всегда ниже, чем надежность баллов в каждом отдельном условии, — возможно, из-за того, что дисперсия разницы баллов ниже, чем дисперсия в каждом из условий, особенно если корреляция показателей высокая [Caruso, 2004; Eide et al., 2002; Edwards, 2001].

Низкая надежность и недостаточный уровень дисперсии в некоторых тестах могут также приводить к снижению или отсутствию связи между результатами тестов, измеряющих, по идее, один и тот же конструкт. Например, показано, что результаты вербально-цветового теста Струпа имеют очень низкую корреляцию с показателями флангового теста, хотя оба инструмента измеряют устойчивость к нерелевантным стимулам или подавление доминирующего стимула [Rey-Mermet, Gade, Oberauer, 2018; Stahl et al., 2014]. Возможны два объяснения низкой корреляции показателей. Первое: эти два типа заданий действительно оценивают разные способности, и поэтому показатели не коррелируют друг с другом. Второе: низкая корреляция связана с низкой надежностью вследствие большой ошибки измерения и/или с низким уровнем дисперсии для одной или двух переменных. При этом степень снижения корреляции зависит от числа заданий и вариабельности между ними [Rouder, Haaf, 2019].

3.3. Проблемы с анализом времени ответа

Применительно ко времени ответа наиболее часто обсуждаются следующие методологические вопросы: время ответа обычно не имеет нормального распределения, эта переменная не имеет отрицательных значений, и ее распределение имеет правый скос [Whelan, 2008]. Кроме того, часто при анализе времени ответа обнаруживаются выбросы, которые могут исказить оценку средних эффектов [Heathcote, Popiel, Mewhort, 1991; Rousselet, Wilcox, 2020; Baayen, Milin, 2010]. В итоге используемые средние показатели времени ответа могут не отражать реальной тенденции и искажать оценку эффектов [Speelman, McGann, 2013].

В одном из исследований с использованием теста Струпа были проанализированы средние показатели времени ответа по каждому из условий. Выявлены значимые различия по времени ответа между неконгруэнтными и нейтральными заданиями (эффект интерференции), но не между конгруэнтными и нейтральными (эффект фасилитации отсутствует). Оказалось, что время ответа соответствует, скорее, экспоненциально модифицированному распределению Гаусса (*ex-Gaussian*). С учетом характера распределения были рассчитаны иные показатели мер средней тенденции и получены иные результаты в отношении эффектов, при этом подтверждены оба эффекта — и фасилитации, и интерференции [Heathcote, Popiel, Mewhort, 1991].

Для того чтобы решить проблему с отклонениями от нормального распределения, некоторые исследователи рекомендуют отказываться от использования средних показателей и параметрических методов анализа, а рассчитывать, напри-

мер, медиану вместо среднего [Whelan, 2008; Speelman, McGann, 2013]. Однако использование медианы тоже не всегда «спасает». В частности, показано, что на маленькой выборке медианные оценки могут быть более смещенными, чем среднее. Кроме того, медиану не рекомендуют использовать, например, если в исследовании сравниваются условия, тестируемые с помощью разного количества заданий [Rousselet, Wilcox, 2020].

Еще один распространенный вариант борьбы с отклонениями от нормального распределения при использовании времени ответа — трансформация переменной. Наиболее часто применяют логарифмирование [Schramm, Rouder, 2019; Lo, Andrews, 2015]. Трансформация позволяет применять параметрические методы анализа, к которым привыкли большинство психологов. Кроме того, она может быть полезна для обнаружения небольших эффектов [Schramm, Rouder, 2019]. Однако следует учитывать, во-первых, что происходящее после трансформации изменение шкалы измерения времени не всегда имеет смысл с точки зрения теории и интерпретации полученных результатов [Lo, Andrews, 2015]. Во-вторых, оценки эффектов на «сырой» шкале и на трансформированной шкале могут различаться, например на логарифмированной шкале могут обнаружиться значимые эффекты, которых нет на «сырой» шкале. Некоторые исследователи также отмечают, что иногда разные виды трансформации используются для *p-hacking* — для манипуляции данными с целью получить значимые эффекты [Moris Fernández, Vadillo, 2020].

Если цель исследования заключается в оценке различий между условиями, нет необходимости стремиться к нормальному распределению (при этом необходимо принять решение о том, что делать с выбросами). Однако если цель состоит в оценке связей между временем ответа и другими переменными, использование трансформации оправданно [Schramm, Rouder, 2019].

Оценка индивидуальных различий только на основании времени ответа может быть проблематичной еще и потому, что надежность разных показателей с использованием времени ответа ниже надежности индикаторов с использованием точности [Draheim et al., 2019; Dietrich et al., 2016; Saville et al., 2011].

3.4 Изменение количества заданий

Варьирование числа заданий может существенно изменить показатели корреляций, надежности и размеров эффекта в случае использования подходов классической теории тестирования вследствие нарушения допущения портативности (*portability*) — неизменности значений, полученных с помощью инструмента, в популяции независимо от размера выборки [Rouder, Haaf,

2019]. Так, показано, что увеличение количества заданий для каждого из условий в вербально-цветовом тесте Струпа ведет к нарастанию размеров эффекта и повышению надежности — этот эффект известен в классической теории тестирования. Увеличение длины шкалы приводит к уменьшению ошибки измерения [Rouder, Haaf, 2019]. Следовательно, для того чтобы сравнивать размеры эффекта, полученные в разных исследованиях, необходимо учитывать разницу в числе заданий, а не только размер выборки.

3.5. Использование агрегированных данных

В большинстве когнитивных исследований — как в экспериментальных, так и при анализе индивидуальных различий — используются агрегированные показатели времени ответа (например, среднее время ответа для респондента или группы респондентов, среднее время для условий) или точности (сумма или пропорция правильных ответов). В частности, для теста Струпа рассчитываются пропорция правильных ответов (или пропорция ошибок) и/или среднее время правильного ответа для каждого условия и средняя разница между конгруэнтными, неконгруэнтными и нейтральными условиями [Schmidt, Besner, 2008; Schichel, Tzeglov, 2018]. Далее эти показатели также могут агрегироваться на уровень выборки или группы [Hepp et al., 1996].

Основная проблема, связанная с использованием агрегированных данных, состоит в потере информации о внутрииндивидуальной вариабельности, т.е. дисперсии времени ответа или точности между заданиями, в то время как такая дисперсия может быть выше межиндивидуальной [Rouder, Haaf, 2019]. В психологии уже давно известны ограничения использования агрегированных данных и возможные негативные эффекты их применения, в частности парадокс Симпсона: связи между двумя агрегированными переменными могут быть прямо противоположными связям между переменными, оцененными не на агрегированном уровне [Kievit et al., 2013]. В когнитивных исследованиях, например, связь между временем ответа и точностью на уровне индивида может отличаться от связи между этими переменными на уровне выборки [Moleenar, Tuerlinckx, van der Maas, 2015]. Для того чтобы учитывать внутрииндивидуальную дисперсию и различия в характере связи между переменными на уровне индивида и на уровне выборки, необходимо применять анализ на уровне заданий с использованием моделей со смешанными эффектами или конфирматорного факторного анализа [Moleenar, Tuerlinckx, van der Maas, 2015; Brauer, Curtin, 2018; Cunnings, 2012]. Однако в когнитивных исследованиях применение таких моделей все еще скорее исключение.

4. Есть ли возможности для взаимодействия

Несмотря на наличие проблем, связанных с измерениями, в когнитивных исследованиях, некоторые исследователи считают, что когнитивная психология и психометрика могли бы наладить более тесное сотрудничество и были бы полезны друг другу. Психометрики могут использовать возможности когнитивной психологии для выдвижения теорий и моделей когнитивных процессов, операционализации конструктов, генерации заданий для тестов [Embretson, Gorin, 2001]. Для когнитивных исследований психометрика может быть источником статистических моделей и подходов для оценки психометрических свойств тестов, а также для анализа полученных результатов с учетом специфики заданий и способности респондентов [Maas van der et al., 2011; Rouder, Haaf, 2019; Heck, Erdfelder, 2016].

Проблема измерений в когнитивной психологии возникает в тот момент, когда экспериментальная парадигма трансформируется в парадигму оценки индивидуальных различий, когда инструменты, обычно работающие в экспериментальных исследованиях, привлекаются для оценки индивидуальных различий. Использование психометрических моделей в когнитивной психологии обсуждается именно применительно к оценке индивидуальных различий, психометрики не претендуют на выявление общих механизмов или законов памяти и восприятия. Основной посыл со стороны психометриков когнитивным психологам можно сформулировать так: если исследователь когнитивных процессов переходит от экспериментов и оценки средних эффектов к тестированию способностей и индивидуальных различий, необходимо пользоваться уже разработанными психометрическими подходами и моделями, чтобы делать это правильно.

Что значит правильно? Психометрики, вероятно, могли бы дать ряд рекомендаций по измерению индивидуальных различий в когнитивных исследованиях.

Во-первых, следует изменить подходы к использованию и разработке инструментов для оценки индивидуальных различий в когнитивных процессах. Не всегда стоит применять методики, работающие в экспериментальных исследованиях, даже если они широко известны и хорошо себя зарекомендовали. Возможно, имеет смысл создавать новые инструменты для исследования индивидуальных различий с учетом ранее выявленных проблем. Например, учитывая сравнительно низкую надежность показателей на основе времени ответа, стоит разрабатывать инструменты, в которых будет изменяться трудность заданий и будет оцениваться точность ответа, а не только время ответа [Draheim et al., 2021]. При использовании популярных экспериментальных методик для оценки индивидуальных различий, возможно, имеет смысл переходить от общих назва-

ний методик к указанию того, что конкретно измеряет данная методика (например, тест Струпа — тест устойчивости к дистракторам) по аналогии со шкалами, применяемыми в исследованиях психологических конструктов. Из имеющихся вариантов экспериментальных методик и условий целесообразно отбирать тот, который имеет наибольшую межиндивидуальную дисперсию и надежность [Goodhew, Edwards, 2019]. Кроме того, для каждой методики необходимо указывать число заданий и учитывать его при интерпретации результатов. Возможно, имеет смысл разрабатывать стандартизированные методики с фиксированным количеством заданий и стимулов. Такие методики есть: например, модифицированная и стандартизированная версия теста Струпа, разработанная в Университете Виктории (*Victoria Stroop test*) [Troyer, Leach, Strauss, 2006], но для анализа результатов используются «классические» методы с агрегированными данными.

Во-вторых, важно при публикации результатов сообщать надежность когнитивных тестов, используемых в исследованиях индивидуальных различий [Parsons, Kruijt, Fox, 2019]. При этом надежность каждого конкретного теста должна быть оценена отдельно, нельзя полагаться на ранее полученные оценки надежности, поскольку они могут зависеть от параметров выборки. Достаточно часто психологи используют в качестве показателя надежности коэффициент альфа Кронбаха, который подразумевает ряд ограничений [Kim, Feldt, 2010; Dunn, Baguley, Brunsden, 2014; Tavakol, Dennick, 2011]. В настоящее время психометрики рекомендуют применять другие показатели надежности, например коэффициент омега [Dunn, Baguley, Brunsden, 2014].

В-третьих, при интерпретации полученных корреляций не стоит полагаться на оценку надежности как на гарантию их точной оценки [Rouder, Kumar, Haaf, 2019]. Высокая надежность используемых шкал может служить такой гарантией только при применении стандартных тестов или шкал, содержащих одни и те же формулировки заданий и одинаковое количество заданий. В экспериментальных методиках такое бывает редко. Поэтому даже при получении высоких показателей надежности нужно учитывать возможность недооценки корреляции на уровне выборки [Ibid.].

В-четвертых, психометрики рекомендуют перестать использовать агрегированные показатели: среднюю точность или среднее время для респондента, или для условия, или для выборки. Их следует заменить моделями, которые учитывают дисперсию между заданиями, например моделями со смешанными эффектами или иерархическими моделями конфирматорного факторного анализа [Rouder, Kumar, Haaf, 2019; Molenaar,

Tuerlinckx, van der Maas, 2015]. В психометрике за последние годы разработано много моделей, учитывающих как точность, так и время ответа, эти модели можно применять для анализа результатов когнитивных тестов. Например, B-GLIRT-модель может учитывать дисперсию как на внутрииндивидуальном уровне, так и на межиндивидуальном. В эту модель могут быть включены параметры ответа и времени ответа для каждого задания, на этой основе можно оценивать латентные способности респондента, его быстроту как латентную характеристику, а также разные типы связи между точностью и скоростью [Molenaar, Tuerlinckx, van der Maas, 2015]. При этом модель допускает возможность включения нелинейных связей между точностью и скоростью и взаимодействие между ними и трудностью задания.

Различия в двух рассмотренных психологических традициях — экспериментальной и дифференциально-измерительной, — возможно, непреодолимы в обозримом будущем не столько из-за разницы в методологических подходах, сколько из-за различий в предмете исследования [Borsboom et al., 2009]. Поэтому, по выражению Д. Борсбума, стоит принять рабочую гипотезу о разделенной психологии.

При этом необходимо учитывать и то общее, что стоит за двумя традициями. Исследователи индивидуальных различий не должны исключать возможность, что некоторые межиндивидуальные различия могут быть порождены системами внутрииндивидуальных процессов, и, наоборот, теории внутрииндивидуальных процессов не исключают возможности межиндивидуальных различий [Borsboom et al., 2009]. Кроме того, важно понимать ограничения каждого подхода. Их наличие означает, что результаты, полученные в экспериментальных исследованиях, не могут и не должны быть приложимы для описания отдельных индивидов. И наоборот, результаты изучения индивидуальных различий не могут быть прямо перенесены на описание внутрииндивидуальных процессов и механизмов [Molenaar, Beltz, 2020].

Благодарности

Исследование реализовано при поддержке факультета социальных наук, Национальный исследовательский университет «Высшая школа экономики».

Литература

1. Воронин И.А., Захаров И.М., Табуева А.О., Мерзон Л.А. (2020) Диффузная модель принятия решения: оценка скорости и точности ответов в задачах выбора из двух альтернатив в исследованиях когнитивных процессов и способностей. *Теоретическая и экспериментальная психология*, т. 13, № 2, сс. 6–23.

2. Шульц Д.П., Шульц С.Э. (1998) *История современной психологии*. СПб.: Евразия.
3. Ackerman T.A., Gierl M.J., Walker C.M. (2003) Using Multidimensional Item Response Theory to Evaluate Educational and Psychological Tests. *Educational Measurement: Issues and Practice*, vol. 22, no 3, pp. 37–51. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x
4. Baayen R.H., Milin P. (2010) Analyzing Reaction Times. *International Journal of Psychological Research*, vol. 3, no 2, pp. 12–28. https://doi.org/10.21500/20112084.807
5. Bindra D., Scheier I.H. (1954) The Relation between Psychometric and Experimental Research in Psychology. *American Psychologist,* vol. 9, no 2, pp. 69–71. https://doi.org/10.1037/h0062472
6. Birnbaum A. (1958) *On the Estimation of Mental Ability. Series Report no 15, Project no 7755–7723*. Texas: Randolph Air Force Base, TX USAF School of Aviation Medicine.
7. Bolsinova M., Tijmstra J. (2018) Improving Precision of Ability Estimation: Getting More from Response Times. *British Journal of Mathematical and Statistical Psychology*, vol. 71, no 1, pp. 13–38. https://doi.org/10.1111/bmsp.12104
8. Borsboom D. (2006) The Attack of the Psychometricians. *Psychometrika*, vol. 71, no 3, pp. 425–440. https://doi.org/10.1007/s11336-006-1447-6
9. Borsboom D., Kievit R.A., Cervone D., Hood S.B. (2009) The Two Disciplines of Scientific Psychology, or: The Disunity of Psychology as a Working Hypothesis. *Dynamic Process Methodology in the Social and Developmental Sciences* (eds J. Valsiner, P. Molenaar, M. Lyra, N. Chaudhary), New York, NY: Springer, pp. 67–97. https://doi.org/10.1007/978-0-387-95922-1_4
10. Braat M., Engelen J., van Gemert T., Verhaegh S. (2020) The Rise and Fall of Behaviorism: The Narrative and the Numbers. *History of Psychology*, vol. 23, no 3, pp. 252–280. https://doi.org/10.1037/hop0000146
11. Brauer M., Curtin J.J. (2018) Linear Mixed-Effects Models and the Analysis of Nonindependent Data: A Unified Framework to Analyze Categorical and Continuous Independent Variables that Vary Within-Subjects and/or Within-Items. *Psychological Methods*, vol. 23, no 3, pp. 389–411. https://doi.org/10.1037/met0000159
12. Brown J. (1992) *The Definition of a Profession: The Authority of Metaphor in the History of Intelligence Testing, 1890–1930*. Princeton, NJ: Princeton University.
13. Caruso J.C. (2004) A Comparison of the Reliabilities of Four Types of Difference Scores for Five Cognitive Assessment Batteries. *European Journal of Psychological Assessment*, vol. 20, no 3, pp. 166–171. https://doi.org/10.1027/1015-5759.20.3.166
14. Cattell J.M., Galton F. (1890) Mental Tests and Measurements. *Mind*, vol. os-XV, iss. 59, pp. 373–381. https://doi.org/10.1093/mind/os-XV.59.373
15. Corneille O., Mierop A., Unkelbach C. (2020) Repetition Increases Both the Perceived Truth and Fakeness of Information: An Ecological Account. *Cognition*, vol. 205, December, Article no 104470. https://doi.org/10.1016/j.cognition.2020.104470
16. Cronbach L.J. (1957) The Two Disciplines of Scientific Psychology. *American Psychologist,* vol. 12, no 11, pp. 671–684. https://doi.org/10.1037/h0043943
17. Cronbach L.J., Shavelson R.J. (2004) My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and Psychological Measurement*, vol. 64, no 3, pp. 391–418. https://doi.org/10.1177/0013164404266386
18. Cunnings I. (2012) An Overview of Mixed-Effects Statistical Models for Second Language Researchers. *Second Language Research*, vol. 28, no 3, pp. 369–382. https://doi.org/10.1177/0267658312443651

19. De Boeck P., Jeon M. (2019) An Overview of Models for Response Times and Processes in Cognitive Tests. *Frontiers in Psychology*, vol. 10, February, Article no 102. https://doi.org/10.3389/fpsyg.2019.00102
20. Dietrich J.F., Huber S., Klein E., Willmes K., Pixner S., Moeller K. (2016) A Systematic Investigation of Accuracy and Response Time Based Measures Used to Index ANS Acuity. *PLoS ONE*, vol. 11, no 9, Article no e0163076. https://doi.org/10.1371/journal.pone.0163076
21. Dietrich J.F., Huber S., Nuerk H.-C. (2015) Methodological Aspects to Be Considered When Measuring the Approximate Number System (ANS): A Research Review. *Frontiers in Psychology*, vol. 6, March, Article no 295. https://doi.org/10.3389/fpsyg.2015.00295
22. Dodonova Yu.A., Dodonov Yu.S. (2013) Faster on Easy Items, More Accurate on Difficult Ones: Cognitive Ability and Performance on a Task of Varying Difficulty. *Intelligence*, vol. 41, no 1, pp. 1–10. https://doi.org/10.1016/j.intell.2012.10.003
23. Draheim C., Mashburn C.A., Martin J.D., Engle R.W. (2019) Reaction Time in Differential and Developmental Research: A Review and Commentary on the Problems and Alternatives. *Psychological Bulletin*, vol. 145, no 5, pp. 508–535. https://doi.org/10.1037/bul0000192
24. Draheim C., Tsukahara J.S., Martin J.D., Mashburn C.A., Engle R.W. (2021) A Toolbox Approach to Improving the Measurement of Attention Control. *Journal of Experimental Psychology: General,* vol. 150, no 2, pp. 242–275. https://doi.org/10.1037/xge0000783
25. Drew T., Vogel E.K. (2008) Neural Measures of Individual Differences in Selecting and Tracking Multiple Moving Objects. *The Journal of Neuroscience*, vol. 28, no 16, pp. 4183–4191. https://doi.org/10.1523/JNEUROSCI.0556-08.2008
26. Dunn T.J., Baguley T., Brunsden V. (2014) From Alpha to Omega: A Practical Solution to the Pervasive Problem of Internal Consistency Estimation. *British Journal of Psychology*, vol. 105, no 3, pp. 399–412. https://doi.org/10.1111/bjop.12046
27. Edwards J.R. (2001) Ten Difference Score Myths. *Organizational Research Methods*, vol. 4, no 3, pp. 265–287. https://doi.org/10.1177/109442810143005
28. Eide P., Kemp A., Silberstein R.B., Nathan P.J., Stough C. (2002) Test-Retest Reliability of the Emotional Stroop Task: Examining the Paradox of Measurement Change. *The Journal of Psychology*, vol. 136, no 5, pp. 514–520. https://doi.org/10.1080/00223980209605547
29. Embretson S. (1994) Applications of Cognitive Design Systems to Test Development. *Cognitive Assessment: A Multidisciplinary Perspective* (ed. C.R. Reynolds), New York, NY: Springer Science+ Business Media, pp. 107–135. https://doi.org/10.1007/978-1-4757-9730-5_6
30. Embretson S., Gorin J. (2001) Improving Construct Validity with Cognitive Psychology Principles. *Journal of Educational Measurement*, vol. 38, no 4, pp. 343–368. https://doi.org/10.1111/j.1745-3984.2001.tb01131.x
31. Friedman N.P., Miyake A. (2017) Unity and Diversity of Executive Functions: Individual Differences as a Window on Cognitive Structure. *Cortex*, no 86, pp. 186–204. https://doi.org/10.1016/j.cortex.2016.04.023
32. Galton F. (1879) Psychometric Experiments. *Brain*, vol. 2, no 2, pp. 149–162. https://doi.org/10.1093/brain/2.2.149
33. Galton F. (1883) *Inquiries into Human Faculty and Its Development.* New York, NY: MacMillan. https://doi.org/10.1037/14178-000
34. Gevins A., Smith M.E. (2000) Neurophysiological Measures of Working Memory and Individual Differences in Cognitive Ability and Cognitive Style. *Cerebral Cortex*, vol. 10, no 9, pp. 829–839. https://doi.org/10.1093/cercor/10.9.829
35. Glaser R. (1981) The Future of Testing: A Research Agenda for Cognitive Psychology and Psychometrics. *American Psychologist*, vol. 36, no 9, pp. 923–936. https://doi.org/10.1037/0003-066X.36.9.923

36. Goldhammer F., Naumann J., Stelter A., Tóth K., Rölke H., Klieme E. (2014) The Time on Task Effect in Reading and Problem Solving Is Moderated by Task Difficulty and Skill: Insights from a Computer-Based Large-Scale Assessment. *Journal of Educational Psychology,* vol. 106, no 3, pp. 608–626. https://doi.org/10.1037/a0034716
37. Goldstein H. (2012) Francis Galton, Measurement, Psychometrics and Social Progress. *Assessment in Education: Principles, Policy & Practice*, vol. 19, no 2, pp. 147–158. https://doi.org/10.1080/0969594X.2011.614220
38. Goodhew S.C., Edwards M. (2019) Translating Experimental Paradigms into Individual-Differences Research: Contributions, Challenges, and Practical Recommendations. *Consciousness and Cognition*, vol. 69, January, pp. 14–25. https://doi.org/10.1016/j.concog.2019.01.008
39. Hambleton R.K. (1989) Principles and Selected Applications of Item Response Theory. *Educational Measurement* (ed. R.L. Linn), New York, NY: Macmillan Publishing Co, Inc; American Council on Education, pp. 147–200.
40. Heathcote A., Popiel S.J., Mewhort D.J. (1991) Analysis of Response Time Distributions: An Example Using the Stroop Task. *Psychological Bulletin,* vol. 109, no 2, pp. 340–347. https://doi.org/10.1037/0033-2909.109.2.340
41. Heck D.W., Erdfelder E. (2016) Extending Multinomial Processing Tree Models to Measure the Relative Speed of Cognitive Processes. *Psychonomic Bulletin & Review*, vol. 23, no 5, pp. 1440–1465. https://doi.org/10.3758/s13423-016-1025-6
42. Hedge C., Powell G., Sumner P. (2018) The Reliability Paradox: Why Robust Cognitive Tasks Do Not Produce Reliable Individual Differences. *Behavior Research Methods*, vol. 50, no 3, pp. 1166–1186. https://doi.org/10.3758/s13428-017-0935-1
43. Hepp H.H., Maier S., Hermle L., Spitzer M. (1996) The Stroop Effect in Schizophrenic Patients. *Schizophrenia Research*, vol. 22, no 3, pp. 187–195. https://doi.org/10.1016/S0920-9964(96)00080-1
44. Jensen A.R. (2006) *Clocking the Mind: Mental Chronometry and Individual Differences*. Amsterdam: Elsevier.
45. Johnson R.C., McClearn G.E., Yuen S., Nagoshi C.T., Ahern F.M., Cole R.E. (1985) Galton's Data a Century Later. *American Psychologist,* vol. 40, no 8, pp. 875–892. https://doi.org/10.1037/0003-066X.40.8.875
46. Kane M.J., Hambrick D.Z., Tuholski S.W., Wilhelm O., Payne T.W., Engle R.W. (2004) The Generality of Working Memory Capacity: A Latent-Variable Approach to Verbal and Visuospatial Memory Span and Reasoning. *Journal of Experimental Psychology: General,* vol. 133, no 2, pp. 189–217. https://doi.org/10.1037/0096-3445.133.2.189
47. Kievit R.A., Frankenhuis W.E., Waldorp L.J., Borsboom D. (2013) Simpson's Paradox in Psychological Science: A Practical Guide. *Frontiers in Psychology*, vol. 4, August, Article no 513. https://doi.org/10.3389/fpsyg.2013.00513
48. Kim E.S., Yoon M. (2011) Testing Measurement Invariance: A Comparison of Multiple-Group Categorical CFA and IRT. *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 18, no 2, pp. 212–228. https://doi.org/10.1080/10705511.2011.557337
49. Kim S., Feldt L.S. (2010) The Estimation of the IRT Reliability Coefficient and Its Lower and Upper Bounds, with Comparisons to CTT Reliability Statistics. *Asia Pacific Education Review*, vol. 11, no 2, pp. 179–188. https://doi.org/10.1007/s12564-009-9062-8
50. Kleka P., Soroko E. (2018) How to Avoid the Sins of Questionnaire Abridgement — Guideline. *PsyArXiv*. https://doi.org/10.31234/osf.io/8jg9u
51. Lamiell J.T. (1992) Personality Psychology and the Second Cognitive Revolution. *American Behavioral Scientist*, vol. 36, no 1, pp. 88–101. https://doi.org/10.1177/0002764292036001008

52. Lazarsfeld P.F. (1950) The Logical and Mathematical Foundation of Latent Structure Analysis. *Studies in Social Psychology in World War II. Vol. IV: Measurement and Prediction* (eds S. Stouffer, L. Guttman, E. Suchman, P. Lazarsfeld), Princeton: Princeton University, pp. 362–412.
53. Leitgöb H., Seddig D., Asparouhov T., Behr D., Davidov E., de Roover K. et al. (2023) Measurement Invariance in the Social Sciences: Historical Development, Methodological Challenges, State of the Art, and Future Perspectives. *Social Science Research*, vol. 110, January, Article no 102805. https://doi.org/10.1016/j.ssresearch.2022.102805
54. Linden van der W.J. (2009) Conceptual Issues in Response-Time Modeling. *Journal of Educational Measurement*, vol. 46, no 3, pp. 247–272. https://doi.org/10.1111/j.1745-3984.2009.00080.x
55. Linden van der W.J. (2007) A Hierarchical Framework for Modeling Speed and Accuracy on Test Items. *Psychometrika*, vol. 72, no 3, pp. 287–308. https://doi.org/10.1007/s11336-006-1478-z
56. Lo S., Andrews S. (2015) To Transform Or Not to Transform: Using Generalized Linear Mixed Models to Analyse Reaction Time Data. *Frontiers in Psychology*, vol. 6, August, Article no 1171. https://doi.org/10.3389/fpsyg.2015.01171
57. Lord F.M. (1953) The Relation of Test Score to the Trait Underlying the Test. *Educational and Psychological Measurement*, vol. 13, no 4, pp. 517–549. https://doi.org/10.1177/001316445301300401
58. Ludlow L.H. (1998) Galton: The First Psychometrician. *Popular Measurement*, vol. 1, no 1, pp. 13–14.
59. Maas van der H.L.J., Molenaar D., Maris G., Kievit R.A., Borsboom D. (2011) Cognitive Psychology Meets Psychometric Theory: On the Relation between Process Models for Decision Making and Latent Variable Models for Individual Differences. *Psychological Review,* vol. 118, no 2, pp. 339–356. https://doi.org/10.1037/a0022749
60. Meade A.W., Lautenschlager G.J. (2004) A Comparison of Item Response Theory and Confirmatory Factor Analytic Methodologies for Establishing Measurement Equivalence/Invariance. *Organizational Research Methods*, vol. 7, no 4, pp. 361–388. https://doi.org/10.1177/1094428104268027
61. Miller G.A. (2003) The Cognitive Revolution: A Historical Perspective. *Trends in Cognitive Sciences*, vol. 7, no 3, pp. 141–144. https://doi.org/10.1016/S1364-6613(03)00029-9
62. Molenaar D., Tuerlinckx F., van der Maas H.L. (2015) A Bivariate Generalized Linear Item Response Theory Modeling Framework to the Analysis of Responses and Response Times. *Multivariate Behavioral Research*, vol. 50, no 1, pp. 56–74. https://doi.org/10.1080/00273171.2014.962684
63. Molenaar P.C.M., Beltz A.M. (2020) Modeling the Individual: Bridging Nomothetic and Idiographic Levels of Analysis. *The Cambridge Handbook of Research Methods in Clinical Psychology* (eds A.G.C. Wright, M.N. Hallquist), Cambridge: Cambridge University, pp. 327–336. https://doi.org/10.1017/9781316995808.031
64. Moore J. (1999) The Basic Principles of Behaviorism. *The Philosophical Legacy of Behaviorism* (ed. B.A. Thyer), Dordrecht: Springer Science+Business Media, pp. 41–68. https://doi.org/10.1007/978-94-015-9247-5_2
65. Moore J. (1996) On the Relation between Behaviorism and Cognitive Psychology. *The Journal of Mind and Behavior*, vol. 17, no 4, pp. 345–367.
66. Morís Fernández L., Vadillo M.A. (2020) Flexibility in Reaction Time Analysis: Many Roads to a False Positive? *Royal Society Open Science*, vol. 7, no 2, Article no 190831. https://doi.org/10.1098/rsos.190831
67. Parsons S., Kruijt A.-W., Fox E. (2019) Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measure-

ments. *Advances in Methods and Practices in Psychological Science*, vol. 2, no 4, pp. 378–395. https://doi.org/10.1177/2515245919879695
68. Passolunghi M.C., Siegel L.S. (2001) Short-Term Memory, Working Memory, and Inhibitory Control in Children with Difficulties in Arithmetic Problem Solving. *Journal of Experimental Child Psychology*, vol. 80, no 1, pp. 44–57. https://doi.org/10.1006/jecp.2000.2626
69. Pronk T., Hirst R.J., Wiers R.W., Murre J.M.J. (2023) Can We Measure Individual Differences in Cognitive Measures Reliably via Smartphones? A Comparison of the Flanker Effect across Device Types and Samples. *Behavior Research Methods*, vol. 55, no 4, pp. 1641–1652. https://doi.org/10.3758/s13428-022-01885-6
70. Putnick D.L., Bornstein M.H. (2016) Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research. *Developmental Review*, vol. 41, June, pp. 71–90. https://doi.org/10.1016/j.dr.2016.06.004
71. Rasch G. (1968) A Mathematical Theory of Objectivity and Its Consequences for Model Construction. *Report from European Meeting on Statistics, Econometrics and Management Sciences, Amsterdam*.
72. Rasch G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danmarks Paedagogiske Institut.
73. Ratcliff R., Smith P.L., McKoon G. (2015) Modeling Regularities in Response Time and Accuracy Data with the Diffusion Model. *Current Directions in Psychological Science*, vol. 24, no 6, pp. 458–470. https://doi.org/10.1177/0963721415596228
74. Rey-Mermet A., Gade M., Oberauer K. (2018) Should We Stop Thinking about Inhibition? Searching for Individual and Age Differences in Inhibition Ability. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* vol. 44, no 4, pp. 501–526. https://doi.org/10.1037/xlm0000450
75. Rouder J.N., Haaf J.M. (2019) A Psychometrics of Individual Differences in Experimental Tasks. *Psychonomic Bulletin & Review*, vol. 26, no 2, pp. 452–467. https://doi.org/10.3758/s13423-018-1558-y
76. Rouder J., Kumar A., Haaf J.M. (2019) Why Most Studies of Individual Differences with Inhibition Tasks Are Bound to Fail. *PsyArXiv*. https://doi.org/10.31234/osf.io/3cjr5
77. Rousselet G.A., Wilcox R.R. (2020) Reaction Times and Other Skewed Distributions: Problems with the Mean and the Median. *Meta-Psychology*, vol. 4, Article no MP.2019.1630. https://doi.org/10.15626/MP.2019.1630
78. Royer J.M. (ed.) (2006) *The Cognitive Revolution on Educational Psychology*: *Current Perspectives on Cognition, Learning and Instruction.* Charlotte, NC: Information Age Publishing.
79. Saville C.W.N., Pawling R., Trullinger M., Daley D., Intriligator J., Klein C. (2011) On the Stability of Instability: Optimising the Reliability of Intra-Subject Variability of Reaction Times. *Personality and Individual Differences*, vol. 51, no 2, pp. 148–153. https://doi.org/10.1016/j.paid.2011.03.034
80. Scarpina F., Tagini S. (2017) The Stroop Color and Word Test. *Frontiers in Psychology*, vol. 8, April, Article no 557. https://doi.org/10.3389/fpsyg.2017.00557
81. Schmidt J.R., Besner D. (2008) The Stroop Effect: Why Proportion Congruent Has Nothing to Do with Congruency and Everything to Do with Contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* vol. 34, no 3, pp. 514–523. https://doi.org/10.1037/0278-7393.34.3.514
82. Schramm P., Rouder J.N. (2019) Are Reaction Time Transformations Really Beneficial? *PsyArXiv*. https://doi.org/10.31234/osf.io/9ksa6
83. Shichel I., Tzelgov J. (2018) Modulation of Conflicts in the Stroop Effect. *Acta Psychologica*, vol. 189, pp. 93–102. https://doi.org/10.1016/j.actpsy.2017.10.007

84. Simon H.A. (1979) Information Processing Models of Cognition. *Annual Review of Psychology*, vol. 30, no 1, pp. 363–396. https://doi.org/10.1146/annurev.ps.30.020179.002051
85. Smedt de B., Gilmore C.K. (2011) Defective Number Module or Impaired Access? Numerical Magnitude Processing in First Graders with Mathematical Difficulties. *Journal of Experimental Child Psychology*, vol. 108, no 2, pp. 278–292. https://doi.org/10.1016/j.jecp.2010.09.003
86. Sokal M.M. (1987) Psychological Testing and American Society 1890–1930. New Brunswick: Rutgers University.
87. Speelman C.P., McGann M. (2013) How Mean is the Mean? *Frontiers in Psychology*, vol. 4, July, Article no 451. https://doi.org/10.3389/fpsyg.2013.00451
88. Spence R., Owens M., Goodyer I. (2012) Item Response Theory and Validity of the NEO-FFI in Adolescents. *Personality and Individual Differences*, vol. 53, no 6, pp. 801–807. https://doi.org/10.1016/j.paid.2012.06.002
89. Stahl C., Voss A., Schmitz F., Nuszbaum M., Tüscher O., Lieb K., Klauer K.C. (2014) Behavioral Components of Impulsivity. *Journal of Experimental Psychology: General,* vol. 143, no 2, pp. 850–886. https://doi.org/10.1037/a0033981
90. Sternberg R.J. (1981) Testing and Cognitive Psychology. *American Psychologist*, vol. 36, no 10, pp. 1181–1189. https://doi.org/10.1037/0003-066X.36.10.1181
91. Stroop J.R. (1935) Studies of Interference in Serial Verbal Reactions. *Journal of Experimental Psychology,* vol. 18, no 6, pp. 643–662. https://doi.org/10.1037/h0054651
92. Tavakol M., Dennick R. (2011) Making Sense of Cronbach's Alpha. *International Journal of Medical Education*, vol. 2, June, pp. 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd
93. Terman L.M. (1924) The Mental Test as a Psychological Method. *Psychological Review,* vol. 31, no 2, pp. 93–117. https://doi.org/10.1037/h0070938
94. Troyer A.K., Leach L., Strauss E. (2006) Aging and Response Inhibition: Normative Data for the Victoria Stroop Test. *Aging, Neuropsychology, and Cognition*, vol. 13, no 1, pp. 20–35. https://doi.org/10.1080/138255890968187
95. Tuerlinckx F., De Boeck P.D. (2005) Two Interpretations of the Discrimination Parameter. *Psychometrika*, vol. 70, no 4, pp. 629–650. https://doi.org/10.1007/s11336-000-0810-3
96. Watrin J.P., Darwich R. (2012) On Behaviorism in the Cognitive Revolution: Myth and Reactions. *Review of General Psychology*, vol. 16, no 3, pp. 269–282. https://doi.org/10.1037/a0026766
97. Watson J.B. (1913) Psychology as the Behaviorist Views It. *Psychological Review*, vol. 20, no 2, pp. 158–177. https://doi.org/10.1037/h0074428
98. Wells F.L. (1912) The Relation of Practice to Individual Differences. *The American Journal of Psychology*, vol. 23, no 1, pp. 75–88. https://doi.org/10.2307/1413115
99. Whelan R. (2008) Effective Analysis of Reaction Time Data. *The Psychological Record*, vol. 58, no 3, pp. 475–482. https://doi.org/10.1007/BF03395630
100. Wicherts J.M. (2016) The Importance of Measurement Invariance in Neurocognitive Ability Testing. *The Clinical Neuropsychologist*, vol. 30, no 7, pp. 1006–1016. https://doi.org/10.1080/13854046.2016.1205136
101. Wijsen L.D., Borsboom D., Alexandrova A. (2022) Values in Psychometrics. *Perspectives on Psychological Science*, vol. 17, no 3, pp. 788–804. https://doi.org/10.1177/17456916211014183
102. Willoughby M.T., Wirth R.J., Blair C.B. (2012) Executive Function in Early Childhood: Longitudinal Measurement Invariance and Developmental Change. *Psychological Assessment*, vol. 24, no 2, pp. 418–431. https://doi.org/10.1037/a0025779
103. Wissler C. (1901) The Correlation of Mental and Physical Tests. *The Psychological Review: Monograph Supplements,* vol. 3, no 6, pp. i–62. https://doi.org/10.1037/h0092995

References

Ackerman T.A., Gierl M.J., Walker C.M. (2003) Using Multidimensional Item Response Theory to Evaluate Educational and Psychological Tests. *Educational Measurement: Issues and Practice*, vol. 22, no 3, pp. 37–51. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x

Baayen R.H., Milin P. (2010) Analyzing Reaction Times. *International Journal of Psychological Research*, vol. 3, no 2, pp. 12–28. https://doi.org/10.21500/20112084.807

Bindra D., Scheier I.H. (1954) The Relation between Psychometric and Experimental Research in Psychology. *American Psychologist,* vol. 9, no 2, pp. 69–71. https://doi.org/10.1037/h0062472

Birnbaum A. (1958) *On the Estimation of Mental Ability. Series Report no 15, Project no 7755–7723*. Texas: Randolph Air Force Base, TX USAF School of Aviation Medicine.

Bolsinova M., Tijmstra J. (2018) Improving Precision of Ability Estimation: Getting More from Response Times. *British Journal of Mathematical and Statistical Psychology*, vol. 71, no 1, pp. 13–38. https://doi.org/10.1111/bmsp.12104

Borsboom D. (2006) The Attack of the Psychometricians. *Psychometrika*, vol. 71, no 3, pp. 425–440. https://doi.org/10.1007/s11336-006-1447-6

Borsboom D., Kievit R.A., Cervone D., Hood S.B. (2009) The Two Disciplines of Scientific Psychology, or: The Disunity of Psychology as a Working Hypothesis. *Dynamic Process Methodology in the Social and Developmental Sciences* (eds J. Valsiner, P. Molenaar, M. Lyra, N. Chaudhary), New York, NY: Springer, pp. 67–97. https://doi.org/10.1007/978-0-387-95922-1_4

Braat M., Engelen J., van Gemert T., Verhaegh S. (2020) The Rise and Fall of Behaviorism: The Narrative and the Numbers. *History of Psychology*, vol. 23, no 3, pp. 252–280. https://doi.org/10.1037/hop0000146

Brauer M., Curtin J.J. (2018) Linear Mixed-Effects Models and the Analysis of Nonindependent Data: A Unified Framework to Analyze Categorical and Continuous Independent Variables that Vary Within-Subjects and/or Within-Items. *Psychological Methods*, vol. 23, no 3, pp. 389–411. https://doi.org/10.1037/met0000159

Brown J. (1992) *The Definition of a Profession: The Authority of Metaphor in the History of Intelligence Testing, 1890–1930*. Princeton, NJ: Princeton University.

Caruso J.C. (2004) A Comparison of the Reliabilities of Four Types of Difference Scores for Five Cognitive Assessment Batteries. *European Journal of Psychological Assessment*, vol. 20, no 3, pp. 166–171. https://doi.org/10.1027/1015-5759.20.3.166

Cattell J.M., Galton F. (1890) Mental Tests and Measurements. *Mind*, vol. os-XV, iss. 59, pp. 373–381. https://doi.org/10.1093/mind/os-XV.59.373

Corneille O., Mierop A., Unkelbach C. (2020) Repetition Increases Both the Perceived Truth and Fakeness of Information: An Ecological Account. *Cognition*, vol. 205, December, Article no 104470. https://doi.org/10.1016/j.cognition.2020.104470

Cronbach L.J. (1957) The Two Disciplines of Scientific Psychology. *American Psychologist,* vol. 12, no 11, pp. 671–684. https://doi.org/10.1037/h0043943

Cronbach L.J., Shavelson R.J. (2004) My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and Psychological Measurement*, vol. 64, no 3, pp. 391–418. https://doi.org/10.1177/0013164404266386

Cunnings I. (2012) An Overview of Mixed-Effects Statistical Models for Second Language Researchers. *Second Language Research*, vol. 28, no 3, pp. 369–382. https://doi.org/10.1177/0267658312443651

De Boeck P., Jeon M. (2019) An Overview of Models for Response Times and Processes in Cognitive Tests. *Frontiers in Psychology*, vol. 10, February, Article no 102. https://doi.org/10.3389/fpsyg.2019.00102

Dietrich J.F., Huber S., Klein E., Willmes K., Pixner S., Moeller K. (2016) A Systematic Investigation of Accuracy and Response Time Based Measures Used to Index ANS Acuity. *PLoS ONE*, vol. 11, no 9, Article no e0163076. https://doi.org/10.1371/journal.pone.0163076

Dietrich J.F., Huber S., Nuerk H.-C. (2015) Methodological Aspects to Be Considered When Measuring the Approximate Number System (ANS): A Research Review. *Frontiers in Psychology*, vol. 6, March, Article no 295. https://doi.org/10.3389/fpsyg.2015.00295

Dodonova Yu.A., Dodonov Yu.S. (2013) Faster on Easy Items, More Accurate on Difficult Ones: Cognitive Ability and Performance on a Task of Varying Difficulty. *Intelligence*, vol. 41, no 1, pp. 1–10. https://doi.org/10.1016/j.intell.2012.10.003

Draheim C., Mashburn C.A., Martin J.D., Engle R.W. (2019) Reaction Time in Differential and Developmental Research: A Review and Commentary on the Problems and Alternatives. *Psychological Bulletin*, vol. 145, no 5, pp. 508–535. https://doi.org/10.1037/bul0000192

Draheim C., Tsukahara J.S., Martin J.D., Mashburn C.A., Engle R.W. (2021) A Toolbox Approach to Improving the Measurement of Attention Control. *Journal of Experimental Psychology: General,* vol. 150, no 2, pp. 242–275. https://doi.org/10.1037/xge0000783

Drew T., Vogel E.K. (2008) Neural Measures of Individual Differences in Selecting and Tracking Multiple Moving Objects. *The Journal of Neuroscience*, vol. 28, no 16, pp. 4183–4191. https://doi.org/10.1523/JNEUROSCI.0556-08.2008

Dunn T.J., Baguley T., Brunsden V. (2014) From Alpha to Omega: A Practical Solution to the Pervasive Problem of Internal Consistency Estimation. *British Journal of Psychology*, vol. 105, no 3, pp. 399–412. https://doi.org/10.1111/bjop.12046

Edwards J.R. (2001) Ten Difference Score Myths. *Organizational Research Methods*, vol. 4, no 3, pp. 265–287. https://doi.org/10.1177/109442810143005

Eide P., Kemp A., Silberstein R.B., Nathan P.J., Stough C. (2002) Test-Retest Reliability of the Emotional Stroop Task: Examining the Paradox of Measurement Change. *The Journal of Psychology*, vol. 136, no 5, pp. 514–520. https://doi.org/10.1080/00223980209605547

Embretson S. (1994) Applications of Cognitive Design Systems to Test Development. *Cognitive Assessment: A Multidisciplinary Perspective* (ed. C.R. Reynolds), New York, NY: Springer Science+ Business Media, pp. 107–135. https://doi.org/10.1007/978-1-4757-9730-5_6

Embretson S., Gorin J. (2001) Improving Construct Validity with Cognitive Psychology Principles. *Journal of Educational Measurement*, vol. 38, no 4, pp. 343–368. https://doi.org/10.1111/j.1745-3984.2001.tb01131.x

Friedman N.P., Miyake A. (2017) Unity and Diversity of Executive Functions: Individual Differences as a Window on Cognitive Structure. *Cortex*, no 86, pp. 186–204. https://doi.org/10.1016/j.cortex.2016.04.023

Galton F. (1879) Psychometric Experiments. *Brain*, vol. 2, no 2, pp. 149–162. https://doi.org/10.1093/brain/2.2.149

Galton F. (1883) *Inquiries into Human Faculty and Its Development.* New York, NY: MacMillan. https://doi.org/10.1037/14178-000

Gevins A., Smith M.E. (2000) Neurophysiological Measures of Working Memory and Individual Differences in Cognitive Ability and Cognitive Style. *Cerebral Cortex*, vol. 10, no 9, pp. 829–839. https://doi.org/10.1093/cercor/10.9.829

Glaser R. (1981) The Future of Testing: A Research Agenda for Cognitive Psychology and Psychometrics. *American Psychologist*, vol. 36, no 9, pp. 923–936. https://doi.org/10.1037/0003-066X.36.9.923

Goldhammer F., Naumann J., Stelter A., Tóth K., Rölke H., Klieme E. (2014) The Time on Task Effect in Reading and Problem Solving Is Moderated by Task Difficulty and Skill: Insights from a Computer-Based Large-Scale Assessment. *Journal*

of Educational Psychology, vol. 106, no 3, pp. 608–626. https://doi.org/10.1037/a0034716

Goldstein H. (2012) Francis Galton, Measurement, Psychometrics and Social Progress. *Assessment in Education: Principles, Policy & Practice*, vol. 19, no 2, pp. 147–158. https://doi.org/10.1080/0969594X.2011.614220

Goodhew S.C., Edwards M. (2019) Translating Experimental Paradigms into Individual-Differences Research: Contributions, Challenges, and Practical Recommendations. *Consciousness and Cognition*, vol. 69, January, pp. 14–25. https://doi.org/10.1016/j.concog.2019.01.008

Hambleton R.K. (1989) Principles and Selected Applications of Item Response Theory. *Educational Measurement* (ed. R.L. Linn), New York, NY: Macmillan Publishing Co, Inc; American Council on Education, pp. 147–200.

Heathcote A., Popiel S.J., Mewhort D.J. (1991) Analysis of Response Time Distributions: An Example Using the Stroop Task. *Psychological Bulletin,* vol. 109, no 2, pp. 340–347. https://doi.org/10.1037/0033-2909.109.2.340

Heck D.W., Erdfelder E. (2016) Extending Multinomial Processing Tree Models to Measure the Relative Speed of Cognitive Processes. *Psychonomic Bulletin & Review*, vol. 23, no 5, pp. 1440–1465. https://doi.org/10.3758/s13423-016-1025-6

Hedge C., Powell G., Sumner P. (2018) The Reliability Paradox: Why Robust Cognitive Tasks Do Not Produce Reliable Individual Differences. *Behavior Research Methods*, vol. 50, no 3, pp. 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hepp H.H., Maier S., Hermle L., Spitzer M. (1996) The Stroop Effect in Schizophrenic Patients. *Schizophrenia Research*, vol. 22, no 3, pp. 187–195. https://doi.org/10.1016/S0920-9964(96)00080-1

Jensen A.R. (2006) *Clocking the Mind: Mental Chronometry and Individual Differences*. Amsterdam: Elsevier.

Johnson R.C., McClearn G.E., Yuen S., Nagoshi C.T., Ahern F.M., Cole R.E. (1985) Galton's Data a Century Later. *American Psychologist,* vol. 40, no 8, pp. 875–892. https://doi.org/10.1037/0003-066X.40.8.875

Kane M.J., Hambrick D.Z., Tuholski S.W., Wilhelm O., Payne T.W., Engle R.W. (2004) The Generality of Working Memory Capacity: A Latent-Variable Approach to Verbal and Visuospatial Memory Span and Reasoning. *Journal of Experimental Psychology: General,* vol. 133, no 2, pp. 189–217. https://doi.org/10.1037/0096-3445.133.2.189

Kievit R.A., Frankenhuis W.E., Waldorp L.J., Borsboom D. (2013) Simpson's Paradox in Psychological Science: A Practical Guide. *Frontiers in Psychology*, vol. 4, August, Article no 513. https://doi.org/10.3389/fpsyg.2013.00513

Kim E.S., Yoon M. (2011) Testing Measurement Invariance: A Comparison of Multiple-Group Categorical CFA and IRT. *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 18, no 2, pp. 212–228. https://doi.org/10.1080/10705511.2011.557337

Kim S., Feldt L.S. (2010) The Estimation of the IRT Reliability Coefficient and Its Lower and Upper Bounds, with Comparisons to CTT Reliability Statistics. *Asia Pacific Education Review*, vol. 11, no 2, pp. 179–188. https://doi.org/10.1007/s12564-009-9062-8

Kleka P., Soroko E. (2018) How to Avoid the Sins of Questionnaire Abridgement — Guideline. *PsyArXiv*. https://doi.org/10.31234/osf.io/8jg9u

Lamiell J.T. (1992) Personality Psychology and the Second Cognitive Revolution. *American Behavioral Scientist*, vol. 36, no 1, pp. 88–101. https://doi.org/10.1177/0002764292036001008

Lazarsfeld P.F. (1950) The Logical and Mathematical Foundation of Latent Structure Analysis. *Studies in Social Psychology in World War II. Vol. IV: Measurement and Prediction* (eds S. Stouffer, L. Guttman, E. Suchman, P. Lazarsfeld), Princeton: Princeton University, pp. 362–412.

Leitgöb H., Seddig D., Asparouhov T., Behr D., Davidov E., de Roover K. et al. (2023) Measurement Invariance in the Social Sciences: Historical Development, Methodological Challenges, State of the Art, and Future Perspectives. *Social Science Research*, vol. 110, January, Article no 102805. https://doi.org/10.1016/j.ssresearch.2022.102805

Linden van der W.J. (2009) Conceptual Issues in Response-Time Modeling. *Journal of Educational Measurement*, vol. 46, no 3, pp. 247–272. https://doi.org/10.1111/j.1745-3984.2009.00080.x

Linden van der W.J. (2007) A Hierarchical Framework for Modeling Speed and Accuracy on Test Items. *Psychometrika*, vol. 72, no 3, pp. 287–308. https://doi.org/10.1007/s11336-006-1478-z

Lo S., Andrews S. (2015) To Transform Or Not to Transform: Using Generalized Linear Mixed Models to Analyse Reaction Time Data. *Frontiers in Psychology*, vol. 6, August, Article no 1171. https://doi.org/10.3389/fpsyg.2015.01171

Lord F.M. (1953) The Relation of Test Score to the Trait Underlying the Test. *Educational and Psychological Measurement*, vol. 13, no 4, pp. 517–549. https://doi.org/10.1177/001316445301300401

Ludlow L.H. (1998) Galton: The First Psychometrician. *Popular Measurement*, vol. 1, no 1, pp. 13–14.

Maas van der H.L.J., Molenaar D., Maris G., Kievit R.A., Borsboom D. (2011) Cognitive Psychology Meets Psychometric Theory: On the Relation between Process Models for Decision Making and Latent Variable Models for Individual Differences. *Psychological Review,* vol. 118, no 2, pp. 339–356. https://doi.org/10.1037/a0022749

Meade A.W., Lautenschlager G.J. (2004) A Comparison of Item Response Theory and Confirmatory Factor Analytic Methodologies for Establishing Measurement Equivalence/Invariance. *Organizational Research Methods*, vol. 7, no 4, pp. 361–388. https://doi.org/10.1177/1094428104268027

Miller G.A. (2003) The Cognitive Revolution: A Historical Perspective. *Trends in Cognitive Sciences*, vol. 7, no 3, pp. 141–144. https://doi.org/10.1016/S1364-6613(03)00029-9

Molenaar D., Tuerlinckx F., van der Maas H.L. (2015) A Bivariate Generalized Linear Item Response Theory Modeling Framework to the Analysis of Responses and Response Times. *Multivariate Behavioral Research*, vol. 50, no 1, pp. 56–74. https://doi.org/10.1080/00273171.2014.962684

Molenaar P.C.M., Beltz A.M. (2020) Modeling the Individual: Bridging Nomothetic and Idiographic Levels of Analysis. *The Cambridge Handbook of Research Methods in Clinical Psychology* (eds A.G.C. Wright, M.N. Hallquist), Cambridge: Cambridge University, pp. 327–336. https://doi.org/10.1017/9781316995808.031

Moore J. (1999) The Basic Principles of Behaviorism. *The Philosophical Legacy of Behaviorism* (ed. B.A. Thyer), Dordrecht: Springer Science+Business Media, pp. 41–68. https://doi.org/10.1007/978-94-015-9247-5_2

Moore J. (1996) On the Relation between Behaviorism and Cognitive Psychology. *The Journal of Mind and Behavior*, vol. 17, no 4, pp. 345–367.

Morís Fernández L., Vadillo M.A. (2020) Flexibility in Reaction Time Analysis: Many Roads to a False Positive? *Royal Society Open Science*, vol. 7, no 2, Article no 190831. https://doi.org/10.1098/rsos.190831

Parsons S., Kruijt A.-W., Fox E. (2019) Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. *Advances in Methods and Practices in Psychological Science*, vol. 2, no 4, pp. 378–395. https://doi.org/10.1177/2515245919879695

Passolunghi M.C., Siegel L.S. (2001) Short-Term Memory, Working Memory, and Inhibitory Control in Children with Difficulties in Arithmetic Problem Solving. *Journal of Experimental Child Psychology*, vol. 80, no 1, pp. 44–57. https://doi.org/10.1006/jecp.2000.2626

Pronk T., Hirst R.J., Wiers R.W., Murre J.M.J. (2023) Can We Measure Individual Differences in Cognitive Measures Reliably via Smartphones? A Comparison of the Flanker Effect across Device Types and Samples. Behavior Research Methods, vol. 55, no 4, pp. 1641–1652. https://doi.org/10.3758/s13428-022-01885-6

Putnick D.L., Bornstein M.H. (2016) Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research. *Developmental Review*, vol. 41, June, pp. 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Rasch G. (1968) A Mathematical Theory of Objectivity and Its Consequences for Model Construction. *Report from European Meeting on Statistics, Econometrics and Management Sciences, Amsterdam*.

Rasch G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danmarks Paedagogiske Institut.

Ratcliff R., Smith P.L., McKoon G. (2015) Modeling Regularities in Response Time and Accuracy Data with the Diffusion Model. *Current Directions in Psychological Science*, vol. 24, no 6, pp. 458–470. https://doi.org/10.1177/0963721415596228

Rey-Mermet A., Gade M., Oberauer K. (2018) Should We Stop Thinking about Inhibition? Searching for Individual and Age Differences in Inhibition Ability. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* vol. 44, no 4, pp. 501–526. https://doi.org/10.1037/xlm0000450

Rouder J.N., Haaf J.M. (2019) A Psychometrics of Individual Differences in Experimental Tasks. *Psychonomic Bulletin & Review*, vol. 26, no 2, pp. 452–467. https://doi.org/10.3758/s13423-018-1558-y

Rouder J., Kumar A., Haaf J.M. (2019) Why Most Studies of Individual Differences with Inhibition Tasks Are Bound to Fail. *PsyArXiv*. https://doi.org/10.31234/osf.io/3cjr5

Rousselet G.A., Wilcox R.R. (2020) Reaction Times and Other Skewed Distributions: Problems with the Mean and the Median. *Meta-Psychology*, vol. 4, Article no MP.2019.1630. https://doi.org/10.15626/MP.2019.1630

Royer J.M. (ed.) (2006) *The Cognitive Revolution on Educational Psychology*: *Current Perspectives on Cognition, Learning and Instruction.* Charlotte, NC: Information Age Publishing.

Saville C.W.N., Pawling R., Trullinger M., Daley D., Intriligator J., Klein C. (2011) On the Stability of Instability: Optimising the Reliability of Intra-Subject Variability of Reaction Times. *Personality and Individual Differences*, vol. 51, no 2, pp. 148–153. https://doi.org/10.1016/j.paid.2011.03.034

Scarpina F., Tagini S. (2017) The Stroop Color and Word Test. *Frontiers in Psychology*, vol. 8, April, Article no 557. https://doi.org/10.3389/fpsyg.2017.00557

Schmidt J.R., Besner D. (2008) The Stroop Effect: Why Proportion Congruent Has Nothing to Do with Congruency and Everything to Do with Contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* vol. 34, no 3, pp. 514–523. https://doi.org/10.1037/0278-7393.34.3.514

Schramm P., Rouder J.N. (2019) Are Reaction Time Transformations Really Beneficial? *PsyArXiv*. https://doi.org/10.31234/osf.io/9ksa6

Schultz D.P., Schultz S.E. (1998) *Istoriya sovremennoy psikhologii* [A History of Modern Psychology]. Saint-Petersburg: Evraziya.

Shichel I., Tzelgov J. (2018) Modulation of Conflicts in the Stroop Effect. *Acta Psychologica*, 189, pp. 93–102. https://doi.org/10.1016/j.actpsy.2017.10.007

Simon H.A. (1979) Information Processing Models of Cognition. *Annual Review of Psychology*, vol. 30, no 1, pp. 363–396. https://doi.org/10.1146/annurev.ps.30.020179.002051

Smedt de B., Gilmore C.K. (2011) Defective Number Module or Impaired Access? Numerical Magnitude Processing in First Graders with Mathematical Difficulties. *Journal of Experimental Child Psychology*, vol. 108, no 2, pp. 278–292. https://doi.org/10.1016/j.jecp.2010.09.003

Sokal M.M. (1987) Psychological Testing and American Society 1890–1930. New Brunswick: Rutgers University.

Speelman C.P., McGann M. (2013) How Mean is the Mean? *Frontiers in Psychology*, vol. 4, July, Article no 451. https://doi.org/10.3389/fpsyg.2013.00451

Spence R., Owens M., Goodyer I. (2012) Item Response Theory and Validity of the NEO-FFI in Adolescents. *Personality and Individual Differences*, vol. 53, no 6, pp. 801–807. https://doi.org/10.1016/j.paid.2012.06.002

Stahl C., Voss A., Schmitz F., Nuszbaum M., Tüscher O., Lieb K., Klauer K.C. (2014) Behavioral Components of Impulsivity. *Journal of Experimental Psychology: General,* vol. 143, no 2, pp. 850–886. https://doi.org/10.1037/a0033981

Sternberg R.J. (1981) Testing and Cognitive Psychology. *American Psychologist*, vol. 36, no 10, pp. 1181–1189. https://doi.org/10.1037/0003-066X.36.10.1181

Stroop J.R. (1935) Studies of Interference in Serial Verbal Reactions. *Journal of Experimental Psychology,* vol. 18, no 6, pp. 643–662. https://doi.org/10.1037/h0054651

Tavakol M., Dennick R. (2011) Making Sense of Cronbach's Alpha. *International Journal of Medical Education*, vol. 2, June, pp. 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Terman L.M. (1924) The Mental Test as a Psychological Method. *Psychological Review,* vol. 31, no 2, pp. 93–117. https://doi.org/10.1037/h0070938

Troyer A.K., Leach L., Strauss E. (2006) Aging and Response Inhibition: Normative Data for the Victoria Stroop Test. *Aging, Neuropsychology, and Cognition*, vol. 13, no 1, pp. 20–35. https://doi.org/10.1080/138255890968187

Tuerlinckx F., De Boeck P.D. (2005) Two Interpretations of the Discrimination Parameter. *Psychometrika*, vol. 70, no 4, pp. 629–650. https://doi.org/10.1007/s11336-000-0810-3

Voronin I.A., Zakharov I.M., Tabueva A.O., Merzon L.A. (2020) Diffuznaya model´ prinyatiya resheniya: otsenka skorosti i tochnosti otvetov v zadachakh vybora iz dvukh al´ternativ v issledovaniyakh kognitivnykh protsessov i sposobnostey [Diffuse Decision-Making Model: Assessment of the Speed and Accuracy of Answers in the Problems of Choosing from Two Alternatives in the Study of Cognitive Processes and Abilities]. *The Theoretical and Experimental Psychology,* vol. 13, no 2, pp. 6–23.

Watrin J.P., Darwich R. (2012) On Behaviorism in the Cognitive Revolution: Myth and Reactions. *Review of General Psychology*, vol. 16, no 3, pp. 269–282. https://doi.org/10.1037/a0026766

Watson J.B. (1913) Psychology as the Behaviorist Views It. *Psychological Review*, vol. 20, no 2, pp. 158–177. https://doi.org/10.1037/h0074428

Wells F.L. (1912) The Relation of Practice to Individual Differences. *The American Journal of Psychology*, vol. 23, no 1, pp. 75–88. https://doi.org/10.2307/1413115

Whelan R. (2008) Effective Analysis of Reaction Time Data. *The Psychological Record*, vol. 58, no 3, pp. 475–482. https://doi.org/10.1007/BF03395630

Wicherts J.M. (2016) The Importance of Measurement Invariance in Neurocognitive Ability Testing. *The Clinical Neuropsychologist*, vol. 30, no 7, pp. 1006–1016. https://doi.org/10.1080/13854046.2016.1205136

Wijsen L.D., Borsboom D., Alexandrova A. (2022) Values in Psychometrics. *Perspectives on Psychological Science*, vol. 17, no 3, pp. 788–804. https://doi.org/10.1177/17456916211014183

Willoughby M.T., Wirth R.J., Blair C.B. (2012) Executive Function in Early Childhood: Longitudinal Measurement Invariance and Developmental Change. *Psychological Assessment*, vol. 24, no 2, pp. 418–431. https://doi.org/10.1037/a0025779

Wissler C. (1901) The Correlation of Mental and Physical Tests. *The Psychological Review: Monograph Supplements,* vol. 3, no 6, pp. i–62. https://doi.org/10.1037/h0092995

Опыт использования бифакторных моделей для снижения эффектов социальной желательности на материале нормативного опросника универсальных компетенций

Егор Сагитов, Ирина Брун, Станислав Павлов

Статья поступила в редакцию в феврале 2023 г.

Сагитов Егор Борисович — аспирант Института образования, Национальный исследовательский университет «Высшая школа экономики». Адрес: 101000, Москва, Потаповский пер., 16, стр. 10. E-mail: esagitov@hse.ru. ORCID: https://orcid.org/0009-0006-6163-480X (контактное лицо для переписки)

Брун Ирина Викторовна — психометрик, ООО «Форматта». E-mail: brun.i@formatta.ru

Павлов Станислав Витальевич — директор, ООО «Форматта». E-mail: pavlov.s@formatta.ru

Аннотация

Одним из существенных недостатков опросных методов психологического тестирования является искажение итоговых баллов по измеряемым конструктам, обусловленное эффектами социальной желательности. Угроза валидности решений, принимаемых на основании результатов опроса, которую создает социальная желательность, особенно значима в условиях высоких ставок, например при отборе на должность. При этом дискуссионным остается вопрос о связи разных компонентов социальной желательности с наиболее часто измеряемыми личностными конструктами. На материале авторского нормативного опросника универсальных компетенций рассматривается возможность внесения корректировок в итоговые баллы по измеряемым конструктам с использованием разработанных шкал эгоистической и моралистической социальной желательности. Обсуждается использование формулировок утверждений, нейтральных к социальной желательности и отражающих наиболее высокую степень выраженности измеряемого индикатора, в качестве способа минимизировать актуализацию намерения давать социально желательные ответы у респондента.

Эмпирическую основу исследования составили данные апробации опросника, проведенной весной 2022 г., в ходе которой получены ответы 579 респондентов по 49 компетенциям. Выполнена оценка качества разработанных шкал социальной желательности и проведено моделирование каждой из шкал универсальных компетенций с включением шкалы социальной желательности. Данные анализировались в рамках структурного моделирования — конфирматорного факторного анализа с использованием бифакторных моделей для каждой из измеряемых компетенций.

Установлено, что использование шкалы эгоистической социальной желательности в качестве основания для корректировки факторных баллов по измеряемым компетенциям имеет в целом удовлетворительные психометрические показатели, однако опасение вызывает сравнительно большая ошибка измерения. Рассматриваются достоинства и недостатки как используемого подхода, так и других наиболее часто применяемых практик, направленных на снижение эффектов социальной желательности в академической и бизнес-среде.

Ключевые слова эгоистическая социальная желательность, моралистическая социальная желательность, бифакторные модели, нормативный формат опросника, ипсативный формат опросника, универсальные компетенции

Для цитирования Сагитов Е.Б., Брун И.В., Павлов С.В. (2023) Опыт использования бифакторных моделей для снижения эффектов социальной желательности на материале нормативного опросника универсальных компетенций. *Вопросы образования / Educational Studies Moscow*, № 3, сс. 145–171. https://doi.org/10.17323/vo-2023-16827

Experience of Using Bifactor Models to Reduce the Effects of Social Desirability on the Normative Questionnaire of Universal Competencies

Egor Sagitov, Irina Brun, Stanislav Pavlov

Egor B. Sagitov — PhD Student, Institute of Education, National Research University Higher School of Economics. Address: 101000, Moscow, Potapovsky per., 16, build. 10. E-mail: esagitov@hse.ru. ORCID: https://orcid.org/0009-0006-6163-480X (corresponding author)

Irina V. Brun — psychometrician, LLC "Formatta". E-mail: brun.i@formatta.ru

Stanislav V. Pavlov — head, LLC "Formatta". E-mail: pavlov.s@formatta.ru

Abstract One of the significant lack of questionnaires is a scores distortion for the measured constructs, associated with the social desirability effects. An even greater threat to the validity of decisions is social desirability in high-stakes evaluation, such as selection for a position. Moreover the issue of the relationship between different components of social desirability and the most frequently measured personal constructs remains debatable. In the material of the author's normative questionnaire of universal competencies, an approach is considered for making adjustments to the final scores for measured constructs using the developed scales of egoistic and moralistic social desirability. Also discussed the prospect of using statement formulations that are neutral to social desirability or express the most positive degree of measured indicators.

The empirical basis of this study is data gathered within a pilot conducted in the spring of 2022, during which data were obtained from 579 respondents in 49 measurable competencies. The analysis was aimed at assessing the quality of the developed scales of social desirability and modeling of each of the universal competencies scales was carried out with the inclusion of a scale of social desirability. The data were analyzed in the framework of structural modeling — confirmatory factor analysis (CFA) using bifactor models for each of the measured competencies.

According to the results of this study, the use of the scale of egoistic social desirability as a measure for adjusting factor scores for the competencies has generally satisfactory psychometric statistics, but there is concern about the relatively large measurement error. The paper discusses the advantages and disadvantages of this approach and other practices that are most often used to reduce the effects of social desirability in the academic and business environment.

Keywords egoistic social desirability, moralistic social desirability, bifactorial models, normative approach, ipsative approach, universal competencies

For citing Sagitov E.B., Brun I.V., Pavlov S.V. (2023) Opyt ispol´zovaniya bifaktornykh modeley dlya snizheniya effektov sotsial´noy zhelatel´nosti na materiale normativnogo oprosnika universal´nykh kompetentsiy [Experience of Using Bifactor Models to Reduce the Effects of Social Desirability on the Normative Questionnaire of Universal Competencies]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 3, pp. 145–171. https://doi.org/10.17323/vo-2023-16827

Социальная желательность (СЖ) ответов состоит в склонности респондентов давать необоснованно (чрезмерно) положительные самоописания [Paulhus, 1998] и является одним из значимых факторов искажения результатов, получаемых с помощью опросных методов [Allport, 1937; Anderson, 1976; Grimm, 2010]. Учитывать влияние СЖ критически важно в условиях, когда нет возможности сделать измерение обезличенным и когда на основании ответов респондентов делаются выводы, определяющие их будущее, например о принятии на должность.

За десятилетия развития психометрики разработано множество способов измерения СЖ и минимизации ее эффектов на разных этапах создания измерительного инструмента, проведения оценочной процедуры и анализа полученных ответов [Paulhus, 1991; Salgado, 2005; Larson, 2018]. При этом исследователи отмечают, что стратегия минимизации эффектов СЖ имеет шансы быть эффективной только в случае индивидуального подхода разработчиков к созданию каждого отдельно взятого инструмента с учетом конкретных условий его использования [Osin, 2009; Larson, 2018].

В данной статье рассматривается применение комплекса мер по снижению эффектов СЖ на материале разрабатываемого инструмента измерения универсальных компетенций, который в дальнейшем планируется ввести в батарею методик для отбора кандидатов на должность. Преимущественное внимание уделяется бифакторным моделям с использованием скорректированной шкалы СЖ из *Balanced Inventory of Desirable Responding* [Paulhus, 1998; Osin, 2009]. Рассматриваются достоинства и недостатки данного подхода.

1. Обзор литературы

1.1. Почему СЖ угрожает валидности результатов оценивания с высокими ставками, полученных с помощью опросников

Последние несколько десятилетий личностные опросники широко используются в организациях при принятии важных решений, например при отборе кандидатов на должность или в кадровый резерв [García-Izquierdo, Ramos-Villagrasa, Lubiano, 2020; Golubovich et al., 2020; Sackett et al., 2017]. Большинство из этих опросников разработаны в рамках нормативного подхода [Martínez, Moscoso, Lado, 2021], который предполагает, что респондент оценивает степень своего согласия с каждым отдельным утверждением опросника, чаще всего по шкале Ликерта. Однако полученные таким образом сведения о респонденте не вызывают доверия ни у специалистов в области оценивания, ни у представителей бизнеса, поскольку респондент может легко завысить свои баллы [Christiansen, Burns, Montgomery, 2005; Heggestad et al., 2006; Kreitchmann et al., 2019], тем самым получив необоснованное преимущество перед другими кандидатами. Так, например, в нескольких исследованиях сравнивались итоговые баллы тех респондентов, которым в ситуации эксперимента была дана инструкция отвечать честно, и тех, которым разрешили «немного приукрасить» свои ответы. Результаты респондентов во второй группе в среднем оказались на 0,6 стандартного отклонения (σ) выше, чем в первой [Goffin, Christiansen, 2003]. О том, что респонденты могут искажать свои баллы при заполнении нормативных опросников, если они мотивированы создать позитивный образ себя, свидетельствуют и метааналитические исследования [Birkeland et al., 2006; Salgado, 2016]. В организационном контексте респонденты стремятся завысить баллы по характеристикам, которые, на их взгляд, положительно связаны с эффективностью деятельности и успешностью на желаемой позиции [Martínez, Moscoso, Lado, 2021; Anglim et al., 2017].

При этом искажением измеряемых показателей эффекты СЖ не ограничиваются. Результаты многочисленных исследований показывают, что СЖ может лежать в основе завышения корреляций между измеряемыми личностными чертами, например в одном из наиболее распространенных опросников *Big Five* [Costa, McCrae, 1992; Anusic et al., 2009; Bäckström, Björklund, Larsson, 2009; Bäckström, Björklund, 2020]. Кроме того, результаты метаанализа данных свидетельствуют в пользу того, что СЖ значимо (хотя и умеренно) коррелирует со многими психологическими конструктами в исследованиях организационного поведения [Moorman, Podsakof, 1992].

Риски возникновения вышеперечисленных эффектов заставляют значительную часть исследователей сомневаться в релевантности личностных показателей для прогнозирования успешности деятельности и других организационных показателей, поскольку валидность этих измерений будет существен-

но снижена из-за СЖ [Goffin, Christiansen, 2003; Mueller-Hanson, Heggestad, Thornton, 2003; Neeley, Cronley, 2004]. Некоторые специалисты полностью отвергают личностные показатели как основание для принятия кадровых решений. При этом доступных альтернативных способов прогнозирования эффективности сотрудников на рабочем месте пока не разработано.

1.2. Какие способы минимизации эффектов СЖ чаще всего используются

Изучение СЖ продолжается уже более пяти десятилетий, за это время разработано множество способов минимизации ее эффектов на разных этапах разработки и использования измерительных инструментов [Salgado, 2005; Larson, 2018]. Рассмотрим наиболее известные и часто применяемые практики, релевантные в условиях оценивания с высокими ставками, когда нет возможности провести измерение анонимно и респонденты знают о том, что на основании их ответов будут приниматься значимые для их жизни решения.

1.2.1. Минимизация эффектов СЖ на этапе выбора дизайна опросника

Первое решение, которое должен принять разработчик опросника, определяя стратегию минимизации эффектов СЖ, — сделать выбор между нормативным и ипсативным подходом. Учитывая обозначенные выше недостатки нормативного подхода, многие исследователи отдают предпочтение ипсативному [Hicks, 1970; Bowen, Martin, Hunt, 2002; Brown, Maydeu-Olivares, 2013; Bäckström, Björklund, 2020]. Классический вариант опросника в ипсативном формате предполагает попарное предъявление утверждений, из которых респонденту необходимо выбрать то, которое описывает его наилучшим образом. При этом популярность приобретают варианты опросников, в которых респондент должен совершить принудительный выбор (*forced-choice*) из трех или четырех утверждений: выбрать среди них то, которое описывает его наилучшим образом, и еще одно — в наименьшей степени относящееся к нему. Такой дизайн опросника позволяет уменьшить количество сравнений и тем самым сократить время прохождения опроса. Таким образом, ипсативный подход исключает ситуацию, в которой респондент выражает высокую степень согласия с каждым утверждением (как это часто бывает в нормативных опросниках), что помимо снижения эффектов СЖ решает проблему дифференциации итогового профиля [Brown, Maydeu-Olivares, 2013].

Кроме того, благодаря разработанным в последнее десятилетие IRT-моделям стало возможным сравнение респондентов по выраженности измеряемых конструктов [Brown, Maydeu-Olivares, 2013; Lee et al., 2019], тем самым было устранено препятствие к использованию ипсативного формата опроса при необ-

ходимости сравнения кандидатов. Широко известны примеры опоры на результаты применения личностных опросников, разработанных в рамках ипсативного подхода, при принятии кадровых решений, например OPQ 32 [SHL, 1999], DISC (Thomas International), D5D [Rolland, Mogenet, 2001].

Однако значительную трудность при разработке опросников в ипсативном формате представляет выравнивание блоков утверждений по уровню СЖ [Meade, 2004], к тому же разработка такого опросника более затратна. При этом ипсативный формат вовсе не исключает для респондента возможность давать социально желательные ответы, хотя многие исследователи считают, что условия принудительного выбора уменьшают количество «наименее изощренных» способов завышения собственных баллов по сравнению с нормативным форматом [Ibid.]. А меньшая в ипсативном формате по сравнению с нормативным корреляция между шкалами может быть связана не столько с минимизацией эффектов СЖ, сколько со свойствами самой ипсативной матрицы корреляций, в которой изначально предполагаются близкие к нулю или отрицательные значения [Ibid.].

Если же разработчик избирает для создания опросника нормативный формат, то для минимизации эффектов СЖ он может включить в опросник дополнительную шкалу для ее измерения и коррекции итоговых баллов на основе полученных оценок [Larson, 2018; Bäckström, Björklund, Larsson, 2009; Salgado, 2005]. Такая стратегия использовалась в трех самых популярных личностных опросниках — MMPI, 16-факторном опроснике Кеттела и опроснике *Big Five*, и результатом ее применения стали достаточно хорошие психометрические показатели [Salgado, 2005].

В ходе создания как этих, так и других инструментов разработано множество разных вариантов корректировки баллов. Так, например, в случае выявления высоких баллов по шкале СЖ практикуется прямое вычитание некоторого количества баллов из итогового показателя по измеряемому конструкту, однако такая практика часто основана на субъективном опыте исследователей и практиков [Ibid.]. Относительно новый способ внесения корректировок состоит в использовании бифакторных моделей [Holzinger, Swineford, 1937; Reise, 2012]. Данный подход основан на предположении, что дисперсия баллов по каждому утверждению частично объясняется конструктом, на измерение которого направлено это утверждение, и частично — СЖ. Таким образом, СЖ рассматривается как общий фактор, а измеряемые конструкты — как специфичные, что позволяет получить более «очищенные» от СЖ факторные баллы по измеряемым конструктам [Ferrando, Lorenzo-Seva, Chico, 2009; Biderman et al., 2011; Chen et al., 2016]. Бифакторные модели применялись, например, при разработке инструмента для оцен-

ки персонала HEXACO [Ashton, Lee, De Vries, 2014; Anglim et al., 2017], при изучении факторной структуры опросника *Big Five* [Anusic et al., 2009; Biderman et al., 2011] и при изучении самооценки [Kam, 2020].

В данном исследовании мы рассчитываем достичь удовлетворительных психометрических показателей разрабатываемого нормативного опросника, внося корректировки в итоговые баллы с использованием шкалы СЖ. Для этого мы будем применять бифакторные модели. Индикаторами минимизации эффектов СЖ будут служить:

а) корреляция между измеряемыми компетенциями. Сравнивая корреляции между баллами по компетенциям, рассчитанными на основе сырых сумм баллов и скорректированных факторных баллов, мы ожидаем, что включение в математическую модель фактора СЖ снизит эти корреляции, поскольку рассчитываемые факторные баллы будут учитывать связь балла за каждое утверждение как с компетенцией, так и с СЖ [Ferrando, Lorenzo-Seva, Chic, 2009] (гипотеза 1а);

б) завышение баллов. Сравнивая скорректированные факторные баллы между ответными профилями, одинаковыми по компетенциям, но различающимися по уровню СЖ, мы ожидаем, что скорректированные факторные баллы у ответных профилей, одинаковых по компетенциям, будут ниже у тех испытуемых, кто получит более высокие баллы по СЖ (гипотеза 1б).

1.2.2. Минимизация эффектов СЖ на этапе разработки утверждений

Многие исследователи предпринимали попытки создать опросники, устойчивые к эффектам СЖ, фокусируясь на формулировках утверждений, которыми измеряются целевые конструкты. Формулировки разрабатывались с таким расчетом, чтобы они минимально актуализировали СЖ [Jackson, 1984]. В частности, существует такая версия опросника *Big Five* [Bäckström, Björklund, 2020]. При этом средние значения по нейтрализированным утверждениям были ниже, чем в изначальной версии опросника, и приближались к среднему значению используемой шкалы Ликерта.

Однако исследований, в которых оценивалось снижение эффектов СЖ при перефразировании утверждений, мало, и их результаты противоречивы. Мы предлагаем альтернативный подход, при котором формулировки утверждений разрабатываются с таким расчетом, чтобы они ярко отражали максимальную выраженность индикатора. При этом мы ожидаем, что средние баллы по утверждениям будут приближаться к средним значениям шкалы Ликерта, как и в приведенных выше исследованиях, поскольку респондентам будет труднее выражать высокую степень согласия с такими утверждениями (гипотеза 2).

1.2.3. Минимизация эффектов СЖ на этапе проведения измерения

Эффективным средством снижения влияния СЖ является мотивирование респондентов к тому, чтобы отвечать честно [Bryan, Adams, Monin, 2013], для этого их, например, предупреждают, что в случае выявления завышения баллов результаты оценивания будут аннулированы [Bryan, Adams, Monin, 2013; McFarland, Ryan, 2006]. Эта практика, очень простая и не требующая особых затрат, — одна из самых надежных, при ее использовании имеет место минимальное завышение баллов [Salgado, 2005], и с учетом целей измерения ее можно применять как базовую.

1.3. Как измеряют социальную желательность ответов респондентов

Для измерения СЖ создано множество шкал — прежде всего разработчиками измерительных инструментов в области психологии и оценки персонала. При этом долгое время СЖ рассматривалась как одномерный конструкт: это предположение лежит в основе оценки СЖ в таких широко используемых инструментах, как шкала Краун — Марлоу [Crowne, Marlowe, 1960], шкала искренности Айзенка [Eysenck, Eysenck, 1964], шкала позитивных мотивационных искажений 16-факторного опросника Кеттела [Cattell, Mead, 2008] и шкала социальной желательности Профессионального личностного опросника [SHL, 1999].

Однако впоследствии наибольшее признание получила модель Д. Паулхуса [Paulhus, 1998; Paulhus, Vazire, 2007], согласно которой СЖ включает два компонента — эгоистический и моралистический. Эгоистический компонент СЖ отражает стремление обладать властью и иметь высокий социальный статус — так называемый нормальный нарциссизм [Paulhus, 1998]. Моралистический компонент СЖ связан с потребностью в общении, которое предполагает избегание неодобрения со стороны общества и соблюдение социальных норм. При этом имеются свидетельства того, что эгоистический компонент СЖ более тесно связан с ответами респондентов на личностные опросники и поведением сотрудников на рабочем месте [Salgado, 2005].

На основе этой теоретической модели Д. Паулхус [Paulhus, 1991; 1998] разработал инструмент измерения *Balanced Inventory of Desirable Responding* (BIDR), который показал достаточно хорошие психометрические свойства. Разрабатываемая нами шкала СЖ основана на переводе данной методики на русский язык, выполненном С. Лебедевым [Osin, 2009]. В соответствии с ранее полученными результатами мы предполагаем, что эгоистический компонент СЖ будет в среднем сильнее связан с измеряемыми универсальными компетенциями, чем моралистический компонент (гипотеза 3). Поэтому мы считаем оправданным использование эгоистического компонента СЖ в качестве общего фактора в бифакторных моделях.

2. Данные

2.1. Подготовительное исследование

В качестве материала для проверки гипотез использовался разработанный нами опросник универсальных компетенций, предназначенный для тестирования при отборе кандидатов на должность. На основе анализа распространенных теоретических моделей универсальных компетенций в академической и бизнес-среде разработана модель, которая изначально включала 49 компетенций и 237 индикаторов, описывающих поведение сотрудников в рабочем контексте (полный перечень компетенций см. в табл. 2). Под универсальными компетенциями мы понимаем те знания, умения и установки, которые способствуют созданию человеком дохода и других полезных эффектов как для себя, так и для работодателя и общества в целом [Kuzminov et al., 2018].

Далее мы разработали 481 утверждение (по 2–3 утверждения на каждый индикатор) и провели пилотное исследование психометрических характеристик этих утверждений, оценив их соответствие теоретической модели на выборке из 158 человек. Респонденты для пилотного исследования отбирались так же, как для основного (см. раздел 2.2). На основании таких показателей, как средний балл по утверждению, дискриминативность и стандартизированная факторная нагрузка, мы итеративно исключали утверждения до того момента, пока на каждый индикатор не оставалось одно утверждение с наилучшими психометрическими показателями: минимальный средний балл, максимальные дискриминативность и стандартизированная факторная нагрузка. После пилотного этапа исследования теоретическая модель опросника включала 232 индикатора и соответственно 232 измеряющих их утверждения. Пять индикаторов из тех, что присутствовали в изначальной теоретической модели, были исключены из модели, поскольку оба репрезентирующих их утверждения имели стандартизированные факторные нагрузки значительно ниже 0,5 и теоретически имели наименьшую связь с измеряемыми компетенциями.

2.2. Основное исследование

Опрос проводился летом 2022 г. при помощи специального программного обеспечения «Анкетолог». Принять участие в исследовании мог любой зарегистрированный пользователь, достигший 18-летнего возраста. Мы ввели возрастное ограничение для того, чтобы отсечь респондентов, которые с высокой долей вероятности не имели или практически не имели опыта профессиональной деятельности. За участие в исследовании предлагалось вознаграждение в размере 150 рублей.

Начальная выборка составила 670 респондентов. Однако с учетом большого количества утверждений в опроснике и высокого риска немотивированности респондентов к тому, что-

бы добросовестно пройти весь опросник, мы приняли решение провести дополнительный отсев.

Для отсева недобросовестных заполнений использованы три критерия. Во-первых, респондентам в двух тестовых утверждениях, выбранных случайным образом, предъявлялись проверочные указания типа «Если вы прочитали текст этого вопроса, выберите вариант ответа "Совершенно НЕ согласна"». Если респондент выбирал неверный вариант ответа хотя бы на одно из этих утверждений, все ответы данного респондента исключались из дальнейшего анализа. Во-вторых, контролировалось время заполнения опросника, исходя из того, что для осознанного прочтения и ответа на каждое утверждение минимально необходимо 2 секунды [Huang et al., 2011]. В-третьих, из дальнейшего рассмотрения исключались ответы тех респондентов, которые выражали одинаковую степень согласия с шестью утверждениями на пяти и более листах. Это критическое значение выбрано на основании распределения количества одинаково заполненных страниц (всего 39 страниц).

Таким образом, итоговую выборку исследования составили 579 респондентов, среди которых 64% — женщины, 52% входят в возрастную группу от 36 до 60 лет, и 63% имеют высшее образование (табл. 1).

Таблица 1. **Описательные статистики выборки (*N* = 579)**

Переменная	Ответные категории	Доля (%)
Пол	Женщина	64,2
	Мужчина	35,8
Возраст	18–25 лет	2,1
	26–35 лет	27,8
	36–50 лет	52,5
	51–64 года	17,1
	65 лет или старше	0,1
Уровень образования	Неполное среднее (9 классов или меньше)	0,1
	Полное среднее общее (10 или 11 классов)	3,8
	Начальное или среднее профессиональное (техникум, колледж, ПТУ и др.)	16,1
	Неоконченное высшее (3 курса вуза или больше)	5,9
	Высшее профессиональное (институт, университет), включая бакалавриат, специалитет, магистратуру	62,5
	Послевузовское профессиональное образование (аспирантура, повышение квалификации и др.) или несколько высших	10,9

Переменная	Ответные категории	Доля (%)
Общий рабочий стаж	Нет опыта работы	0,0
	Менее 1 года	0,0
	От 1 года до 3 лет	2,6
	От 4 до 10 лет	19,0
	От 11 до 20 лет	38,5
	От 21 года до 30 лет	26,8
	Более 31 года	12,6
Основная деятельность респондента за последние 3 месяца	Учусь	1,7
	Занимаюсь неоплачиваемым трудом	8,1
	Работаю	90,5
	Временно не работаю	2,8
Уровень должности	Рабочий	9,7
	Специалист	48,0
	Линейный руководитель / тим-лидер	5,5
	Руководитель среднего звена	20,6
	Топ-менеджер	4,5
	Собственник бизнеса	9,0

2.3. Измерения

Для измерения универсальных компетенций и СЖ использована 7-балльная шкала Ликерта со следующими вариантами ответа: 0 — «совершенно НЕ согласна», 1 — «почти полностью НЕ согласна», 2 — «скорее НЕ согласна», 3 — «отчасти согласна, отчасти нет», 4 — «скорее согласна», 5 — «почти полностью согласна», 6 — «совершенно согласна».

Для того чтобы респонденты точно увидели частицу «не», она была выделена прописными буквами. Варианты ответов сформулированы в женском роде, поскольку мы предполагали, что большую часть выборки составят женщины.

2.4. Разработка шкал СЖ

В качестве основы для создания шкал СЖ использовался подготовленный С. Лебедевым [Osin, 2009] перевод версии опросника BIDR, состоящей из 60 пунктов [Paulhus, 1998]. С разрешения автора статьи мы отобрали 24 утверждения (13 на эгоистический компонент СЖ и 11 — на моралистический), которые наиболее близки по содержанию к рабочим ситуациям и которые по результатам исследования имели хорошие психометрические показатели [Osin, 2009]. Далее формулировки этих утверж-

дений были переработаны таким образом, чтобы по лексике и синтаксису они были максимально похожи на утверждения, измеряющие компетенции, — мы рассчитывали, что в восприятии респондента утверждения, направленные на оценку СЖ, не будут сильно «выбиваться» из контекста опросника. Например, одно из утверждений, направленных на измерение компетенции «Понимает людей», выглядит следующим образом: «При возникновении даже малейших сомнений в правильном понимании слов другого человека я обязательно переспрашиваю и уточняю». А одно из утверждений, направленных на измерение эгоистического компонента СЖ, сформулировано так: «Я всегда мыслю совершенно здраво и рационально».

3. Аналитическая стратегия

Поскольку главная цель исследования состоит в оценке психометрических характеристик опросника и разработке математической модели снижения эффектов СЖ, для анализа данных мы обратились к методам структурного моделирования — к конфирматорному факторному анализу (КФА). Одно из его достоинств заключается в сравнительной простоте установления степени согласия между теоретически ожидаемой факторной структурой и полученными эмпирическими данными. Для оценки качества математических моделей мы использовали следующие стандартные статистики: *Comparative Fit Index* (CFI), *Tucker — Lewis Index* (TLI), *Root Mean Square Error of Approximation* (RMSEA) и *Standardized Root Mean Square Residual* (SRMR). В соответствии с критическими значениями, принятыми в исследовательском сообществе [Brown, 2015], статистики CFI и TLI считаются удовлетворительными, если принимают значения >0,90, и хорошими — при значениях >0,95. Статистика RMSEA считается удовлетворительной, если принимает значения <0,08, и хорошей — при значениях <0,06. Статистика SRMR считается удовлетворительной, если принимает значения <0,06. В качестве показателя надежности используется коэффициент омега, который рассчитывается через значения стандартизированных факторных нагрузок структурной модели [Viladrich, Angulo-Brunet, Doval, 2017]. Анализ проводился с помощью пакета *lavaan v*. 0.6-7 для R v.3.6.3.

Анализ выполнялся в следующей последовательности. Сначала оценивались психометрические показатели утверждений и каждой отдельной компетенции. Затем определялись психометрические показатели двух компонентов СЖ (эгоистический и моралистический). После этого проведен анализ связи компонентов СЖ со шкалами, измеряющими компетенции.

Далее для того, чтобы учесть долю дисперсии, которую объясняет общий фактор СЖ, строились бифакторные модели, ка-

ждая из которых включает два ортогональных фактора: фактор компетенции и совмещенные факторы компетенции и СЖ, после чего рассчитывались факторные баллы с помощью регрессий [Thurstone, 1935; DiStefano, Zhu, Mîndrilă, 2009]. Таким способом удается сравнительно легко получить факторные баллы каждого респондента по измеряемым конструктам с учетом сложной структуры математической модели.

В качестве индикаторов минимизации эффектов СЖ мы оценивали:

а) корреляции между компетенциями — сравнивались значения корреляций между баллами по компетенциям, рассчитанные на основе сумм сырых баллов до внесения корректировок и скорректированных факторных баллов;

б) завышение баллов — сравнивались скорректированные факторные баллы между ответными профилями, одинаковыми по компетенциям, но различающимися по уровню СЖ, с помощью симуляции данных.

4. Результаты анализа

На этапе подготовительного анализа полученных данных обнаружен небольшой скос в сторону большего согласия респондентов с утверждениями, поэтому при проведении конфирматорного факторного анализа мы приняли решение использовать метод максимального правдоподобия с робастными стандартными ошибками, устойчивый к отклонениям от нормального распределения. Средний балл по всем утверждениям равен 3,97 при среднем значении используемой шкалы, равном 3 (медиана равна 4). Этот результат можно расценивать как свидетельство того, что использованные нами формулировки утверждений, ярко отражающие максимальную выраженность индикатора, затрудняли для респондентов проявление наибольшей степени согласия с утверждениями, направленными на измерение компетенций (гипотеза 2).

4.1. Оценка факторной структуры универсальных компетенций до корректировки баллов

Сначала оценивалась факторная структура каждой компетенции отдельно (табл. 2). Эта процедура необходима, поскольку дизайн инструмента предполагает свободный выбор заказчиком набора измеряемых конструктов. Большинство компетенций имеет хорошие или приемлемые значения по статистикам CFI, TLI и SRMR, однако в шести компетенциях значения статистики RMSEA превышают критические (выделены в табл. 2 жирным шрифтом), что может быть связано с использованием преимущественно коротких шкал. Компетенции, которые измеряются тремя утверждениями, имеют наилучшие статистические показатели, что обусловлено особенностями математи-

ческих расчетов. В семи компетенциях среди тестирующих их утверждений есть одно с факторной нагрузкой от 0,4 до 0,5. Однако на этом этапе мы сочли важным сохранить эти утверждения, поскольку они отражают ценные с точки зрения портрета респондента индикаторы компетенций. На следующих этапах анализа мы исключили шесть утверждений, наличие которых в бифакторных моделях значительно ухудшало их показатели, в том числе из-за относительно более сильной связи со шкалой СЖ. Таким образом, итоговая теоретическая модель опросника включает 49 компетенций и 226 индикаторов. В целом можно считать изначальные шкалы достаточно хорошо функционирующими с учетом их малой длины (3–8 утверждений).

Таблица 2. **Психометрические показатели компетенций до корректировки баллов (*N* = 579)**

Компетенция (количество утверждений)	CFI	TLI	RMSEA	SRMR	Минимальная стандартизированная факторная нагрузка*	Средняя стандартизированная факторная нагрузка*	Корреляция с эгоистической СЖ**	Корреляция с моралистической СЖ**
Сохраняет самообладание (5)	1,00	1,00	0,03	0,02	0,64	0,73	0,70	0,49
Проявляет выносливость (3)	1,00	1,00	0,00	0,00	0,56	0,67	0,55	0,48
Работает энергично (6)	0,95	0,92	**0,09**	0,04	0,62	0,66	0,63	0,50
Адаптируется к изменениям (5)	0,99	0,98	0,05	0,02	0,63	0,68	0,64	0,46
В ситуации неопределенности действует (5)	0,96	0,93	0,08	0,04	0,53	0,61	0,60	0,46
Проявляет инициативу (4)	1,00	1,00	0,00	0,01	0,52	0,66	0,56	0,45
Верит в лучший исход (4)	0,96	0,89	**0,13**	0,03	0,61	0,68	0,59	0,50
Самосовершенствуется (6)	0,98	0,97	0,05	0,03	0,52	0,60	0,67	0,58
Поощряет обратную связь в свой адрес (3)	1,00	1,00	0,00	0,00	0,57	0,59	0,44	0,46
Конкурирует (4)	0,99	0,98	0,05	0,02	0,56	0,67	0,55	0,33
Транслирует уверенность в себе (3)	1,00	1,00	0,00	0,00	0,57	0,64	0,58	0,39
Мотивирует других (4)	1,00	0,99	0,04	0,01	0,67	0,70	0,52	0,52
Создает и укрепляет команду (7)	0,99	0,98	0,05	0,03	0,65	0,70	0,53	0,51
Занимает лидирующую позицию (5)	0,94	**0,87**	**0,12**	0,05	0,40	0,63	0,56	0,43
Дает сотрудникам возможности для развития (5)	1,00	1,00	0,00	0,02	0,54	0,61	0,54	0,53
Понимает людей (7)	0,99	0,99	0,03	0,03	0,56	0,64	0,55	0,61
Поддерживает других людей (6)	0,97	0,95	0,08	0,04	0,63	0,67	0,50	0,55
Работает в команде (5)	0,99	0,97	0,05	0,03	0,56	0,59	0,47	0,53

Компетенция (количество утверждений)	CFI	TLI	RMSEA	SRMR	Минимальная стандартизированная факторная нагрузка*	Средняя стандартизированная факторная нагрузка*	Корреляция с эгоистической СЖ**	Корреляция с моралистической СЖ**
Поддерживает широкий круг общения (5)	1,00	0,99	0,03	0,02	0,42	0,56	0,59	0,49
Проявляет социальную смелость (3)	1,00	1,00	0,00	0,00	0,70	0,76	0,58	0,39
Ценит разнообразие людей и взглядов (5)	0,96	0,91	**0,09**	0,04	0,49	0,59	0,56	0,61
Управляет конфликтами (4)	0,99	0,98	0,05	0,02	0,56	0,61	0,62	0,58
Влияет на других людей (7)	0,99	0,99	0,03	0,03	0,54	0,64	0,64	0,45
Ведет переговоры (5)	0,98	0,97	0,07	0,03	0,63	0,69	0,65	0,48
Выступает на публике (4)	0,99	0,97	**0,09**	0,02	0,71	0,75	0,57	0,38
Ясно излагает мысли (3)	1,00	1,00	0,00	0,00	0,72	0,79	0,57	0,35
Производит впечатление на людей (4)	0,97	0,96	0,06	0,03	0,52	0,6	0,33	0,40
Учитывает интересы стейкхолдеров и клиентов (4)	1,00	1,00	0,00	0,01	0,59	0,63	0,56	0,56
Говорит прямо (4)	1,00	0,99	0,03	0,02	0,56	0,63	0,54	0,39
Организует работу других (8)	0,99	0,98	0,04	0,03	0,54	0,66	0,53	0,50
Планирует работу (6)	0,98	0,97	0,05	0,03	0,57	0,63	0,62	0,50
Следует плану (3)	1,00	1,00	0,00	0,00	0,44	0,60	0,46	0,47
Преследует цели (4)	1,00	1,00	0,00	0,01	0,60	0,65	0,58	0,39
Выбирает амбициозные цели (3)	1,00	1,00	0,00	0,00	0,71	0,73	0,59	0,51
Добивается высокого качества (6)	0,97	0,95	0,08	0,03	0,62	0,68	0,56	0,54
Принимает решение действовать (4)	1,00	1,00	0,00	0,01	0,61	0,64	0,64	0,43
Поддерживает социальные и этические нормы (5)	1,00	1,00	0,00	0,01	0,58	0,63	0,54	0,65
Поддерживает правила и процедуры (5)	0,98	0,96	0,06	0,03	0,56	0,6	0,52	0,56
Внедряет и использует цифровые инструменты (4)	1,00	1,00	0,00	0,02	0,45	0,63	0,42	0,37
Ищет и использует новые знания (3)	1,00	1,00	0,00	0,00	0,54	0,66	0,55	0,41
Мыслит нестандартно (3)	1,00	1,00	0,00	0,00	0,67	0,75	0,61	0,40
Мыслит практически (4)	1,00	1,00	0,00	0,01	0,40	0,59	0,55	0,49
Мыслит предпринимательски (5)	1,00	1,00	0,00	0,01	0,64	0,69	0,62	0,52
Ищет информацию (4)	0,99	0,97	0,07	0,02	0,62	0,70	0,54	0,47
Опирается на достоверную информацию (3)	1,00	1,00	0,00	0,00	0,43	0,59	0,54	0,44

Компетенция (количество утверждений)	CFI	TLI	RMSEA	SRMR	Минимальная стандартизированная факторная нагрузка*	Средняя стандартизированная факторная нагрузка*	Корреляция с эгоистической СЖ**	Корреляция с моралистической СЖ**
Анализирует информацию (7)	1,00	1,00	0,00	0,01	0,62	0,68	0,66	0,45
Мыслит стратегически (6)	0,97	0,95	0,08	0,03	0,60	0,68	0,69	0,51
Рассматривает альтернативы (4)	1,00	1,00	0,00	0,01	0,50	0,63	0,52	0,45
Мыслит абстрактно (4)	0,97	0,91	**0,09**	0,03	0,55	0,62	0,58	0,51

* Из-за большого количества компетенций и утверждений мы приводим значения минимальной и средней стандартизированной факторной нагрузки по каждому фактору. Значения всех факторных нагрузок значимы ($p < 0,01$).
** Все значения коэффициентов корреляции Спирмена значимы ($p < 0,01$).

4.2. Анализ компонентов СЖ

Прежде всего из шкал СЖ мы исключили плохо функционирующие утверждения, для которых значения стандартизированных факторных нагрузок были значительно ниже 0,5 или которые имели слишком сильную связь с другим утверждением шкалы. Итоговую версию составили 10 утверждений на эгоистический компонент СЖ и 7 утверждений на моралистический. Психометрические показатели обеих шкал приведены в табл. 3. В целом их можно охарактеризовать как достаточно хорошие.

Затем мы проанализировали связь обоих компонентов СЖ с каждой компетенцией. В среднем корреляционная связь компетенций с эгоистическим компонентом оказалась сильнее (0,57), чем с моралистическим (0,48) (см. табл. 2), что свидетельствует в пользу гипотезы 3. Только для семи компетенций связь с моралистическим компонентом СЖ оказалась выше, чем с эгоистическим (отмечены курсивом в табл. 2). На основании этих результатов мы приняли решение скорректировать баллы по компетенциям, используя только шкалу эгоистической СЖ, поскольку этот компонент сильнее связан с компетенциями и, вероятно, может сильнее искажать итоговые баллы по компетенциям.

Таблица 3. **Психометрические показатели двух компонентов СЖ**

Компонент СЖ (количество утверждений)	CFI	TLI	RMSEA	SRMR	Минимальная стандартизированная факторная нагрузка	Средняя стандартизированная факторная нагрузка
Эгоистический (10)	0,96	0,95	0,05	0,04	0,50	0,58
Моралистический (7)	0,95	0,94	0,05	0,04	0,53	0,59

4.3. Корректировка факторных баллов по компетенциям

В табл. 4 приведены психометрические показатели построенных бифакторных моделей по каждой из измеряемых компетенций. Все значения по статистикам CFI, TLI, RMSEA, SRMR хорошие или удовлетворительные, а по большинству выделенных выше компетенций показатели улучшились по сравнению с изначальными моделями. При этом значения стандартизированных факторных нагрузок стали значительно хуже в сравнении с моделями без включения фактора эгоистического компонента СЖ. Для большинства компетенций имеются одно-два утверждения, у которых нагрузки на фактор компетенции существенно ниже 0,5, что связано с введением в модель фактора СЖ, который «забрал на себя» часть объясняемой компетенцией дисперсии. Именно поэтому наблюдается достаточно высокая ошибка измерения, при которой погрешность оценки может быть до одного стена в сторону больших или меньших значений. Использование же для корректировки шкалы моралистической СЖ или объединенного фактора, состоящего из обоих компонентов, не привело к улучшению статистических показателей, поэтому мы не приводим их в данной работе. Пример бифакторной модели для компетенции «Понимает людей» представлен на рис. 1. Аналогичные модели были построены для каждой из компетенций.

Таблица 4. **Психометрические показатели компетенций после корректировки баллов (*N* = 579)**

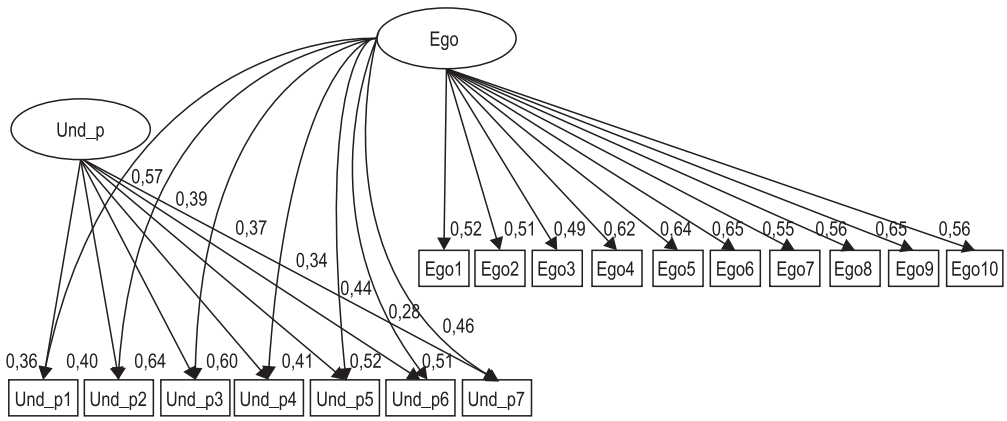
Компетенция (количество утверждений)	CFI	TLI	RMSEA	SRMR	Надежность Rel-omega	Станд. ошибка	Минимальная стандартизированная факторная нагрузка*	Средняя стандартизированная факторная нагрузка*
Сохраняет самообладание (5)	0,95	0,94	0,05	0,04	0,57	0,61	0,31	0,46
Проявляет выносливость (3)	0,97	0,96	0,04	0,04	0,49	0,61	0,39	0,49
Работает энергично (6)	0,96	0,94	0,05	0,04	0,57	0,62	0,27	0,42
Адаптируется к изменениям (5)	0,95	0,94	0,05	0,04	0,54	0,66	0,34	0,44
В ситуации неопределенности действует (5)	0,94	0,93	0,05	0,04	0,51	0,68	0,35	0,41
Проявляет инициативу (4)	0,95	0,94	0,05	0,04	0,53	0,64	0,47	0,35
Верит в лучший исход (4)	0,95	0,94	0,05	0,04	0,53	0,64	0,37	0,47
Самосовершенствуется (6)	0,93	0,91	0,06	0,05	0,43	0,68	0,23	0,33
Поощряет обратную связь в свой адрес (3)	0,95	0,94	0,05	0,04	0,46	0,72	0,43	0,47
Конкурирует (4)	0,94	0,92	0,06	0,05	0,56	0,61	0,34	0,49
Транслирует уверенность в себе (3)	0,95	0,94	0,05	0,04	0,37	0,76	0,34	0,40

Компетенция (количество утверждений)	CFI	TLI	RM-SEA	SRMR	Надежность Rel-omega	Станд. ошибка	Минимальная стандартизированная факторная нагрузка*	Средняя стандартизированная факторная нагрузка*
Мотивирует других (4)	0,93	0,92	0,06	0,05	0,64	0,57	0,46	0,55
Создает и укрепляет команду (7)	0,95	0,94	0,05	0,05	0,76	0,49	0,42	0,55
Занимает лидирующую позицию (5)	0,93	0,91	0,06	0,05	0,55	0,63	0,32	0,44
Дает сотрудникам возможности для развития (5)	0,94	0,93	0,05	0,05	0,56	0,64	0,29	0,45
Понимает людей (7)	0,94	0,93	0,05	0,05	0,70	0,53	0,40	0,49
Поддерживает других людей (6)	0,93	0,92	0,06	0,05	0,70	0,53	0,41	0,53
Работает в команде (5)	0,92	0,90	0,06	0,05	0,60	0,62	0,41	0,48
Поддерживает широкий круг общения (5)	0,94	0,93	0,05	0,05	0,58	0,62	0,37	0,47
Проявляет социальную смелость (3)	0,96	0,95	0,05	0,04	0,57	0,59	0,49	0,55
Ценит разнообразие людей и взглядов (5)	0,94	0,93	0,05	0,05	0,55	0,64	0,29	0,44
Управляет конфликтами (4)	0,91	0,90	0,07	0,05	0,40	0,75	0,29	0,38
Влияет на других людей (7)	0,92	0,90	0,06	0,05	0,56	0,66	0,31	0,40
Ведет переговоры (5)	0,93	0,91	0,06	0,05	0,55	0,62	0,27	0,44
Выступает на публике (4)	0,95	0,94	0,05	0,04	0,63	0,57	0,05	0,55
Ясно излагает мысли (3)	0,95	0,94	0,06	0,05	0,58	0,58	0,47	0,56
Производит впечатление на людей (4)	0,93	0,91	0,06	0,05	0,61	0,60	0,47	0,53
Учитывает интересы стейкхолдеров и клиентов (4)	0,92	0,91	0,06	0,05	0,50	0,69	0,43	0,45
Говорит прямо (4)	0,94	0,93	0,05	0,05	0,49	0,67	0,29	0,44
Организует работу других (8)	0,96	0,95	0,04	0,05	0,75	0,50	0,41	0,52
Планирует работу (6)	0,92	0,9	0,06	0,05	0,56	0,65	0,32	0,42
Следует плану (3)	0,94	0,92	0,06	0,05	0,46	0,64	0,33	0,47
Преследует цели (4)	0,95	0,94	0,05	0,04	0,43	0,73	0,32	0,42
Выбирает амбициозные цели (3)	0,96	0,95	0,05	0,04	0,52	0,65	0,43	0,51
Добивается высокого качества (6)	0,93	0,91	0,06	0,05	0,68	0,56	0,42	0,51
Принимает решение действовать (4)	0,96	0,95	0,05	0,04	0,42	0,70	0,23	0,39

Компетенция (количество утверждений)	CFI	TLI	RM-SEA	SRMR	Надежность Rel-omega	Станд. ошибка	Минимальная стандартизированная факторная нагрузка*	Средняя стандартизированная факторная нагрузка*
Поддерживает социальные и этические нормы (5)	0,94	0,93	0,05	0,05	0,58	0,61	0,33	0,46
Поддерживает правила и процедуры (5)	0,94	0,92	0,05	0,05	0,56	0,65	0,35	0,45
Внедряет и использует цифровые инструменты (4)	0,95	0,94	0,05	0,04	0,62	0,57	0,33	0,53
Ищет и использует новые знания (3)	0,95	0,94	0,05	0,04	0,46	0,69	0,37	0,47
Мыслит нестандартно (3)	0,95	0,94	0,06	0,04	0,53	0,30	0,35	0,51
Мыслит практически (4)	0,95	0,94	0,05	0,04	0,46	0,70	0,29	0,42
Мыслит предпринимательски (5)	0,97	0,96	0,04	0,04	0,58	0,63	0,45	0,47
Ищет информацию (4)	0,95	0,93	0,05	0,05	0,63	0,56	0,34	0,54
Опирается на достоверную информацию (3)	0,94	0,92	0,06	0,05	0,40	0,67	0,19	0,42
Анализирует информацию (7)	0,96	0,95	0,05	0,05	0,62	0,61	0,31	0,43
Мыслит стратегически (6)	0,94	0,93	0,06	0,05	0,50	0,69	0,28	0,39
Рассматривает альтернативы (4)	0,95	0,93	0,04	0,05	0,56	0,63	0,42	0,49
Мыслит абстрактно (4)	0,95	0,93	0,05	0,04	0,46	0,70	0,30	0,42

* Из-за большого количества компетенций и утверждений мы приводим значения минимальной и средней стандартизированной факторной нагрузки по каждому фактору. Значения всех факторных нагрузок значимы ($p < 0,01$).

Рис. 1. **Пример моделирования компетенции «Понимает людей», где Und_p — специфический фактор компетенции «Понимает людей», Ego — общий фактор эгоистической СЖ, числами указаны значения стандартизированных факторных нагрузок соответствующих индикаторов ($p < 0,01$)**



4.4. Проверка минимизации эффектов СЖ

Для проверки эффективности используемого подхода к минимизации влияния СЖ мы использовали два индикатора: сравнение значений корреляций между всеми компетенциями до и после внесения корректировок; сравнение ответных профилей, одинаковых по компетенциям, но различающихся по уровню СЖ, на основе симуляции данных.

Среднее значений корреляций Спирмена между баллами по всем компетенциям, рассчитанных на основе суммы сырых баллов до внесения корректировок, равняется 0,6. Среднее значений корреляции между факторными баллами по всем компетенциям после внесения корректировок равняется 0,41. Различие существенное и свидетельствует в пользу гипотезы 1а. В целом связь между компетенциями после внесения корректировок можно охарактеризовать как слабую.

Поскольку измеряемых компетенций много, мы приводим пример симуляции для случайно выбранной компетенции «Понимает людей». Мы использовали три ответных профиля по этой компетенции, различающихся по выраженности измеряемого конструкта на основе суммы сырых баллов в единицах стандартного отклонения: низкий уровень (–1σ), средний уровень (0σ), высокий уровень (1σ).

По каждому из трех ответных профилей мы симулировали ответы по шкале СЖ, как если бы респонденты с ответными профилями, одинаковыми по компетенции, имели разные по уровню выраженности ответы по шкале СЖ (в действительности мы проверяли большее количество уровней, но в рамках данной работы приводим часть значений): низкий уровень (–1,5σ), средний уровень (0σ), высокий уровень (1,5σ).

В табл. 5 приведены значения факторных баллов по компетенции «Понимает людей» после внесения корректировки. Из этих значений видно, что при одинаковых сырых баллах по компетенции, чем выше уровень СЖ, тем ниже факторный балл по компетенции. Это справедливо для всех ответных профилей, независимо от уровня выраженности конструкта. Разница в факторных баллах между одинаковыми по компетенции ответными профилями, имеющими низкий и высокий уровень

Таблица 5. **Факторные баллы по компетенции «Понимает людей» для разных ответных профилей**

Ответные профили по компетенции «Понимает людей»	Низкий уровень СЖ (–1,55)	Средний уровень СЖ (0,03)	Высокий уровень СЖ (1,49)
Низкий уровень компетенции	0,21	–0,69	–1,44
Средний уровень компетенции	1,03	0,13	–0,61
Высокий уровень компетенции	2,11	1,21	0,47

СЖ, достигает примерно 1–1,5σ, т.е. при использовании модели, включающей общий фактор СЖ, происходит значительная коррекция итогового балла, что свидетельствует в пользу гипотезы 1б. Для остальных компетенций эти различия составляют от 0,8σ до 1,7σ в зависимости от силы связи между компетенцией и эгоистическим компонентом СЖ.

5. Ограничения исследования

Данное исследование имеет несколько ограничений, которые необходимо учитывать при интерпретации полученных результатов.

Во-первых, условия, в которых проводилось измерение компетенций, отличаются от ситуации оценивания кандидатов на должность: ответы респондентов анонимны и не служат основанием для принятия значимых для респондента решений. При дальнейшем использовании инструмента измерения в соответствии с его целями актуализация компонентов СЖ и их эффекты могут значительно отличаться от имевших место в данном исследовании, а полученные показатели должны быть вновь проверены и при необходимости переоценены в новой ситуации использования инструмента.

Во-вторых, практически невозможно определить, насколько сознательно респондент чрезмерно завышает баллы по измеряемым конструктам. Так, например, исследования показывают, что часть ответов, расцениваемых как проявление СЖ, обусловлена случайными ошибками в воспоминаниях [Krosnick, 1999], а не преднамеренной ложью. Критически важно то, что завышающий баллы респондент получает несправедливое преимущество перед другими кандидатами. Однако если кандидат ставит перед собой цель продемонстрировать определенный профиль ответов (*faking*), то оценка СЖ и корректировка итоговых баллов малоэффективны в минимизации рисков завышения баллов.

В-третьих, респонденты, которые имеют большой опыт прохождения оценочных процедур, способны идентифицировать утверждения, направленные на измерение СЖ, и могут сознательно занизить баллы по этим утверждениям, чтобы не быть «пойманными» [Larson, 2018]. Мы постарались максимально приблизить формулировки этих утверждений к формулировкам утверждений по компетенциям.

6. Заключение и дискуссия

Измерение и контроль СЖ остаются одной из значимых проблем при оценивании психологических конструктов с помощью опросных методов [Grimm, 2010]. При этом, как резонно заметил Т. Трейси [Tracey, 2016], вместо того чтобы безосновательно

предполагать, что измерение и контроль СЖ необходимы, исследователи должны объяснять, почему контроль СЖ улучшит результаты исследования. Использование показателей СЖ с неизвестными статистическими свойствами для корректировки данных с неизвестной степенью систематической ошибки может принести больше вреда, чем пользы [Uziel, 2010]. Поэтому стоит крайне внимательно подходить к разработке шкал СЖ с целью их использования в качестве корректирующих мер. Мы считаем, что для нашего опросника контроль СЖ необходим, поскольку на его основе будут приниматься значимые для жизни респондентов решения. В таких условиях возрастают риски необоснованного завышения баллов и тем самым получения тем или иным респондентом несправедливого преимущества перед другими кандидатами, что подтверждается значимыми корреляциями как эгоистического, так и моралистического компонента СЖ с большинством измеряемых конструктов, и эти данные согласуются с результатами предыдущих исследований [Moorman, Podsakof, 1992].

Главная цель проведенного исследования состояла в создании такой математической модели обработки данных, которая бы минимизировала эффект СЖ при измерении компетенций и имела бы по крайней мере удовлетворительные психометрические показатели. Использование бифакторных моделей и расчет факторных баллов в целом позволили достичь этой цели. Эффективность используемого подхода к минимизации влияния СЖ подтверждена двумя индикаторами: снижением корреляций между компетенциями после внесения корректировок и снижением факторных баллов у симулированных ответных профилей, одинаковых по компетенциям и различающихся только по уровню СЖ. Однако существенным недостатком такого подхода стала сравнительно большая ошибка измерения. Вероятно, ее величина обусловлена использованием достаточно коротких шкал при измерении компетенций — от 3 до 8 утверждений. Шкала измерения СЖ при этом состояла из 10 утверждений, из-за чего могла увеличиться доля объясненной шкалой СЖ дисперсии по этим компетенциям, несмотря на изначально хорошие психометрические показатели. Возможно, если бы количество утверждений в шкалах было более сбалансированным, статистические показатели были бы лучше. При этом уменьшение количества утверждений в шкале эгоистической СЖ не привело к существенному улучшению показателей моделей, поэтому мы приняли решение не сокращать шкалу, чтобы измерять СЖ более полно с точки зрения содержания конструкта.

Другим ожидаемым результатом стала более сильная связь эгоистического компонента СЖ с измеряемыми компетенциями

по сравнению с моралистическим. Возможное объяснение состоит в том, что большинство измеряемых компетенций характеризуют респондента в качестве профессионала, в то время как моралистический компонент СЖ скорее связан со стремлением респондента выглядеть человеком, соблюдающим моральные нормы [Paulhus, 1991; 1998; Salgado, 2005]. Вероятно, поэтому такие компетенции, как, например, «Работает в команде» или «Поддерживает других людей», по результатам данного исследования оказались сильнее связаны с моралистической СЖ: поведение, соответствующее этим компетенциям, стимулируется скорее мотивацией общения, чем стремлением получить власть и признание. При этом связь обоих компонентов СЖ с измеряемыми компетенциями оказалась умеренной, что также могло стать одним из факторов возникновения большой ошибки измерения при использовании бифакторных моделей. Такая связь может объясняться общим фактором личности [Musek, 2007; Van der Linden, te Nijenhuis, Bakker, 2010], артефактами ответных стилей [Ashton, Lee, De Vries, 2014; Bäckström, Björklund, Larsson, 2009] или тем, что респонденты могут действительно вести себя социально желательным образом [Osin, 2009], т.е. СЖ может проявляться не только в ответах респондентов, но в их поведении в рабочем контексте.

Отдельного внимания заслуживают результаты, которые свидетельствуют о перспективности с точки зрения минимизации эффектов СЖ разработки таких формулировок, которые отражают наиболее яркую положительную сторону измеряемых индикаторов. Традиционно минимизация влияния СЖ на этапе разработки утверждений состоит в нейтрализации формулировок с таким расчетом, чтобы они не провоцировали стремление отвечать социально желательным образом [Jackson, 1984; Bäckström, Björklund, 2020]. Сталкиваясь с максимально яркими формулировками, респонденты, вероятно, опасаются приукрашивать положение вещей, поскольку считают, что завышение баллов будет выглядеть слишком явным и их «поймают» — притом что в инструкции к опроснику есть соответствующее предупреждение. Однако это предположение должно быть проверено на выборке в условиях реальных высоких ставок. Дополнительно при использовании обеих стратегий необходимо удостовериться в том, что утверждения по-прежнему измеряют изначально интересующий конструкт с точки зрения его содержания.

Многие разработчики тестов неохотно удаляют дисперсию, связанную с СЖ, поскольку есть данные, что в результате таких действий снижается критериальная [McCrae, Costa, 1983; Hough, 1998] и конструктная валидность [Ones, Viswesvaran, 2001]. Однако эти данные получены преимущественно на материале

опросника *Big Five* и при использовании более простых методов внесения корректировок в итоговые баллы. Данная работа и дальнейшие исследования с использованием бифакторных моделей могут внести вклад в дискуссию о ценности учета СЖ при измерении различных психологических конструктов и о связи СЖ с валидностью инструментов измерения.

Благодарности Исследование проведено при финансовой поддержке ООО «Форматта».

References

Allport G.W. (1937) *Personality: A Psychological Interpretation*. New York, NY: Henry Holt and Company.

Anderson J.R. (1976) *Language, Memory, and Thought*. New York, NY: Lawrence Erlbaum.

Anglim J., Morse G., de Vries R.E., MacCann C., Marty A. (2017) Comparing Job Applicants to Non-Applicants Using an Item-Level Bifactor Model on the HEXACO Personality Inventory. *European Journal of Personality*, vol. 31, no 6, pp. 669–684. https://doi.org/10.1002/per.2120

Anusic I., Schimmack U., Pinkus R.T., Lockwood P. (2009) The Nature and Structure of Correlations among Big Five Ratings: The Halo-Alpha-Beta Model. *Journal of Personality and Social Psychology*, vol. 97, no 6, pp. 1142–1156. https://doi.org/10.1037/a0017159

Ashton M.C., Lee K., de Vries R.E. (2014) The HEXACO Honesty-Humility, Agreeableness, and Emotionality Factors: A Review of Research and Theory. *Personality and Social Psychology Review*, vol. 18, no 2, pp. 139–152. https://doi.org/10.1177/1088868314523838

Bäckström M., Björklund F. (2020) The Properties and Utility of Less Evaluative Personality Scales: Reduction of Social Desirability; Increase of Construct and Discriminant Validity. *Frontiers in Psychology*, vol. 11, October, Article no 560271. https://doi.org/10.3389/fpsyg.2020.560271

Bäckström M., Björklund F., Larsson M.R. (2009) Five-Factor Inventories Have a Major General Factor Related to Social Desirability Which Can Be Reduced by Framing Items Neutrally. *Journal of Research in Personality*, vol. 43, no 3, pp. 335–344. https://doi.org/10.1016/j.jrp.2008.12.013

Biderman M.D., Nguyen N.T., Cunningham C.J., Ghorbani N. (2011) The Ubiquity of Common Method Variance: The Case of the Big Five. *Journal of Research in Personality*, vol. 45, no 5, pp. 417–429. https://doi.org/10.1016/j.jrp.2011.05.001

Birkeland S.A., Manson T.M., Kisamore J.L., Brannick M.T., Smith M.A. (2006) A Meta-Analytic Investigation of Job Applicant Faking on Personality Measures. *International Journal of Selection and Assessment*, vol. 14, no 4, pp. 317–335. https://doi.org/10.1111/j.1468-2389.2006.00354.x

Bowen C., Martin B.A., Hunt S.T. (2002) A Comparison of Ipsative and Normative Approaches for Ability to Control Faking In Personality Questionnaires. *The International Journal of Organizational Analysis*, vol. 10, no 3, pp. 240–259. https://doi.org/10.1108/eb028952

Brown T.A. (2015) *Confirmatory Factor Analysis for Applied Research*. New York, NY: Guilford.

Brown A., Maydeu-Olivares A. (2013) How IRT Can Solve Problems of Ipsative Data in Forced-Choice Questionnaires. *Psychological Methods*, vol. 18, no 1, pp. 36–52. https://doi.org/10.1037/a0030641

Bryan C. J., Adams G. S., Monin B. (2013) When Cheating Would Make You a Cheater: Implicating the Self Prevents Unethical Behavior. *Journal of Experimental Psychology: General*, vol. 142, no 4, pp. 1001–1005. https://doi.org/10.1037/a0030655.supp

Cattell H.E.P., Mead A.D. (2008) The Sixteen Personality Factor Questionnaire (16PF). *The Sage Handbook of Personality Theory and Assessment. Vol. 2. Personality Measurement and Testing* (eds G.J. Boyle, G. Matthews, D.H. Saklofske), Los Angeles, CA: Sage, pp. 135–159. https://doi.org/10.4135/9781849200479.n7

Chen Z., Watson P., Biderman M., Ghorbani N. (2016) Investigating the Properties of the General Factor (M) in Bifactor Models Applied to Big Five or HEXACO Data in Terms of Method or Meaning. *Imagination, Cognition and Personality*, vol. 35, June, pp. 216–243. https://doi.org/10.1177/0276236615590587

Christiansen N.D., Burns G.N., Montgomery G.E. (2005) Reconsidering Forced-Choice Item Formats for Applicant Personality Assessment. *Human Performance*, vol. 18, no 3, pp. 267–307. https://doi.org/10.1207/s15327043hup1803_4

Costa P.T., McCrae R.R. (1992) The Five-Factor Model of Personality and Its Relevance to Personality Disorders. *Journal of Personality Disorders*, vol. 6, no 4, pp. 343–359. https://doi.org/10.1521/pedi.1992.6.4.343

Crowne D.P., Marlowe D. (1960) A New Scale of Social Desirability Independent of Psychopathology. *Journal of Consulting Psychology*, vol. 24, no 4, pp. 349–354. https://doi.org/10.1037/h0047358

DiStefano C., Zhu M., Mîndrilă D. (2009) Understanding and Using Factor Scores: Considerations for the Applied Researcher. *Practical Assessment, Research & Evaluation*, vol. 14, no 20. https://doi.org/10.7275/da8t-4g52

Eysenck S.B.G., Eysenck H.J. (1964) Personality of Judges as a Factor in the Validity of Their Judgments of Extraversion-Introversion. *British Journal of Social & Clinical Psychology*, vol. 3, no 2, pp. 141–148. https://doi.org/10.1111/j.2044-8260.1964.tb00418.x

Ferrando P.J., Lorenzo-Seva U., Chico E. (2009) A General Factor-Analytic Procedure for Assessing Response Bias in Questionnaire Measures. *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 16, no 2, pp. 364–381. https://doi.org/10.1080/10705510902751374

García-Izquierdo A.L., Ramos-Villagrasa P.J., Lubiano M.A. (2020) Developing Biodata for Public Manager Selection Purposes: A Comparison between Fuzzy Logic and Traditional Methods. *Revista de Psicología del Trabajo y de las Organizaciones*, vol. 36, no 3, pp. 231–242. https://doi.org/10.5093/jwop2020a22

Goffin R.D., Christiansen N.D. (2003) Correcting Personality Tests for Faking: A Review of Popular Personality Tests and an Initial Survey of Researchers. *International Journal of Selection and Assessment*, vol. 11, no 4, pp. 340–344. https://doi.org/10.1111/j.0965-075x.2003.00256.x

Golubovich J., Lake C.J., Anguiano-Carrasco C., Seybert J. (2020) Measuring Achievement Striving via a Situational Judgment Test: The Value of Additional Context. *Journal of Work and Organizational Psychology*, vol. 36, no 2, pp. 157–168. https://doi.org/10.5093/jwop2020a15

Grimm P. (2010) Social Desirability Bias. *Wiley International Encyclopedia of Marketing* (eds J. Sheth, N. Malhotra), Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9781444316568.wiem02057

Heggestad E.D., Morrison M., Reeve C.L., McCloy R.A. (2006) Forced-Choice Assessments of Personality for Selection: Evaluating Issues of Normative Assessment and Faking Resistance. *Journal of Applied Psychology*, vol. 91, no 1, pp. 9–24. https://doi.org/10.1037/0021-9010.91.1.9

Hicks L.E. (1970) Some Properties of Ipsative, Normative, and Forced-Choice Normative Measures. *Psychological Bulletin*, vol. 74, no 3, pp. 167–184. https://doi.org/10.1037/h0029780

Holzinger K.J., Swineford F. (1937) The Bi-Factor Method. *Psychometrika*, vol. 2, March, pp. 41–54. https://doi.org/10.1007/BF02287965

Hough L. (1998) Effects of Intentional Distortion in Personality Measurement and Evaluation of Suggested Palliatives. *Human Performance*, vol. 11, no 2, pp. 209–244. https://doi.org/10.1207/s15327043hup1102&3_6

Huang J.L., Curran P.G., Keeney J., Poposki E.M., DeShon R.P. (2011) Detecting and Deterring Insufficient Effort Responding to Surveys. *Journal of Business and Psychology*, vol. 27, no 1, pp. 99–114. https://doi.org/10.1007/s10869-011-9231-8

Jackson D.N. (1984) Personality Research from Manual. Port Huron, MI: Research Psychologists.

Kam C.C.S. (2020) Bifactor Model Is Not the Best-Fitting Model for Self-Esteem: Investigation with a Novel Technique. *Assessment*, vol. 28, no 7, Article no 1073191120949916. https://doi.org/10.1177/107319112094991

Kreitchmann R.S., Abad F.J., Ponsoda V., Nieto M.D., Morillo D. (2019) Controlling for Response Biases in Self-Report Scales: Forced-Choice vs Psychometric Modeling of Likert Items. *Frontiers in Psychology*, vol. 10, October, Article no 2309. https://doi.org/10.3389/fpsyg.2019.02309

Krosnick J.A. (1999) Survey research. *Annual Review of Psychology*, vol. 50, pp. 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

Kuzminov Y., Sorokin P., Froumin I. (2019) Generic and Specific Skills as Components of Human Capital: New Challenges for Education Theory and Practice. *Foresight and STI Governance*, vol. 13, no 2, pp. 19–41. https://doi.org/10.17323/2500-2597.2019.2.19.41

Larson R.B. (2018) Controlling Social Desirability Bias. *International Journal of Market Research*, vol. 61, no 5, pp. 534–547. https://doi.org/10.1177/1470785318805305

Lee M.D., Criss A.H., Devezer B., Donkin C., Etz A., Leite F.P. et al. (2019) *Robust Modeling in Cognitive Science*. Paper presented at Workshop on Robust Social Science (St. Petersburg, FL, 2018, June). https://doi.org/10.31234/osf.io/dmfhk

Martínez A., Moscoso S., Lado M. (2021) Faking Effects on the Factor Structure of a Quasi-Ipsative Forced-Choice Personality Inventory. *Journal of Work and Organizational Psychology*, vol. 37, no 1, pp. 1–10. https://doi.org/10.5093/jwop2021a7

McCrae R.R., Costa P.T. (1983) Social Desirability Scales: More Substance Than Style. *Journal of Consulting and Clinical Psychology*, vol. 51, no 6, pp. 882–888. https://doi.org/10.1037/0022-006x.51.6.882

McFarland L.A., Ryan A.M. (2006) Toward an Integrated Model of Applicant Faking Behavior. *Journal of Applied Social Psychology*, vol. 36, no 4, pp. 979–1016. https://doi.org/10.1111/j.0021-9029.2006.00052.x

Meade A.W. (2004) Psychometric Problems and Issues Involved with Creating and Using Ipsative Measures for Selection. *Journal of Occupational and Organizational Psychology*, vol. 77, no 4, pp. 531–551. https://doi.org/10.1348/0963179042596504

Moorman R.H., Podsakoff P.M. (1992) A Meta-Analytic Review and Empirical Test of the Potential Confounding Effects of Social Desirability Response Sets in Organizational Behaviour Research. *Journal of Occupational and Organizational Psychology*, vol. 65, no 2, pp. 131–149. https://doi.org/10.1111/j.2044-8325.1992.tb00490.x

Mueller-Hanson R., Heggestad E.D., Thornton G.C. (2003) Faking and Selection: Considering the Use of Personality from Select-In and Select-Out Perspectives. *Journal of Applied Psychology*, vol. 88, no 2, pp. 348–355. https://doi.org/10.1037/0021-9010.88.2.348

Musek J. (2007) A General Factor of Personality: Evidence for the Big One in the Five-Factor Model. *Journal of Research in Personality*, vol. 41, no 6, pp. 1213–1233. https://doi.org/10.1016/j.jrp.2007.02.003

Neeley S.M., Cronley M.L. (2004) When Research Participants Don`t Tell It Like It Is: Pinpointing the Effects of Social Desirability Bias Using Self vs Indirect-Questioning. *ACR North American Advances*, vol. 31, pp. 432–433.

Ones D.S., Viswesvaran C. (2001) Integrity Tests and Other Criterion-Focused Occupational Personality Scales (COPS) Used in Personnel Selection. *International Journal of Selection and Assessment*, vol. 9, no 1–2, pp. 31–39. https://doi.org/10.1111/1468-2389.00161

Osin E.N. (2009) Social Desirability in Positive Psychology: Bias or Desirable Sociality? *Understanding Positive Life. Research and Practice on Positive Psychology* (ed. T. Freire), Lisbon: Climepsi Editores, pp. 421–442.

Paulhus D.L. (1998) *Manual for the Paulhus Deception Scales: BIDR Version 7*. Toronto, Canada: Multi-Health Systems.

Paulhus D.L. (1991) Measurement and Control of Response Bias. *Measures of Personality and Social Psychological Attitudes* (eds J.P. Robinson, P. Shaver, L.S. Wrightsman), San Diego, CA: Academic Press, pp. 17–59.

Paulhus D.L., Vazire S. (2007) The Self-Report Method. *Handbook of Research Methods in Personality Psychology* (eds R.W. Robins, R.C. Fraley, R.F. Krueger), New York; London: The Guilford, pp. 224–239.

Reise S.P. (2012) The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research*, vol. 47, no 5, pp. 667–696. https://doi.org/10.1080/00273171.2012.715555

Rolland J.P., Mogenet J.L. (2001) *Système de Description en Cinq Dimensions (D5D). Manuel Réservé aux Psychologues*. Paris: Les Editions du Centre de Psychologie Appliquée.

Sackett P.R., Lievens F., van Iddekinge C.H., Kuncel N.R. (2017) Individual Differences and Their Measurement: A Review of 100 Years of Research. *Journal of Applied Psychology*, vol. 102, no 3, pp. 254–273. https://doi.org/10.1037/apl0000151

Salgado F.J. (2005) Personality and Social Desirability in Organizational Settings: Practical Implications for Work and Organizational Psychology. *Papeles del Psicologo*, vol. 26, January, pp. 115–128.

Salgado J.F. (2016) A Theoretical Model of Psychometric Effects of Faking on Assessment Procedures: Empirical Findings and Implications for Personality at Work. *International Journal of Selection and Assessment*, vol. 24, no 3, pp. 209–228. https://doi.org/10.1111/ijsa.12142

SHL (1999) *OPQ32 Manual and User's Guide*. Thames Ditton, Surrey: SHL Group.

Thurstone L.L. (1935) *The Vectors of Mind*. Chicago, IL: University of Chicago.

Tracey T.J. (2016) A Note on Socially Desirable Responding. *Journal of Counseling Psychology*, vol. 63, no 2, pp. 224–232. https://doi.org/10.1037/cou0000135

Uziel L. (2010) Rethinking Social Desirability Scales: From Impression Management to Interpersonally Oriented Self-Control. *Perspectives on Psychological Science*, vol. 5, no 3, pp. 243–262. https://doi.org/10.1177/1745691610369465

Van der Linden D., te Nijenhuis J., Bakker A.B. (2010) The General Factor of Personality: A Meta-Analysis of Big Five Intercorrelations and a Criterion-Related Validity Study. *Journal of Research in Personality*, vol. 44, no 3, pp. 315–327. https://doi.org/10.1016/j.jrp.2010.03.003

Viladrich C., Angulo-Brunet A., Doval E. (2017) A Journey around Alpha and Omega to Estimate Internal Consistency Reliability. *Anales de Psicología*, vol. 33, no 3, pp. 755–782. https://doi.org/10.6018/analesps.33.3.268401

Измерение образовательного прогресса на основе когнитивных операций

Сергей Тарасов, Ирина Зуева, Денис Федерякин

Статья поступила в редакцию в марте 2023 г.

Тарасов Сергей Владимирович — аспирант, стажер-исследователь Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики». Адрес: 101000, Москва, Потаповский пер., 16, стр. 10. E-mail: svtarasov@hse.ru. ORCID: https://orcid.org/0000-0003-4151-115X (контактное лицо для переписки)

Зуева Ирина Олеговна — аспирант, аналитик Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики». E-mail: izueva@hse.ru

Федерякин Денис Александрович — научный сотрудник департамента экономического образования, Университет им. Иоганна Гутенберга (Майнц, Германия). E-mail: denis.federiakin@uni-mainz.de

Аннотация

Измерение образовательного прогресса остается нетривиальной методологической задачей даже при наличии множества описанных в литературе подходов к его концептуализации и моделированию. Рассматривается методологический подход к измерению образовательного прогресса в рамках современной теории тестирования, при этом традиционная концептуализация этого подхода расширяется за счет моделирования когнитивных операций. Показано, что синтез традиционных моделей для измерения образовательного прогресса с одной из самых популярных моделей современной теории тестирования — LLTM, позволяющей моделировать когнитивные операции, — существенно обогащает возможности интерпретации тестовых баллов учеников, сохраняя все достоинства традиционного подхода к измерению образовательного прогресса. Для иллюстрации предлагаемого подхода использована линейка тестов, применявшихся для мониторинга образовательного прогресса в математике в 8–9-х классах средней школы.

Ключевые слова

измерение образовательного прогресса, IRT, LLTM, когнитивные операции, диагностические критериально-ориентированные пороги

Для цитирования

Тарасов С.В., Зуева И.О., Федерякин Д.А. (2023) Измерение образовательного прогресса на основе когнитивных операций. *Вопросы образования / Educational Studies Moscow*, № 3, сс. 172–196. https://doi.org/10.17323/vo-2023-16902

Measuring Learning Progress Based on Cognitive Operations

Sergei Tarasov, Irina Zueva, Denis Federiakin

Sergei V. Tarasov — PhD Student, Research Assistant at the Center for Psychometrics and Measurement in Education, National Research University Higher School of Economics. Address: Bld. 10, 16 Potapovsky Ln, 101000 Moscow, Russian Federation. E-mail: svtarasov@hse.ru. ORCID: https://orcid.org/0000-0003-4151-115X (corresponding author)

Irina O. Zueva — PhD Student, Analyst at the Center for Psychometrics and Measurement in Education, Institute of Education, National Research University Higher School of Economics. E-mail: izueva@hse.ru

Denis A. Federiakin — Research Assistant at the Department of Economic Education, Johannes Gutenberg University (Mainz). E-mail: denis.federiakin@uni-mainz.de

Abstract Measuring students' growth and change is considered one of the main ways for evidence-based development of educational systems. However, it is a non-trivial methodological task, despite the numerous approaches available for its conceptualization and statistical realization. In this article, we describe the main features of measuring students' growth and change using Item Response Theory (IRT) in detail. We then expand this approach to allow for the modeling of cognitive operations with the Linear Logistic Test Model (LLTM). We show that the synthesis of traditional IRT models for measuring growth and change with LLTM significantly enriches the interpretability of ability estimates while preserving the advantages of the traditional approach. To illustrate this approach, we use a set of monitoring tests to measure educational progress in mathematics in secondary school.

Keywords measurement of progress, IRT, LLTM, vertical alignment, cognitive operations, diagnostic thresholds

For citing Tarasov S.V., Zueva I.O., Federiakin D.A. (2023) Izmerenie obrazovatel´nogo progress ana osnove kognitivnykh operatsiy [Measuring Learning Progress Based on Cognitive Operations]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 3, pp. 172–196. https://doi.org/10.17323/vo-2023-16902

Измерение образовательного прогресса — один из основных способов доказательного развития образования. Образовательный прогресс дает основания судить о том, какие образовательные практики и политики положительно ассоциированы с повышением уровня образовательных достижений у учащихся [Slavin, 2005]. Однако создание условий, необходимых для того, чтобы доказательно говорить об измерении образовательного прогресса, представляет собой нетривиальную методологическую задачу. Методологи предложили немало методов измерения прогресса [Саляхутдинова, Федерякин, 2022]. При всем их разнообразии в образовании доминирующим подходом остается современная теория тестирования благодаря ее интуитивной понятности и относительной технической простоте.

В отличие от классической теории тестирования, работающей с суммой первичных баллов за задания, современная теория тестирования (*Item Response Theory*, IRT) [Linden van der, 2018] подразумевает математическое разделение параметров заданий и респондентов. Из-за того что IRT обращается с этими параметрами как с независимыми — или, по крайней мере, отдельными, — этот подход к образовательным измерениям открывает много возможностей, которые нельзя реализовать в классическом подходе. IRT не только включает в математическую модель для оценки уровня способностей респондентов множество дополнительных факторов и аспектов оценки, но и позволяет доказательно говорить об измерении образовательного прогресса. С использованием IRT можно моделировать когнитивные структуры, задействованные в процессе решения заданий, например с помощью оценивания трудностей различных когнитивных операций [Fischer, 1973], — а значит, выяснять, какие когнитивные операции респондент освоил, а какие нет.

Цель данной работы состоит в том, чтобы проиллюстрировать новый подход к измерению образовательного прогресса с помощью когнитивных операций, задействованных в процессе решения заданий теста, в рамках парадигмы IRT. В статье мы прежде всего обосновываем невозможность измерения образовательного прогресса без использования специальных психометрических методов. Далее мы описываем один из подходов к измерению образовательного прогресса в IRT и обсуждаем его соотношение с другими способами измерения прогресса. Затем мы рассматриваем возможности, которые предоставляет IRT для анализа когнитивных операций, заложенных в тест, и синтезируем эти две области IRT, чтобы описать методологию измерения образовательного прогресса на основе когнитивных операций. И наконец, мы приводим пример реализации предложенного подхода на данных мониторинга индивидуального образовательного прогресса в математике учащихся 8–9-х классов и вписываем результаты этого исследования в более широкий исследовательский контекст.

1. Проблемы с измерением образовательного прогресса в классических подходах

Чтобы измерение образовательного прогресса было возможным, необходимы следующие условия:

- наличие нескольких (минимум двух) замеров одной и той же способности на определенной временной дистанции;
- наличие якорных заданий в разных тестах;
- применение специальных методов психометрического моделирования.

Первое условие очевидно. Выполнение второго и третьего условий — именно то, за что подвергаются критике попытки измерить образовательный прогресс в классической теории тестирования.

Наличие якорных заданий дает возможность проверить, действительно ли разные тесты все еще измеряют одну и ту же способность [Waterbury, DeMars, 2021]. Якорными называются абсолютно одинаковые задания, включенные в разные тесты (замеры). Если образовательный прогресс измеряется на длительной дистанции и проводится больше двух замеров, то якорные задания в разных парах замеров могут быть разными, т.е. между первым и вторым замером будут одни якорные задания, а между вторым и третьим — другие; причем между первым и третьим замером может не быть якорных заданий. Дело в том, что содержание образования меняется, поэтому темы из первого замера могут не пересекаться с третьим замером, в то время как между соседними замерами всегда должна находиться область пересечения. Таким образом, постепенное изменение содержания тестов позволяет одновременно сохранить непрерывность шкалы способности и обеспечить гибкость ее интерпретации.

Когда претест и посттест являются абсолютно одинаковыми [Dimitrov, Rumrill, 2003], все задания могут рассматриваться как якорные, что гарантирует полную сопоставимость тестовых баллов даже без применения психометрического моделирования. Однако если образовательный прогресс измеряется на длительной дистанции, при предъявлении ученикам одного и того же множества заданий, например, в начале 1-го и в конце 4-го класса возникнут потолочные эффекты и достоверное измерение образовательного прогресса будет подавлено статистическими артефактами. Таким образом, одинаковые претест и посттест можно применять только для оценки образовательного прогресса на краткосрочной дистанции, например в течение учебной четверти, и только с помощью двух замеров подряд, потому что при большем количестве замеров становится невозможно игнорировать эффект знакомства респондентов с заданиями. На практике такое измерение образовательного прогресса встречается при проведении краткосрочных экспериментальных интервенций, когда требуется сравнить образовательные достижения учащихся, чтобы оценить эффекты новых образовательных практик и политик.

Когда якорных заданий в следующих один за другим замерах нет, происходит «разрыв» шкалы. Если претест и посттест состоят из непересекающихся множеств заданий, то с точки зрения психометрики такая ситуация аналогична использованию в претесте и посттесте тестов по разным предметам. Даже

если претест и посттест разработаны в одной и той же теоретической рамке и спецификации, результаты повторного тестирования совершенно не поддаются интерпретации, так как одни и те же баллы, полученные в двух тестах, не говорят об одинаковой способности ученика. Только наличие якорных заданий обеспечивает сопоставимость оценок способности, полученных в разных замерах. В следующем разделе статьи будет рассмотрен еще один аргумент в пользу наличия якорных заданий — математический.

Для измерения образовательного прогресса недостаточно просто включить якорные задания в соседние замеры. Чтобы сравнивать тестовые баллы из соседних замеров, необходимо обеспечить единство интерпретации этих тестовых баллов, т.е. добиться, чтобы один и тот же тестовый балл соответствовал одной и той же истинной способности респондента [Loyd, Hoover, 1980]. Единства в оценивании результатов двух разных замеров невозможно достичь без применения психометрического моделирования в IRT, так как в классической теории тестирования наблюдаемый балл респондента может интерпретироваться только внутри одного теста относительно конкретного набора заданий. Рассмотрим случай, когда из 30 заданий в претесте и 30 заданий в посттесте 10 являются якорными, а 20 заданий — уникальными для каждого замера. Если респондент набрал условные 15 баллов в претесте и 10 в посттесте, это совсем не обязательно означает отрицательный образовательный прогресс. Возможно, дело в том, что уникальные задания в посттесте были труднее, чем в претесте, — именно так обычно и конструируют задания при разработке инструментов для измерения образовательного прогресса. В классическом подходе сравнение тестовых баллов, полученных на хоть сколько-нибудь отличающихся друг от друга множествах заданий, невозможно. То есть одно и то же количество первичных баллов в претесте и посттесте не соответствует одному и тому же уровню способности. Таким образом, единства интерпретации шкалы невозможно достичь без психометрического моделирования.

2. Применение современной теории тестирования для измерения образовательного прогресса

Одна из самых сложных проблем в измерении образовательного прогресса — его концептуализация. Достаточно понятна процедура измерения прогресса, например, в физкультуре: высота прыжка или количество попаданий баскетбольным мячом в корзину представляют собой элементарные наблюдаемые метрики, интерпретация которых совершенно одинакова как для учеников начальной школы, так и для взрослых людей. Однако концептуализировать когнитивный прогресс чрезвычайно трудно, потому что это латентный процесс. Его обычно не

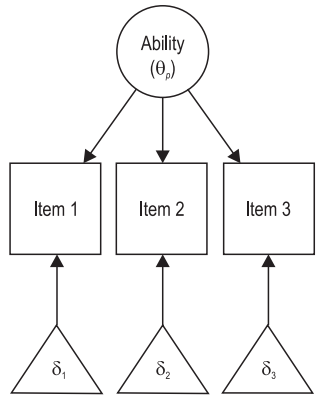
удается наблюдать в поведении учеников, а если и удается, то его количественное измерение составляет большую проблему. В частности, для преодоления этих трудностей и разрабатывался математический аппарат IRT [Sontag, 1984].

Ключевой постулат IRT — математическое разделение параметров респондента и задания. Пример — формула дихотомической модели Раша:

$$P(U_{pi} = 1) = \frac{\exp(\theta_p - \delta_i)}{1 + \exp(\theta_p - \delta_i)}, \quad (1)$$

где U_{pi} — ответ респондента p ϵ (1, 2, ..., P) на задание i ϵ (1, 2, ..., I); $P(U_{pi} = 1)$ — вероятность того, что этот ответ будет равен единице (вероятность правильного ответа); θ_p — оценка латентной способности респондента p на логит-шкале (шкале логарифмических шансов); δ_i — оценка латентной трудности задания i на логит-шкале. Эту модель можно представить с помощью путевой диаграммы (рис. 1).

Рис. 1. **Путевая диаграмма модели Раша**



Примечание: *Ability* — способность, *Item* — задание, кругом обозначена оцениваемая способность респондентов (латентная переменная), квадратами — ответы на задания теста (наблюдаемые переменные), треугольниками — параметры заданий (трудности). Стрелками с одним направлением обозначены регрессионные зависимости («эффект применяется к...»).

В моделях IRT способности респондентов и трудности заданий располагаются и сравниваются на одной метрической шкале. Параметры заданий и параметры респондентов оцениваются отдельно во всех моделях IRT, но только в моделях Раша они сравниваются напрямую без дополнительных модификаций, что позволяет достичь полного разделения этих параметров. Модель из уравнения (1) допускает, что вероятность ответа респондента на задание зависит от того, как соотносятся

уровень способности респондента и уровень трудности здания. Если уровень трудности задания выше, чем уровень способности респондента, вероятность правильного ответа будет меньше 50%. Если уровень способности респондента превышает трудность задания, вероятность правильного ответа больше 50%. Такой способ формулирования процесса ответа на математическом языке обладает интересными математическими свойствами [Andersen, 1977], лежащими в основе современных подходов к измерениям в социальных науках. Однако для целей измерения образовательного прогресса мы заинтересованы только в одном из них: в разделении вариативности ответов «внутри» респондента и между респондентами. Параметр θ_p описывает различия между респондентами, ранжируя их от самых слабых до самых сильных, а параметр δ_i отражает различия «внутри» респондентов, ранжируя задания от самых простых до самых трудных. Различия заданий по параметру δ_i показывают, насколько для любого респондента одно задание труднее другого, — отсюда и интерпретация этого параметра именно «внутри» респондента. Если одно задание труднее другого, оно будет труднее для всех респондентов, т.е. вариация «внутри» респондентов одинаковая (δ_i применяется ко всем респондентам), а все различия между ними вынесены в другой параметр модели — θ_p. И тогда вероятность наблюдать правильный ответ в разных заданиях служит прокси для измерения параметра интереса — θ_p.

Одну из первых моделей для измерения лонгитюдных изменений предложил Э. Андерсен [Andersen, 1985]. Математически она представляет собой многомерное расширение модели Раша на случай с якорными заданиями между замерами:

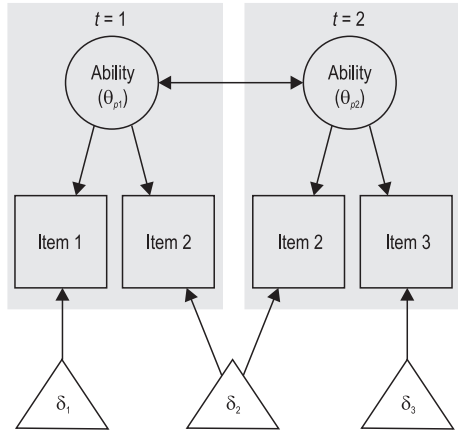
$$P(U_{pit}=1)=\frac{\exp(\theta_{pt}-\delta_i)}{1+\exp(\theta_{pt}-\delta_i)}, \quad (2)$$

где U_{pit} — ответ респондента p на задание i в момент замера $t \in (1, 2, ..., T)$; θ_{pt} — оценка латентной способности респондента p в момент замера t на логит-шкале; δ_i — оценка латентной трудности задания i на логит-шкале.

Путевая диаграмма этой модели представлена на рис. 2.

В отличие от модели Раша, модель Андерсена является многомерной: она содержит столько размерностей, по которым различаются респонденты, сколько было проведено замеров. По каждой из размерностей одни и те же респонденты ранжированы по-разному. При этом от простых многомерных моделей модель Андерсена отличается фиксацией параметров повторяющихся заданий, нагружающих разные размерности респондентов (предъявленных в разные замеры), к одним и тем же значениям. Таким образом, якорные задания имеют

Рис. 2. **Путевая диаграмма модели Андерсена**



Примечание (в дополнение к примечаниям к рис. 1): Двусторонней стрелкой обозначена корреляция, закрашенными областями обозначены разные замеры.

один и тот же уровень трудности во всех замерах — и это условие составляет основу для концептуализации образовательного прогресса. В современной теории тестирования образовательный прогресс понимается как изменение вероятности решения якорных заданий от одного замера к другому. Тот факт, что параметры заданий отделены от параметров респондентов в уравнении модели, позволяет использовать θ_p в простой модели Раша как условную агрегацию вероятности решения заданий на уровень респондента. Изменения этой вероятности, переведенные на логит-шкалу, показывают, как меняется способность респондента. Использование якорных заданий для этой модели — критически необходимое условие сопоставимости шкал соседних замеров: трудность якорных заданий оценивается относительно распределения выборки в первом замере (и не меняется), а во втором (и/или в последующих замерах) уже относительно этих заданий оценивается положение выборки на той же шкале, что и в первом замере. Таким способом обеспечивается соответствие одного и того же значения параметров θ_{pt} для всех замеров одному и тому же уровню способности, что позволяет доказательно говорить об образовательном прогрессе.

Параллельно с отслеживанием образовательного прогресса эта модель позволяет анализировать психометрические характеристики как отдельных замеров, так и всего множества заданий, чтобы убедиться, что измерительный инструментарий соответствует необходимым стандартам качества, а также дает возможность делать индивидуальные и/или групповые выводы о респондентах.

3. Другие возможности измерения образовательного прогресса в современной теории тестирования

Один из наиболее популярных способов измерения прогресса — вертикальное выравнивание [Sontag, 1984; Саляхутдинова, Федерякин, 2022]. В отличие от модели Андерсена, вертикальное выравнивание допускает, что, несмотря на повторяющиеся замеры, множество заданий, используемых для измерения прогресса, является одномерным. То есть здесь отдельные замеры не рассматриваются как отдельные размерности. Такой подход существенно упрощает психометрическое моделирование, используемое для вертикального выравнивания, но предъявляет гораздо более высокие требования к качеству измерительного инструментария: он должен быть нечувствителен к тому, что ответы одного и того же респондента на якорные задания скоррелированы в разные моменты времени, т.е. высока вероятность того, что в каждый следующий замер респондент ответит на якорный вопрос так же, как ответил, столкнувшись с этим вопросом первый раз (или похожим образом). Модель Андерсена учитывает эту вероятность, позволяя оценкам способности из разных замеров коррелировать. При этом вертикальное выравнивание, как метод пост-хок размещения оценок респондентов на одной шкале, может не предполагать установления пороговых баллов. Соответственно этот метод может не предусматривать обогащения интерпретации баллов содержанием образовательного прогресса.

В IRT есть модели для измерения прогресса, в основе которых лежит внутренняя многомерность заданий, в этом случае в заданиях из последующих замеров проявляются также уровни способности из всех предыдущих замеров: например, модель С. Эмбретсон для роста и измерения (*model for measuring growth and change*) [Embretson, 1991]) или экспланаторная полиномиальная модель латентного роста (*explanatory polynomial model of latent growth*) [Wilson, Zheng, McGuire, 2012]). Они позволяют моделировать специфические изменения в латентной способности, в то время как модель Андерсена отражает просто положение респондента на единой шкале. Соответственно в модели Андерсена, чтобы вычислить именно прогресс, необходимо произвести дополнительные операции: вычесть из оценки способности респондента в последующий замер оценку способности в предыдущий. Однако внутренняя многомерность заданий делает модели IRT численно менее стабильными, что затрудняет их практическое применение в случаях, когда после каждого замера респондентам нужно давать обратную связь и сообщать оценки их способности. В частности, в моделях с внутренней многомерностью заданий добавление последующих замеров приводит к некоторому пересчету оценок способности во все предыдущие замеры. Эти изменения относительно неболь-

шие и находятся в пределах ошибки измерения (статистически незначимы), однако респондентам бывает трудно объяснить, почему их балл изменился после нового замера. Модель Андерсена же, как модель с многомерностью между заданиями, характеризуется большей численной стабильностью, что определило ее использование в данной работе.

Безусловно, существуют способы измерения образовательного прогресса, не относящиеся к IRT [Саляхутдинова, Федерякин, 2022], но здесь мы ограничиваемся рассмотрением только этой парадигмы. Все описанные подходы предполагают использование коллатеральной информации о респондентах. Под коллатеральной понимается информация, которая теоретически не обоснована, но которую удобно собирать в компьютерном формате и применять для повышения точности оценивания при психометрической обработке данных с сохранением оригинальной интерпретации параметров модели [Федерякин, Угланова, Скрябин, 2021].

4. Анализ когнитивных операций в современной теории тестирования

Одно из допущений моделей, созданных на основе модели Раша, состоит в том, что у каждого задания есть свой параметр — δ_i (см. уравнения 1–2). Поэтому весь анализ и интерпретация результатов тестирования проводятся на уровне заданий. Однако иногда исследователей интересует интерпретация результатов тестирования относительно не заданий, а стоящих за ними когнитивных операций. Когда теоретическая рамка теста содержит описание когнитивных операций, которые провоцируются заданиями, связь заданий с этими когнитивными операциями может быть отражена в когнитивной карте теста, которую называют *Q*-матрицей. *Q*-матрица представляет собой таблицу с заданиями теста в строках и когнитивными операциями в столбцах, а ячейки показывают, как соотносятся гипотетические когнитивные операции с заданиями: 0 — если когнитивная операция не задействована в задании, 1 — если задействована. Наивный пример когнитивной карты теста по арифметике из трех заданий приведен в табл. 1.

Таблица 1. **Пример когнитивной карты теста**

Задание	Когнитивные операции		
	Сложение	Умножение	Порядок действий
$2 + 3 \times 2$	1	1	1
$4 \times 2 + 1$	1	1	0
$2 + 3 + 5$	1	0	0

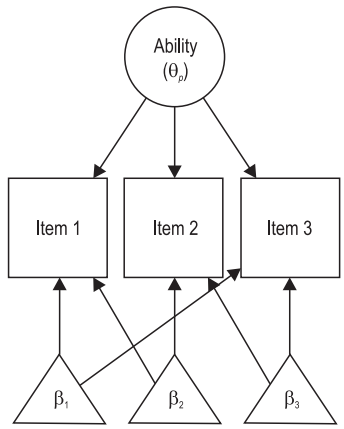
Специально для такого случая разработана (как частный случай модели Раша) одна из самых влиятельных моделей IRT — *Linear Logistic Test Model* (LLTM) [Fischer, 1973; 1995]:

$$P(U_{pi}=1)=\frac{\exp(\theta_{pt}-\sum_{k=1}^{K}\beta_k q_{ik})}{1+\exp(\theta_p-\sum_{k=1}^{K}\beta_k q_{ik})}, \quad (3)$$

где U_{pi} — ответ респондента p на задание i; θ_p — оценка латентной способности респондента p на логит-шкале; β_k — трудность когнитивной операции k ϵ (1, 2, ..., K) на логит-шкале; q_{ik} — элемент Q-матрицы из строки i и столбца k, обозначающий, необходима ли когнитивная операция k для решения задания i.

Путевая диаграмма модели LLTM показана на рис. 3.

Рис. 3. **Пример путевой диаграммы LLTM**



Таким образом, LLTM моделирует не индивидуальные трудности заданий, а трудности когнитивных операций, стоящих за этими заданиями. В LLTM часто добавляют не только когнитивные операции, но и другие атрибуты заданий, например формат или способ презентации информации в задании [Rolfes, Roth, Schnotz, 2018]. Но в данной работе мы концентрируемся только на изучении эффектов когнитивных операций.

В LLTM трудности заданий (δ_i из уравнения 1) допускаются равными сумме трудностей задействованных в задании когнитивных операций:

$$\delta_i=\sum_{k=1}^{K}\beta_k q_{ik}. \quad (4)$$

Ключевое отличие LLTM от модели Раша состоит в том, что в LLTM параметры, оцениваемые со стороны заданий, не уникальны, а «делятся» заданиями на основе общности когнитивных операций между ними. При этом, чтобы LLTM могла быть

оценена, когнитивных операций не должно быть больше, чем заданий. В отличие от обычной модели Раша, LLTM позволяет делать вывод о респондентах в терминах когнитивных операций, сравнивая оценки способности с уровнем трудности этих операций. По аналогии с оригинальной интерпретацией модели Раша, если уровень трудности когнитивной операции выше, чем уровень способности респондента, вероятность ее освоения этим респондентом меньше 50%, если уровень способности респондента превышает уровень трудности когнитивной операции, вероятность ее освоения больше 50%. Такой подход позволяет существенно обогатить интерпретацию тестовых баллов: вместо балла на логит-шкале, трансформированного на какую-то другую шкалу, результатом тестирования становится информация о том, что респондент уже умеет или чего еще не умеет [Lee, 2016]. Таким образом, применение LLTM служит основой для доказательного установления диагностических пороговых баллов на шкале способности: оцененный уровень трудности когнитивных операций становится диагностическим порогом, который можно напрямую сравнить со способностями респондентов.

LLTM допускает прогрессию (иерархию) в освоении когнитивных операций: более простая когнитивная операция будет более простой для всех респондентов, и нет никакой возможности двинуться дальше — к освоению следующей, более трудной когнитивной операции, пока предыдущая не освоена. Модели когнитивной диагностики, в том числе лонгитюдные [Deonovic et al., 2019], и анализ латентных переходов [Nsowaa, 2018] позволяют «расслабить» это допущение (или проверить его), однако они относятся к радикально другой статистической парадигме: они не ранжируют респондентов по способности, а сразу классифицируют их по нескольким основаниям — по степени освоения каждой из когнитивных операций, и их лонгитюдные расширения для измерения прогресса сложнее в применении.

Необходимым условием эффективного использования LLTM является качественная *Q*-матрица. Если она составлена неправильно, LLTM не только будет подходить данным хуже Раш-модели, в этом случае результаты моделирования с большой вероятностью будут содержать различные статистические артефакты [Baker, 1993; Macdonald, 2014]. На практике ситуация, когда LLTM подходит данным лучше, чем модель Раша, практически не встречается, но это не значит, что ее не стоит применять к данным. Теоретическая сила LLTM состоит в ее полезности для объяснения трудности когнитивных операций. При этом использование *Q*-матрицы для расчета LLTM адресует к использованию коллатеральной информации о заданиях [Федерякин, Угланова, Скрябин, 2021].

5. Методологические основы измерения образовательного прогресса на основе когнитивных операций

Для того чтобы проиллюстрировать измерение прогресса на основе когнитивных операций, мы подставляем уравнение (4) в уравнение (2), объединяя модель Андерсена и LLTM. Таким образом мы получаем модель LLTM — Андерсена (5) и противопоставляем ее в дальнейшей работе модели Раша — Андерсена (2):

$$P(U_{pit}=1)=\frac{\exp(\theta_{pt}-\sum_{k=1}^{K}\beta_k q_{ik})}{1+\exp(\theta_{pt}-\sum_{k=1}^{K}\beta_k q_{ik})}, \quad (5)$$

где U_{pit} — ответ респондента p на задание i в замер t; θ_{pt} — оценка латентной способности респондента p в замер t на логит-шкале; β_k — трудность когнитивной операции k на логит-шкале; q_{ik} — элемент Q-матрицы из строки i и столбца k.

Можно сказать, что модель LLTM — Андерсена выравнивает шкалы разных замеров не через трудности якорных заданий, а через трудности якорных когнитивных операций. Кроме выравнивания, модель LLTM — Андерсена обладает всеми достоинствами обычной LLTM, позволяя доказательно устанавливать диагностические пороговые баллы, но уже на вертикальной шкале сквозь все замеры. В этом отношении предлагаемая нами модель LLTM — Андерсена схожа с вертикально интерпретируемыми пороговыми баллами (*vertically moderated standard setting*, VMSS) [Cizek, 2013], однако она более состоятельна с математической точки зрения, поскольку не предполагает таких сильных теоретических допущений, как VMSS, где необходимым условием является равенство пороговых баллов, разделяющих респондентов на качественные группы и установленных внутри каждого теста по отдельности. В модели LLTM — Андерсена пороги на вертикальной шкале — это скорее следствие измерения прогресса, чем средство для него.

Таким образом, предлагаемый нами подход сочетает использование коллатеральной информации о заданиях (в виде Q-матрицы) и коллатеральной информации о респондентах (в виде замеров в разные моменты времени) для повышения надежности измерения при сохранении способа интерпретации способности из оригинальной Раш-модели. При этом возможно также привлечение коллатеральной информации о взаимодействии респондентов и заданий, например сведений о времени решения и/или о попытках решения.

6. Пример измерения образовательного прогресса на основе когнитивных операций

6.1. Инструмент

Для иллюстрации применения описанного подхода мы используем линейку тестов, разработанную в Центре психометрики и измерений в образовании Института образования НИУ «Высшая школа экономики» для мониторинга индивидуального образовательного прогресса учащихся средней школы. Мониторинг предполагает тестирование учащихся в начале и в конце учебного года. Первое тестирование проводится в конце сен-

тября — начале октября, когда учащиеся уже адаптировались к школе после летних каникул, но еще не начали активно изучать новый материал. Второе тестирование проводится в конце апреля — начале мая, когда основная учебная программа уже пройдена. В начале каждого учебного года используется тот же тест, что и в конце предыдущего учебного года.

В данном исследовании мы использовали данные, полученные в четырех волнах мониторинга по математике одной когорты учеников — от начала 8-го класса до конца 9-го класса. Так как тесты, предъявлявшиеся в конце 8-го класса и в начале 9-го класса, совпадают, базу данных составляют результаты выполнения школьниками трех уникальных тестов. Каждый тест состоял из 25 заданий с выбором одного варианта ответа из нескольких предложенных или коротким самостоятельно формулируемым ответом. Каждые два соседних теста содержали 4 якорных задания, а первый и третий тесты имели 2 якорных задания.

В заданиях представлены 12 разных когнитивных операций, причем в первом тесте их было 8, во втором — 11, а третий тест включает все 12. Примеры когнитивных операций, использованных в этой линейке тестов: навыки выполнения вычисления с рациональными числами, навык работы с координатной плоскостью. Весь набор когнитивных операций и их распределение между замерами представлены в табл. 2. Набор когнитивных операций для каждого задания определялся разработчиками теста и впоследствии был скорректирован экспертом в предметной области.

Таблица 2. **Распределение когнитивных операций между замерами**

Когнитивная операция	Замер 1	Замеры 2 и 3	Замер 4
Навыки выполнения вычислений с целыми числами	21	21	19
Навыки выполнения вычислений с рациональными числами	8	6	8
Навыки построения и применения математических моделей	7	6	8
Вспоминание фактов/формул и их последующее применение	11	18	19
Навык преобразования алгебраических выражений	10	7	6
Навыки решения уравнений	5	3	5
Навыки представления и считывания информации из текстового описания	2	5	3
Навыки выстраивания цепочки решения в несколько шагов	8	4	3
Навыки выполнения вычислений с вещественными числами	0	4	1
Навыки работы с координатной плоскостью	0	3	5
Навыки решения систем уравнений	0	2	1
Навыки решения неравенств и систем неравенств	0	0	4

6.2. Выборка и сбор данных

Выборка состоит из 472 учеников, которые принимали участие хотя бы в одном из четырех тестирований в ходе мониторинга. Из них 267 учеников принимали участие во всех четырех замерах, 121 — в трех (пропустили один замер), 53 — в двух (пропустили два замера) и 31 — только в одном. По замерам распределение равномерное: 394 ученика протестированы в начале 8-го класса, 383 — в конце 8-го класса, 388 — в начале 9-го класса и 395 — в конце 9-го класса; т.е. ни на одном из этапов с выборкой не происходило существенных изменений. С географической точки зрения выборка гомогенна, так как все ученики обучаются в школах одного и того же района одного из регионов России. Тестирования проводились в 2020–2022 гг.

Ученики проходили тестирование на персональных компьютерах, предоставленных образовательными учреждениями, на которых им требовалось авторизоваться, чтобы получить доступ к системе тестирования и тестовым материалам. Соблюдение процедуры тестирования, включая контроль списывания, а также безопасность тестовых материалов обеспечивали местные наблюдатели.

Для обработки результатов тестирования использовалась библиотека TAM v.3.7-16 для языка программирования R v.4.1.0.

6.3. Результаты

На общих данных по всем четырем замерам построены две модели: модель Раша — Андерсена, которая с учетом якорных заданий содержит 67 индивидуальных параметров трудности заданий, и модель LLTM — Андерсена, в которой 12 параметров трудности когнитивных операций. Для проверки согласия моделей с данными мы используем статистики глобального согласия: информационный критерий Акаике (*Akaike Information Criterion*, AIC) [Akaike, 1974] и информационный критерий Шварца (*Bayesian Information Criterion*, BIC) [Gideon, 1978]. Эти индексы описывают общее согласие моделей с данными, вводя пенальти за дополнительные параметры (AIC) с учетом размера выборки (BIC). Судя по статистикам согласия, модель Раша — Андерсена подходит нашим данным лучше модели LLTM — Андерсена (табл. 3) — возможно, за счет того, что она точнее описывает трудности заданий из-за большего количества оцениваемых параметров. EAP-надежность для всех замеров [Adams, 2005] незначительно выше в модели Раша — Андерсена.

В обеих моделях средняя способность в первом замере была зафиксирована равной нулю, а в остальных замерах она является оцениваемым параметром, и при этом отслеживается ее изменение по сравнению с первым замером. Во втором замере наблюдается увеличение средней способности, в третьем — ее спад. Такая динамика представляет собой относитель-

Таблица 3. **Сравнение моделей Раша — Андерсена и LLTM — Андерсена**

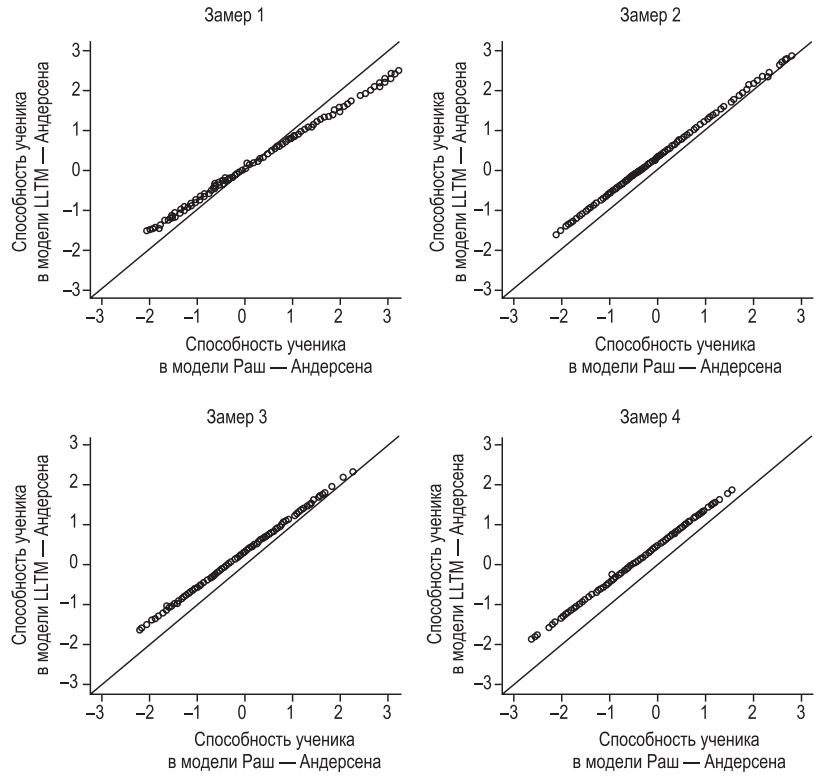
Статистика		Раш — Андерсен	LLTM — Андерсен
Глобальное согласие	AIC	41 828	45 748
	BIC	42 160	45 851
Надежность	Замер 1	0,787	0,767
	Замер 2	0,779	0,772
	Замер 3	0,767	0,760
	Замер 4	0,688	0,673
Средняя способность учеников	Замер 1	0	0
	Замер 2	0,060	0,218
	Замер 3	–0,221	–0,031
	Замер 4	–0,272	0,074
Средняя способность 267 учеников*	Замер 1	0,054	0,076
	Замер 2	–0,026	0,305
	Замер 3	–0,308	0,060
	Замер 4	–0,431	0,104
Средняя трудность заданий	Замер 1	0,88	0,74
	Замер 2	0,62	0,74
	Замер 3	0,62	0,74
	Замер 4	0,53	0,80

* Ученики, участвовавшие во всех четырех замерах.

но хорошо изученный эффект летних каникул: уровень подготовленности учеников школ «проседает» после длительного перерыва в занятиях [Cooper et al., 1996]. Однако в четвертом замере в модели Раша — Андерсена спад средней способности продолжается, а в модели LLTM — Андерсена происходит ее рост. Дело в том, что, несмотря на схожесть, эти модели оценивают способность по-разному: относительно заданий и относительно когнитивных операций. В последнем замере средняя трудность заданий в модели LLTM — Андерсена, полученная как сумма трудностей задействованных в заданиях когнитивных операций, на 0,27 логита выше, чем средняя трудность заданий в модели Раша — Андерсена. Таким образом, трудность заданий последнего замера в модели Раша — Андерсена может быть занижена, так как в данном замере в мониторинг введены новые содержательные области предмета, а задания по этим областям были простыми относительно якорных заданий. Модель LLTM — Андерсена учитывает, что в этих заданиях задействуются более сложные когнитивные операции, уровень трудности которых моделируется на основе заданий всех четырех замеров. Такое расхождение в динамике средней способности, оцененной в разных моделях, означает, что выводы, сделанные на основе одной из них, возможно, невалидны. Мы считаем, что результаты, полученные в рамках модели LLTM — Андер-

сена, более точно отражают реальность. При этом обе модели оценивают способность на одном и том же материале, и корреляции показателей способностей учеников из одного и того же замера, полученных в разных моделях, выше 0,99. Однако оценки способностей учеников для второго, третьего и четвертого замеров в модели Раша — Андерсена ниже, чем оценки в модели LLTM — Андерсена. Причем от замера к замеру эта разница растет (рис. 4).

Рис. 4. **Связь оценок способностей в двух моделях для всех замеров**



Примечание: Диагональная линия показывает численное соотношение способностей 1:1.

Корреляция трудностей заданий из разных моделей равна 0,56. В модели LLTM — Андерсена трудность заданий определена исходя из суммы трудностей задействованных когнитивных операций, в то время как модель Раша — Андерсона оценивает параметр трудности каждого задания напрямую. В принципе разницу между трудностями заданий из двух моделей можно объяснить *Q*-матрицей, которая не включает всех факторов, формирующих трудность заданий, кроме задействованных когнитивных операций. Такими факторами могут быть, например,

формат заданий или ориентация — теоретическая или практическая. Таким образом, оценка трудности заданий в модели LLTM — Андерсена лишена вклада данных факторов, не относящихся к предметной сложности заданий, а полученные в этой модели оценки способностей и прогресса будут характеризовать когнитивные операции.

В результате для модели LLTM — Андерсена мы получили оценки уровня трудности когнитивных операций (табл. 4). Соотнося способность ученика с уровнем трудности когнитивных операций, мы можем сделать вывод о том, какие операции ученик освоил. Например, ученик со способностью 0,5 логита, вероятно, уже освоил навыки выполнения вычислений с целыми числами, но еще не освоил навыки выполнения вычислений с рациональными числами.

Таблица 4. Т**рудности когнитивных операций из модели LLTM — Андерсена**

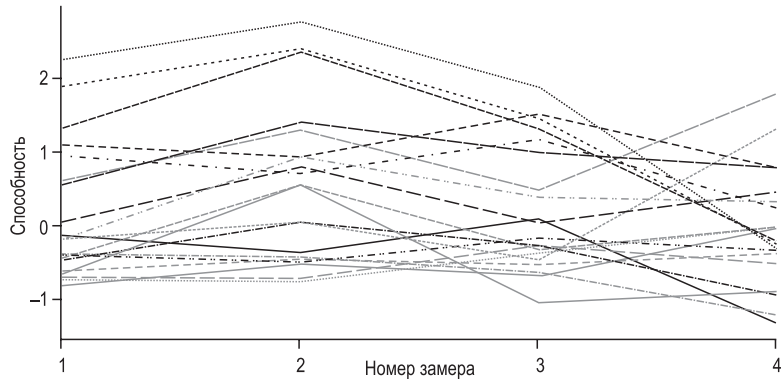
Когнитивная операция	Трудность
Навыки выполнения вычислений с целыми числами	0,026
Навыки выполнения вычислений с рациональными числами	0,647
Навыки построения и применения математических моделей	0,169
Вспоминание фактов/формул и их последующее применение	0,401
Навык преобразования алгебраических выражений	0,143
Навыки решения уравнений	0,426
Навыки представления и считывания информации из текстового описания	–0,503
Навыки выстраивания цепочки решения в несколько шагов	0,575
Навыки выполнения вычислений с вещественными числами	0,021
Навыки работы с координатной плоскостью	0,501
Навыки решения систем уравнений	1,095
Навыки решения неравенств и систем неравенств	–0,360

Полученная иерархия трудностей когнитивных операций по большей части соответствует изначальным предположениям. Трудность около нуля (0,026) имеют задания, в которых необходимо выполнить вычисления с целыми числами, а если требуются вычисления с рациональными числами, трудность задания повышается на 0,647. Минимальную трудность представляет операция «навыки представления и считывания информации из текстового описания», т.е. ученики в первую очередь осваивают данную когнитивную операцию, и наличие в составе задания операции считывания информации из текстового описания снижает трудность задания на 0,503. Если же

задание предполагает выстраивание цепочки решения в несколько шагов, его трудность возрастает на 0,575. Ожидания не подтвердились в отношении двух операций, а именно «навыков выполнения вычислений с вещественными числами» и «навыков решения неравенств и систем неравенств». Оценки их трудности оказались ниже, чем можно было ожидать. Возможно, причина состоит в малом количестве заданий, включающих данные операции (табл. 2). Соответствующая содержательная область только появилась в учебной программе и представлена в мониторинге только базовыми заданиями. Простота этих заданий в рамках своей содержательной области и определила низкие показатели трудности данных когнитивных операций.

Модель LLTM — Андерсена позволяет отслеживать прогресс каждого ученика. На рис. 5 приведены траектории изменения способности 20 случайно выбранных учеников, которые участвовали во всех замерах.

Рис. 5. **Траектории образовательного прогресса учеников**



7. Обсуждение и выводы

Мы рассмотрели способ измерения образовательного прогресса с помощью когнитивных операций, освоение которых проверяется в тесте. Для этой цели применены специальные психометрические методы IRT, а именно модель LLTM, совмещенная с моделью Андерсена для моделирования повторяющихся замеров. В отличие от классической модели Раша — Андерсена, ориентированной на оценку трудности заданий, модель LLTM — Андерсена направлена на анализ трудностей когнитивных операций как составляющих этих заданий. При этом в модели LLTM — Андерсена наличие якорных заданий в последовательных замерах не является обязательным, в отличие от модели Раша — Андерсена. В этом отношении модель LLTM — Андерсена аналогична техникам лонгитюдного моделирования в моделях когнитивной диагностики [Yu, Zhan, Chen, 2023].

Если *Q*-матрица составлена идеально и трудности когнитивных операций, задействованных в заданиях, абсолютно стабильны друг относительно друга и полностью описывают разброс трудностей заданий, то можно использовать непересекающиеся наборы заданий в разных замерах, поскольку в модели LLTM — Андерсена нет индивидуальных трудностей заданий. Однако в реальности составление *Q*-матрицы всегда вносит существенные ограничения в оцениваемые параметры. Соответственно, использование непересекающихся наборов заданий потребует очень сильных допущений, которые могут не выдерживаться в данных.

Использование логики LLTM в моделировании образовательного прогресса означает необходимость учета всех ограничений LLTM. В частности, появляется допущение, что разброс трудностей заданий достаточно хорошо описывается разбросом трудностей когнитивных операций. В реальности такого практически не бывает, из-за чего LLTM никогда не подходит данным лучше Раш-модели и преимущества LLTM перед Раш-моделью являются интерпретационными, а не статистическими. Качество результатов в LLTM очень сильно зависит от качества использованной *Q*-матрицы, что дополнительно актуализирует исследования по валидизации или автоматическому составлению *Q*-матрицы [Sun et al., 2014]. При этом, однако, важно сохранить интерпретируемость LLTM как ее основное достоинство.

Модели LLTM — Андерсена пригодны для анализа лонгитюдной измерительной инвариантности [Vandenberg, Lance, 2000], который не был проведен в данной работе, что является одним из ее ограничений. Анализ измерительной инвариантности — необходимый этап любого лонгитюдного моделирования, предназначенный для подтверждения валидности выводов. Он призван доказать, что психометрические характеристики заданий не меняются со временем. Поскольку в модели LLTM — Андерсена нет индивидуальных характеристик заданий, классические анализы измерительной инвариантности (*Differential Item Functioning* (DIF) анализы) в ней невозможны. Оценить измерительную инвариантность когнитивных операций можно с помощью анализа относительной стабильности их трудностей [Bechger, Maris, 2015]. Для таких анализов необходимо откалибровать LLTM отдельно в каждом замере (без установления общей шкалы), найти разницу между трудностью каждой когнитивной операции и каждой другой, а затем сравнить эти разницы между разными замерами. Если смещения этих разниц несущественны, можно заключить, что трудности когнитивных операций стабильны и все изменения в способности респондентов будут отражены только в изменении оцен-

ки этой способности от одного замера к другому. В силу того, что целью данной статьи было лишь проиллюстрировать и обосновать первое применение такого подхода к измерению образовательного прогресса, мы не проводили полный перечень анализов, необходимых для обеспечения валидности результатов. В этом состоит одно из главных ограничений данного исследования.

Стабильными от замера к замеру в модели LLTM — Андерсена должны оставаться не только параметры заданий (когнитивных операций), но и сама когнитивная карта якорных заданий. Одно и то же задание не может менять состав своих когнитивных операций от одного замера к другому. Например, если задание в первом замере относилось к навыкам выполнения вычислений с целыми числами и навыкам выстраивания цепочки решения в несколько шагов, то и во всех последующих замерах оно должно быть классифицировано точно так же. В таком случае в качестве когнитивных операций не может использоваться какой-либо уровень сложности заданий (базовый или повышенный), так как он может меняться в зависимости от класса: задание, которое представляло повышенную сложность в начале 8-го класса, стало заданием базовой сложности в конце 9-го класса. Соответственно, для составления *Q*-матрицы необходимо использовать только «стабильные» когнитивные операции, интерпретация и проявление которых не меняются. В противном случае они не могут служить порогами на вертикальной шкале интерпретационно и элементами *Q*-матрицы математически.

Одним из главных ограничений предложенного подхода является использование для вынесения суждений о респондентах порогов, которыми в модели LLTM — Андерсена можно считать оцененные трудности когнитивных операций. На том основании, что респондент преодолевает порог, т.е. его способность оказывается выше трудности определенной когнитивной операции, делается вывод об освоении этим респондентом данной когнитивной операции. При принятии решений с высокими ставками в большинстве тестирований пороговые баллы используются с обязательным соблюдением специальных процедур их установления, а именно с участием в этих процедурах представителей всех групп пользователей результатов тестирования: учеников, родителей, учителей, образовательных администраторов и т.д. [Cizek, Bunch, 2007]. Такие процедуры необходимы для минимизации потенциально нежелательных социальных последствий использования, в том числе и некорректного, этих пороговых баллов [Messick, 1998; Hubley, Zumbo, 2011]. Однако рассматриваемые в данной работе пороги являются побочным продуктом предложенной нами методологии,

это пороги исключительно диагностические и предназначены для принятия решений в тестировании с низкими ставками. Такие пороги можно называть критериально-ориентированными диагностическими порогами, потому что они выведены из психометрических параметров заданий, а не на основе распределения групп респондентов, как нормативно ориентированные пороги. Использование данных порогов для принятия решений с высокими ставками влечет за собой риск нежелательных или непредвиденных социальных последствий.

Благодарности Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-29-14110.

Литература

1. Саляхутдинова Д.Р., Федерякин Д.А. (2022) Способы связывания шкал для измерения образовательного прогресса в разных парадигмах анализа данных образовательного тестирования. *Отечественная и зарубежная педагогика*, т. 1, № 3, сс. 98–111. https://doi.org/10.24412/2224-0772-2022-84-98-111
2. Федерякин Д.А., Угланова И.Л., Скрябин М.А. (2021) Новые источники информации в компьютерном тестировании. *Вестник Томского государственного университета*, № 465, сс. 179–187. https://doi.org/10.17223/15617793/465/24
3. Adams R.J. (2005) Reliability as a Measurement Design Effect. *Studies in Educational Evaluation*, vol. 31, no 2, pp. 162–172. https://doi.org/10.1016/j.stueduc.2005.05.008
4. Akaike H. (1974) A New Look at the Statistical Model Identification*. IEEE Transactionson Automatic Control*, vol. 19, no 6, pp. 716–723. https://doi.org/10.1109/TAC.1974.1100705
5. Andersen E.B. (1985) Estimating Latent Correlations between Repeated Testings. *Psychometrika*, vol. 50, March, pp. 3–16. https://doi.org/10.1007/BF02294143
6. Andersen E.B. (1977) Sufficient Statistics and Latent Trait Models. *Psychometrika*, vol. 42, March, pp. 69–81. https://doi.org/10.1007/BF02293746
7. Baker F.B. (1993) Sensitivity of the Linear Logistic Test Model to Misspecification of the Weight Matrix. *Applied Psychological Measurement*, vol. 17, no 3, pp. 201–210. https://doi.org/10.1177/014662169301700301
8. Bechger T.M., Maris G. (2015) A Statistical Test for Differential Item Pair Functioning. *Psychometrika*, vol. 80, June, pp. 317–340. https://doi.org/10.1007/s11336-014-9408-y
9. Cizek G.J. (ed.) (2013) *Vertically Moderated Standard Setting: A Special Issue of Applied Measurement in Education*. New York, NY: Routledge. https://doi.org/10.4324/97813150459008
10. Cizek G.J., Bunch M.B. (2007) *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests.* Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781412985918
11. Cooper H., Nye B., Charlton K., Lindsay J., Greathouse S. (1996) The Effects of Summer Vacation on Achievement Test Scores: A Narrative and Meta-Analytic Review. *Review of Educational Research*, vol. 66, no 3, pp. 227–268. https://doi.org/10.2307/1170523

12. Deonovic B., Chopade P., Yudelson M., de la Torre J., von Davier A.A. (2019) Application of Cognitive Diagnostic Models to Learning and Assessment Systems. *Handbook of Diagnostic Classification Models: Models and Model Extensions, Applications, Software Packages* (eds M. von Davier, Y.-S. Lee), Cham: Springer, pp. 437–460. https://doi.org/10.1007/978-3-030-05584-4_21
13. Dimitrov D.M., Rumrill Jr. P.D. (2003) Pretest-Posttest Designs and Measurement of Change. *Work*, vol. 20, no 2, pp. 159–165.
14. Embretson S.E. (1991) A Multidimensional Latent Trait Model for Measuring Learning and Change. *Psychometrika*, vol. 56, September, pp. 495–515. https://doi.org/10.1007/BF02294487
15. Fischer G.H. (1995) The Linear Logistic Test Model. *Rasch Models* (eds G.H. Fischer, I.W. Molenaar), New York, NY: Springer, pp. 131–155. https://doi.org/10.1007/978-1-4612-4230-7_8
16. Fischer G.H. (1973) The Linear Logistic Test Model as an Instrument in Educational Research. *Acta Psychologica*, vol. 37, no 6, pp. 359–374. https://doi.org/10.1016/0001-6918(73)90003-6
17. Gideon S. (1978) Estimating the Dimension of a Model. *The Annals of Statistics*, vol. 6, no 2, pp. 461–464. https://doi.org/10.1214/aos/1176344136
18. Hubley A.M., Zumbo B.D. (2011) Validity and the Consequences of Test Interpretation and Use. *Social Indicators Research*, vol. 103, no 2, pp. 219–230. https://doi.org/10.1007/s11205-011-9843-4
19. Lee H.K. (2016) *An Application of Item Response Theory to Investigate the Validity of a Learning Progression for Number Sense* (PhD Thesis). Berkeley, CA: University of California.
20. Linden van der W. J. (2018) *Handbook of Item Response Theory: Three Volume Set*. Boca Raton, FL: Chapman and Hall/CRC. https://doi.org/10.1201/9781315119144
21. Loyd B.H., Hoover H.D. (1980) Vertical Equating Using the Rasch Model. *Journal of Educational Measurement*, vol. 17, no 3, pp. 179–193. https://doi.org/10.1111/j.1745-3984.1980.tb00825.x
22. Macdonald G.T. (2014) *The Performance of the Linear Logistic Test Model When the Q-Matrix Is Misspecified: A Simulation Study* (PhD Thesis). Tampa, FL: University of South Florida.
23. Messick S. (1998) Test Validity: A Matter of Consequence. *Social Indicators Research*, vol. 45, November, pp. 35–44. https://doi.org/10.1023/A:1006964925094
24. Nsowaa B. (2018) *The Ordered Latent Transition Analysis Model for the Measurement of Learning* (PhD Thesis). New York, NY: Columbia University.
25. Rolfes T., Roth J., Schnotz W. (2018) Effects of Tables, Bar Charts, and Graphs on Solving Function Tasks. *Journal für Mathematik-Didaktik,* vol. 39, no 1, pp. 97–125. http://dx.doi.org/10.1007/s13138-017-0124-x
26. Slavin R.E. (2005) *Evidence-Based Reform: Advancing the Education of Students at Risk. Report Prepared for Renewing Our Schools, Securing Our Future*. Available at: https://goo.su/vYeO (accessed 20 July 2023).
27. Sontag L.M. (1984) *Vertical Equating Methods: A Comparative Study of Their Efficacy*. New York, NY: Columbia University.
28. Sun Y., Ye S., Inoue S., Sun Y. (2014) Alternating Recursive Method for Q-matrix Learning. Proceedings of the *7th International Conference on Educational Data Mining* (London, July 4–7, 2014), pp. 14–20.
29. Vandenberg R.J., Lance C.E. (2000) A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, vol. 3, no 1, pp. 4–70. https://doi.org/10.1177/109442810031002
30. Waterbury G.T., DeMars C.E. (2021) Anchors Aweigh: How the Choice of Anchor Items Affects the Vertical Scaling of 3PL Data with the Rasch Model.

Educational Assessment, vol. 26, no 3, pp. 175–197. https://doi.org/10.1080/10627197.2020.1858782

31. Wilson M., Zheng X., McGuire L. (2012) Formulating Latent Growth Using an Explanatory Item Response Model Approach. *Journal of Applied Measurement*, vol. 13, no 1, pp. 1–22.
32. Yu X., Zhan P., Chen Q. (2023) Don't Worry about the Anchor-Item Setting in Longitudinal Learning Diagnostic Assessments. *Frontiers in Psychology*, vol. 14, February, Article no 1112463. https://doi.org/10.3389/fpsyg.2023.1112463

References

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

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

Andersen E.B. (1985) Estimating Latent Correlations between Repeated Testings. *Psychometrika*, vol. 50, March, pp. 3–16. https://doi.org/10.1007/BF02294143

Andersen E.B. (1977) Sufficient Statistics and Latent Trait Models. *Psychometrika*, vol. 42, March, pp. 69–81. https://doi.org/10.1007/BF02293746

Baker F.B. (1993) Sensitivity of the Linear Logistic Test Model to Misspecification of the Weight Matrix. *Applied Psychological Measurement*, vol. 17, no 3, pp. 201–210. https://doi.org/10.1177/014662169301700301

Bechger T.M., Maris G. (2015) A Statistical Test for Differential Item Pair Functioning. *Psychometrika*, vol. 80, June, pp. 317–340. https://doi.org/10.1007/s11336-014-9408-y

Cizek G.J. (ed.) (2013) *Vertically Moderated Standard Setting: A Special Issue of Applied Measurement in Education*. New York, NY: Routledge. https://doi.org/10.4324/97813150459008

Cizek G.J., Bunch M.B. (2007) *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests.* Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781412985918

Cooper H., Nye B., Charlton K., Lindsay J., Greathouse S. (1996) The Effects of Summer Vacation on Achievement Test Scores: A Narrative and Meta-Analytic Review. *Review of Educational Research*, vol. 66, no 3, pp. 227–268. https://doi.org/10.2307/1170523

Deonovic B., Chopade P., Yudelson M., de la Torre J., von Davier A.A. (2019) Application of Cognitive Diagnostic Models to Learning and Assessment Systems. *Handbook of Diagnostic Classification Models: Models and Model Extensions, Applications, Software Packages* (eds M. von Davier, Y.-S. Lee), Cham: Springer, pp. 437–460. https://doi.org/10.1007/978-3-030-05584-4_21

Dimitrov D.M., Rumrill Jr. P.D. (2003) Pretest-Posttest Designs and Measurement of Change. *Work*, vol. 20, no 2, pp. 159–165.

Embretson S.E. (1991) A Multidimensional Latent Trait Model for Measuring Learning and Change. *Psychometrika*, vol. 56, September, pp. 495–515. https://doi.org/10.1007/BF02294487

Federiakin D.A. Uglanova I.L., Skryabin M.A. (2021) Novye istochniki informatsii v komp'yuternom testirovanii [New Sources of Information in Computerized Testing]. *Tomsk State University Journal*, no 465, pp. 179–187. https://doi.org/10.17223/15617793/465/24

Fischer G.H. (1995) The Linear Logistic Test Model. *Rasch Models* (eds G.H. Fischer, I.W. Molenaar), New York, NY: Springer, pp. 131–155. https://doi.org/10.1007/978-1-4612-4230-7_8

Fischer G.H. (1973) The Linear Logistic Test Model as an Instrument in Educational Research. *Acta Psychologica*, vol. 37, no 6, pp. 359–374. https://doi.org/10.1016/0001-6918(73)90003-6

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

Hubley A.M., Zumbo B.D. (2011) Validity and the Consequences of Test Interpretation and Use. *Social Indicators Research*, vol. 103, no 2, pp. 219–230. https://doi.org/10.1007/s11205-011-9843-4

Lee H.K. (2016) *An Application of Item Response Theory to Investigate the Validity of a Learning Progression for Number Sense* (PhD Thesis), Berkeley, CA: University of California.

Linden van der W. J. (2018) *Handbook of Item Response Theory: Three Volume Set*. Boca Raton, FL: Chapman and Hall/CRC. https://doi.org/10.1201/9781315119144

Loyd B.H., Hoover H.D. (1980) Vertical Equating Using the Rasch Model. *Journal of Educational Measurement*, vol. 17, no 3, pp. 179–193. https://doi.org/10.1111/j.1745-3984.1980.tb00825.x

Macdonald G.T. (2014) *The Performance of the Linear Logistic Test Model When the Q-Matrix Is Misspecified: A Simulation Study* (PhD Thesis). Tampa, FL: University of South Florida.

Messick S. (1998) Test Validity: A Matter of Consequence. *Social Indicators Research*, vol. 45, November, pp. 35–44. https://doi.org/10.1023/A:1006964925094

Nsowaa B. (2018) *The Ordered Latent Transition Analysis Model for the Measurement of Learning* (PhD Thesis). New York, NY: Columbia University.

Rolfes T., Roth J., Schnotz W. (2018) Effects of Tables, Bar Charts, and Graphs on Solving Function Tasks. *Journal für Mathematik-Didaktik,* vol. 39, no 1, pp. 97–125. http://dx.doi.org/10.1007/s13138-017-0124-x

Salyakhutdinova D.R., Federiakin D.A. (2022) Sposoby svyazyvaniya shkal dlya izmereniya obrazovatel'nogo progressa v raznykh paradigmakh analiza dannykh obrazovatel'nogo testirovaniya [Methods of Linking Scales for Measuring Educational Progress in Different Paradigms of Educational Testing Data Analysis]. *Domestic and Foreign Pedagogy*, vol. 1, no 3, pp. 98–111. https://doi.org/10.24412/2224-0772-2022-84-98-111

Slavin R.E. (2005) *Evidence-Based Reform: Advancing the Education of Students at Risk. Report Prepared for Renewing Our Schools, Securing Our Future*. Available at: https://goo.su/vYeO (accessed 20 July 2023).

Sontag L.M. (1984) *Vertical Equating Methods: A Comparative Study of Their Efficacy*. New York, NY: Columbia University.

Sun Y., Ye S., Inoue S., Sun Y. (2014) Alternating Recursive Method for Q-matrix Learning. Proceedings of the *7th International Conference on Educational Data Mining* (London, July 4–7, 2014), pp. 14–20.

Vandenberg R.J., Lance C.E. (2000) A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, vol. 3, no 1, pp. 4–70. https://doi.org/10.1177/109442810031002

Waterbury G.T., DeMars C.E. (2021) Anchors Aweigh: How the Choice of Anchor Items Affects the Vertical Scaling of 3PL Data with the Rasch Model. *Educational Assessment*, vol. 26, no 3, pp. 175–197. https://doi.org/10.1080/10627197.2020.185878231

Wilson M., Zheng X., McGuire L. (2012) Formulating Latent Growth Using an Explanatory Item Response Model Approach. *Journal of Applied Measurement*, vol. 13, no 1, pp. 1–22.

Yu X., Zhan P., Chen Q. (2023) Don't Worry about the Anchor-Item Setting in Longitudinal Learning Diagnostic Assessments. *Frontiers in Psychology*, vol. 14, February, Article no 1112463. https://doi.org/10.3389/fpsyg.2023.1112463

Так ли полезна психометрика для академической психологии?

Юлия Тюменева

Статья поступила в редакцию в феврале 2023 г.

Тюменева Юлия Алексеевна — кандидат психологических наук, старший научный сотрудник Института образования, Национальный исследовательский университет «Высшая школа экономики». Адрес: 101000, Москва, Потаповский пер., 16, стр. 10. E-mail: jutu@yandex.ru. ORCID: https://orcid.org/0000-0002-2381-917X

Аннотация

Психологические теории относительно способностей и личностных черт часто полагаются на результаты психометрического моделирования. Предполагается, что оно связывает ответы на задания теста с ненаблюдаемым «конструктом» (чертой, способностью), который и «моделируется» на основе данных теста. Однако свидетельствует ли согласие между данными и моделью о том, что модель репрезентирует психологический конструкт? Насколько вообще психометрическое моделирование является моделированием в общенаучном значении этого термина? От ответа на эти вопросы зависит обоснованность использования данных моделирования для понимания психологических феноменов. В статье анализируется логика психометрического моделирования в сравнении с моделированием в других науках и утверждается, что психологические феномены как предмет моделирования не участвуют ни в построении, ни в коррекции моделей. Автор поднимает проблему необоснованных интерпретаций результатов моделирования в психологии и их нежелательных последствий для психологической теории. При этом значение психометрического моделирования как инструмента для решения управленческих задач еще ждет своей оценки.

Ключевые слова

психометрическое моделирование, моделирование латентного конструкта, психологический конструкт, психологическая теория, тест

Для цитирования

Тюменева Ю.А. (2023) Так ли полезна психометрика для академической психологии? *Вопросы образования / Educational Studies Moscow*, № 3, сс. 197–220. https://doi.org/10.17323/vo-2023-16781

Is Psychometrics So Useful for Academic Psychology?

Yulia Tyumeneva

Yulia A. Tyumeneva — PhD, Associate Professor and Senior Research Fellow in Institute of Education, National Research University, Higher School of Economics, Address: Bld. 10, 16 Potapovsky Ln, 101000 Moscow, Russian Federation. E-mail: jutu@yandex.ru. ORCID: https://orcid.org/0000-0002-2381-917X

Abstract Psychological theories regarding ability and personality traits often rely on the results of psychometric modelling. The latter is assumed to link responses to test items to an unobserved 'construct' (trait, ability), which is 'modelled' from the test data. However, does the agreement between the data and the model indicate that the model represents a psychological construct? To what extent is 'psychometric modelling' modelling in the general scientific sense of the term? The validity of using modelling data to understand psychological phenomena depends on the answer to these questions. The article analyses the logic of psychometric modelling in comparison with modelling in other sciences and argues that psychological phenomena as a subject of modelling are not involved neither in the construction nor in the correction of models. The problem of unjustified interpretations of modelling results in psychology and their undesirable consequences for psychological theory is raised. At the same time, the use of psychometric modelling for human resource decision-making is still waiting for its evaluation.

Keywords psychometric modelling, latent construct modelling, psychological construct, psychological theory, test

For citing Tyumeneva Yu.A. (2023) Tak li polezna psikhometrika dlya akademicheskoy psikhologii? [Is Psychometrics So Useful for Academic Psychology?]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 3, pp. 197–220. https://doi.org/10.17323/vo-2023-16902

Предварительные замечания

Предлагаемая вниманию читателей статья вызвала довольно острую полемику между мной и одним из моих рецензентов. При том что я не могу согласиться со многими его аргументами, я считаю состоявшееся обсуждение чрезвычайно важным, поскольку оно подсвечивает всю сложность затронутой проблемы и отсутствие ее однозначного решения. Чтобы проследить логику доводов обеих сторон, мы согласились на их полную публикацию в будущем номере журнала. Таким образом ничье авторство не нарушается, а читатель может получить удовольствие от интригующей дискуссии. Я благодарна обоим рецензентам и редакторам за публикацию этой статьи, несмотря на то что она очевидно диссонирует с их личной позицией, а также, возможно, с другими материалами этого выпуска.

Введение

Психометрика, по определению Американской психометрической ассоциации, занимается количественной оценкой и измерением психических качеств, поведения, успешности и т.п., а также разработкой, анализом и совершенствованием тестов, опросников и других инструментов, используемых для такого измерения[1]. Более аккуратное определение предлагает Международная энциклопедия социальных и поведенческих наук: психометрика — это научная дисциплина, занимающаяся вопросами конструирования психометрических моделей психо-

[1] APA Dictionary of Psychology: https://dictionary.apa.org/psychometrics

логических данных [Borsboom, Molenaar, 2015]. В этих моделях теоретический конструкт, например интеллект, систематически координируется с наблюдениями, например с баллами по тесту интеллекта. Чаще всего это делается с помощью так называемых моделей латентных переменных, которые работают как общие детерминанты набора баллов по тесту[2].

Идея моделирования как метода изучения психологических конструктов пронизывает всю современную психометрическую литературу. Утверждения о том, что психометрика может моделировать латентный психологический конструкт, звучат в литературе совершенно явно. Выводы или гипотезы о конструкте делаются по результатам психометрического моделирования в отношении самых разных феноменов. Идет ли речь о мотивах [Freund, Lohbeck, 2021], настойчивости [Credé, 2018; Tyumeneva, Kardanova, Kuzmina, 2019], личности [Franić et al., 2014; Walton et al., 2008; Streckert, Kurtz, Kajonius, 2023], компетентности [Hartig, Höhler, 2009], мышлении [Dumas, Dong, 2022; Wagner, Harvey, 2006; Alexander et al., 2016; Zhao, Alexander, Sun, 2021; Araujo et al., 2019], креативности [Qian, Plucker, Yang, 2019; Shaw, Kapnek, Morelli, 2021], субъективном благополучии [Nima et al., 2020], эмоциях [Lange et al., 2020; Power, 2006], установках [Hauwaert van, Schimpf, Azevedo, 2020] — во всех случаях психометрические модели составляют существенную часть методического инструментария. Заявления о возможности предоставлять информацию об измеряемом конструкте — мышлении, мотивах, установках и т.д., — более или менее явно сформулированные, можно найти во многих учебниках по психометрике и в методической периодике [Ackerman, Gierl, Walker, 2003; Fox, 2005; Linden van der, Hambleton, 2013; Nering, Ostini, 2010; Sijtsma, Ark van der, 2020]. Моделирование латентных конструктов стало дополнительным способом валидизации шкал, в том числе использующихся в международных программах оценки качества образования, таких как PISA или PIRLS, средством поддержки интерпретации результатов тестирования в терминах черт и способностей и даже источником новых конструктов и теорий [Buchholz, Hartig, 2020; Kunina-Habenicht, Goldhammer, 2020].

У термина «конструкт» до сих пор нет исчерпывающего и общепринятого определения, при этом из-за смешивания с близкими понятиям, такими как «атрибут», «фактор», «ла-

[2] Не все психометрические модели имеют дело с латентными переменными, например есть сетевые или формативные модели, когнитивные диагностические модели [Templin, Henson, 2006; Schmittmann et al., 2013]. В этой работе мы будем обсуждать моделирование конструктов в парадигме современной теории тестирования и факторного анализа, хотя эта же логика применима и к другим моделям.

тентная переменная», «ненаблюдаемая переменная», его содержание становится еще более туманным. Не будет ошибкой утверждать, что часто под конструктом подразумевается предположительная ненаблюдаемая детерминанта какого-то наблюдаемого паттерна поведения, в том числе и выполнения заданий теста. Способности, черты характера, установки — все эти характеристики описываются как латентные конструкты. Латентные в том смысле, что они не наблюдаются непосредственно, но предположительно объясняют поведение.

Латентный / психологический / ненаблюдаемый конструкт

Из-за отсутствия однозначного определения понятия «конструкт» его включение в профессиональный жаргон имело противоречивые последствия. С одной стороны, понятие «конструкт» позволило обойти ограничения операционализма, открыв доступ к теоретическим обобщениям в количественной психологии, а с другой — затруднило коммуникацию, смешав языки статистического вывода, лингвистических концепций и психологических теорий [Cronbach, Meehl, 1955; MacCorquodale, Meehl, 1948; Markus, Borsboom, 2013; Michell, 2013].

Перечислим типичные значения, в которых термин «конструкт» употребляется в психометрической литературе, в том числе в учебниках[3]. Во-первых, понятие «конструкт» используется, чтобы репрезентировать вещи (объекты, процессы, механизмы, состояния, отношения), объективно существующие, но принципиально ненаблюдаемые, независимо от способности ученого их заметить. Например, интеллект может мыслиться как реальность, влияющая на ответы по тесту способностей и доступная изучению через эти ответы. Во-вторых, термин «конструкт» может использоваться как чисто лингвистическое образование, необходимое для обозначения классов объектов, процессов или состояний, используемое в том числе для нужд познания или деятельности. В этом значении «интеллект» является понятием, обобщающим умения и знания, нужные для решения задач теста. Наконец, термин «конструкт» может использоваться в техническом смысле для описания и объяснения статистических феноменов, например когда нужно охарактеризовать распределение переменных. Здесь он часто подменяет понятия «фактор», «латентная переменная» и «размерность». Именно в этом значении Ч. Спирмен ввел понятие общего фактора интеллекта. В последнее время предпринимаются попытки определить конструкт как эмерджентное понятие [Lange et al., 2020]. В общем, такая непроясненность важнейше-

[3] Интереснейшее описание значений, в которых употребляется термин «конструкт», можно найти в специальном выпуске журнала *New Ideas in Psychology* (№ 1 за 2013 г.).

го психометрического термина имеет серьезные последствия для интерпретации результатов моделирования конструктов в психометрике [Maraun, Gabriel, 2013].

Имея в виду неопределенность термина «конструкт», будем все же считать, что в заявлениях о возможности построить «измерительную модель латентного конструкта» последний понимается как некоторый психологический феномен (черта, способность, установки и проч.), лежащий за пределами теста, детерминирующий ответы на тест и за счет этого доступный для моделирования и измерения. В этой статье «конструкт» будет взят именно в этом, весьма распространенном, значении. Например,

> «в психологии и образовании, IRT — один из главных инструментов измерения способностей и установок» [Linden van der, Hambleton, 2013]; «целью является проиллюстрировать, как пользователи теста и исследователи могут применять многомерную IRT (MIRT) для того, чтобы понять, что их тесты измеряют...» [Ackerman, Gierl, Walker, 2003]; «обсуждаются выводы для исследования критического мышления на основе измерений в рамках IRT» [Wagner, Harvey, 2006]; «Является ли креативность специфической или общей способностью? В этом исследовании использованы многоуровневые эксплораторные IRT модели <...> Результаты предполагают, что креативность является скорее общей, а не специфической способностью» [Qian, Plucker, Yang, 2019].

Насколько обоснованны подобные утверждения? Эта статья посвящена анализу логики психометрического моделирования как способа репрезентировать латентный конструкт или обнаружить его структуру. Будет показано, что, во-первых, то, что в психометрике получило название «моделирование», или «построение измерительной модели», имеет весьма отдаленное отношение к моделированию как к общенаучному методу; во-вторых, что психометрическое моделирование в принципе не может дать информацию, которую можно было бы использовать для понимания того феномена, который подлежит моделированию.

Моделирование как метод исследования

Поскольку моделирование — общенаучный метод, применяемый для решения широкого круга задач, например в экологии [Arhonditsis et al., 2006], эволюционной биологии [Pugesek, Tomer, von Eye, 2003], генетике [Franić et al., 2012], медицине [Ottensen, 2000], в исследованиях экономического поведения [Birnbaum, 2008], спорте, рассмотрим общую логику этого метода.

Любая модель используется, чтобы описывать, объяснять и предсказывать поведение реальной системы. Во всех случаях моделируемый феномен тщательно изучается на предмет внутренних закономерностей его существования. Эти закономерности должны быть репрезентированы моделью, поэтому именно результаты предварительных исследований составляют параметры модели. Далее на основе сравнения расчетных данных с реальными модель корректируется или выбирается лучший ее вариант. Параметры, которые не повышают точность прогноза, будут, скорее всего, исключаться из модели. Поэтому обратная связь от поведения реальной системы абсолютно необходима для моделирования и коррекции модели. Ясно также, что эта обратная связь относительно точности модели должна генерироваться независимо от модели [Oberkampf et al., 2022].

Проиллюстрируем принципы моделирования и его возможности двумя примерами. В климатологии математические модели эффективны для проверки предположений о факторах, приводящих к тем или иным экологическим последствиям. Например, *Soil Water Atmosphere Plant* (SWAP) моделирует перенос воды и тепла в зоне грунтовых вод во взаимодействии с развитием растительности [Айзель, Гусев, Насонова, 2017; Dam van et al., 2008]. В модели используется уравнение, которое включает показатели корневой экстракции воды, гидрофильности почвы (которая зависит, в свою очередь, от глубины ее промерзания), движения почвенной влаги, водосброса, а также теплоемкости и теплопроводности почвы. Для оценки переноса воды и растворенных веществ SWAP учитывает основные количественные закономерности процессов конвекции, дисперсии, адсорбции и др., также известные до и независимо от модели. В итоге модель позволяет объяснить явления засухи на изучаемой территории, а также обеспечение влагой растений и их рост. Модель многократно тестировалась в отношении разных климатических условий и показала свои преимущества перед другими моделями в прогнозе многих событий.

Другой пример: в фигурном катании, чтобы проверить гипотезы о влиянии специфических движений на высоту прыжка, может использоваться его математическая модель. Для этого сначала собирается информация о параметрах прыжка, например из видеозаписей его выполнения. У этих же фигуристов дополнительно оценивается сила плеч и мышц ног при разных угловых скоростях и типичная высота прыжков [Podolsky et al., 1990]. На основе корреляционных данных выявляются параметры, наиболее сильно связанные с высотой прыжка, и используются для параметризации математической модели. Качество модели оценивается по ее предсказательной силе в отношении

высоты реальных прыжков на соревнованиях, т.е. в отношении информации, полученной независимо от модели. Данные моделирования находят применение при разработке программ силовых тренировок для фигуристов, а также при прогнозировании риска ошибок или падений при выполнении прыжков [Rhodes, Putkaradze, 2022].

Психометрическое моделирование латентных конструктов иногда описывается и в другой логике — логике выявления ненаблюдаемых объектов [Maraun, 2017]. Она лежит в основе методологии изготовления и коррекции детекторов, самые известные из них — лакмусовая бумага и металлодетектор. В самом деле, психометрика прибегает к терминологии, близкой именно этой логике: «наблюдаемая переменная», «латентный конструкт», «манифестируемая переменная», «индикатор» и проч. Что пытается сделать психометрика по отношению к некоторой ненаблюдаемой реальности — это выработать и обосновать метод детекции этой реальности.

М. Мараун выделяет следующие принципы в технологии создания и работы любого детектора.

1. При необходимости принять решение относительно присутствия специфического ненаблюдаемого объекта создается протокол обнаружения.
2. Такой протокол включает спецификацию объектов класса *U* путем установления правила, определяющего свойства, которыми должен обладать объект, чтобы считаться элементом *U*.
3. Задаются логические суждения, которые связывают наличие элемента *u* из *U* с наблюдаемым «сигналом» *O*. Возможны три типа таких суждений:
   - «Если *U*, тогда *O* (*O* — необходимое условие *u*)»;
   - «Если *O*, тогда *U* (*O* достаточное условие *u*)»;
   - «*U* присутствует, если и только если *O* (*O* — необходимое и достаточное условие *u*)».
4. Инструмент обнаружения (детектор) позволяет принимать решение о наличии или отсутствии *u*. Детектор — это реализация конкретного логического суждения.
5. Определяются дополнительные условия, которые должны быть выполнены для того, чтобы детектор функционировал должным образом [Ibid.].

Проиллюстрируем эту логику на примере работы металлодетектора. В этом случае объектами обнаружения являются металлические предметы, т.е. элементы, обладающие свойствами металлов, а правила определения металлов известны заранее

и независимо от детектора. Суждение о присутствии искомого объекта относится к третьему типу: электромагнитный импульс определенной длительности возникает тогда и только тогда, когда вблизи есть металлический объект. Импульсно-индукционный металлодетектор является реализацией этого логического суждения. Побочные условия: импульсно-индукционный металлодетектор не работает вблизи телевизоров, радиоприемников, сотовых телефонов и других устройств, производящих радиоволны [Maraun, 2017].

Обобщая вышесказанное, можно заключить, что моделирование и/или детекция как общенаучные методы исследования требуют наличия предварительно установленных количественных закономерностей в эмпирической системе (объекте, процессах, взаимодействий и проч.), в отношении которой строится модель или детектор. Они нужны для параметризации модели или установления правил принятия решения при обнаружении сигналов детектором, а также для понимания граничных условий работоспособности модели или детектора. Кроме того, обратная связь для оценки модели должна обеспечиваться независимо от данных, поставляемых этой моделью. Так, информация о наличии металла должна поступить независимо от данных детектора, чтобы можно было оценить эффективность последнего.

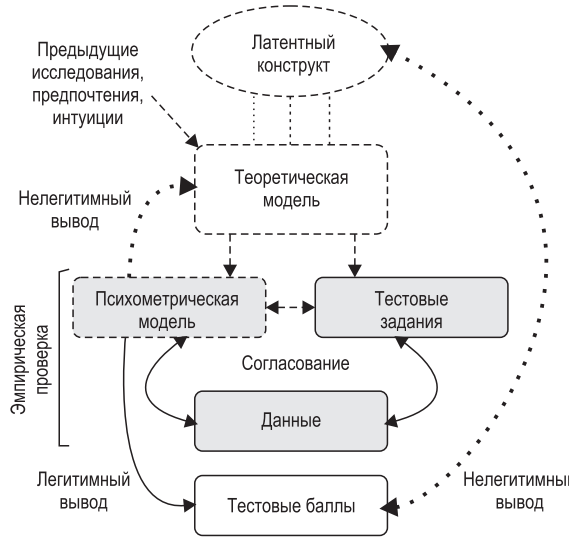
Моделирование в психометрике

Современные психометрические методы моделирования очень разнообразны (обзор см. в [Sen, Cohen, 2019]). Для целей статьи мы остановимся на моделировании латентных конструктов в рамках современной теории тестирования (IRT-моделирование), близкородственной факторному анализу. Однако описываемая ниже логика во многом распространяется и на другие типы психометрических моделей, в том числе сетевые.

Процесс моделирования можно описать как последовательность шагов или содержательных блоков (рис. 1). Первый, предварительный, шаг — это просто называние и описание психологического феномена (конструкта) или процесса, который должен быть замоделирован и измерен. К примеру, научное мышление может быть описано как мышление с тенденцией формулировать суждения относительно наблюдаемых явлений в форме альтернативных гипотез и условий принятия этих гипотез после проверки. Второй шаг — это формулирование психологической теории или гипотезы относительно природы конструкта, его аспектов, структуры и компонентов, внутренних связей, а также взаимодействия этого конструкта с другими психологическими феноменами или объективными

событиями в жизни человека. Такая теоретическая модель относительно природы «латентного конструкта» может строиться на основе предыдущих исследований, наблюдений, интуиций или суждений экспертов об этой черте/способности[4]. Например, теоретическая модель научного мышления может включать комбинацию двух умений: формулировать гипотезу и выбирать метод ее проверки.

Рис. 1. **Моделирование в психометрике**



Примечание: Блоки и стрелки — компоненты моделирования, их последовательность и взаимовлияние; штриховой линией помечены гипотетические компоненты. Только затененные компоненты вовлечены в эмпирическую проверку и потому дают легитимные выводы. Выводы из психометрической модели, сделанные в отношении теоретической модели и латентного конструкта, нелегитимны (пунктирные стрелки).

Третий шаг — это проверка теоретической модели. Для этой проверки используется психометрическая модель. Если она является формализованным вариантом психологической теории конструкта, будут сформированы соответствующие статистические ожидания от распределения данных по тесту, например о размерности (факторной структуре) теста и соответствии опре-

[4] Вообще говоря, теории относительно психологических феноменов могут быть вполне сформированными, например теория Ж. Пиаже о развитии мышления ребенка, которая разрабатывалась много лет. Такие теории уже не полагаются на интуиции, а имеют в своей основе систематически собранный экспериментальный или другой эмпирический материал. Однако такие теории нечасто используются для представления «латентных конструктов». Причины этого, а также различия между экспериментальными методами психологии и психометрическим моделированием не будут обсуждаться в этой работе, хотя они чрезвычайно важны.

деленных заданий «своей» размерности. Так, данные по тесту научного мышления в соответствии с вышеуказанной теорией должны образовывать два фактора: включающие задания на построение гипотез и на знание различных методов исследования.

Технически проверка теоретической модели конструкта реализуется через психометрическое описание полученных данных — через проверку размерности, распределения заданий по размерностям и проч. Причем важно, чтобы формально-статистическое описание соответствовало полученным данным как можно более точно. Наилучшая модель затем сравнивается с теоретически ожидаемой.

При выборе наилучшей модели исследователь опирается на результаты анализа согласия модели с полученными данными. Оценка согласия, как правило, включает оценку размерности, согласия данных с моделью на уровне отдельных заданий, согласия модели с данными с принятием решения о модели, лучше всего подходящей данным, а также локальной независимости заданий, предсказаний модели в отношении получаемых данных по тесту и др. [Hambleton, Swaminathan, 2013; Yen, Fizpatrick, 2006].

Остановимся на философии выбора модели чуть подробнее, потому что это единственный способ принять, отклонить или скорректировать психометрическую модель — а следовательно, и модель теоретическую. В целом выбор модели может быть описан как итеративный процесс, опции которого располагаются в континууме: с одной стороны, модель определяет, какие должны быть данные, чтобы соответствовать модели, так что несогласующиеся данные следует отбросить [Rasch, 1960], а с другой стороны, собранные по тесту данные определяют то, какая модель должна быть найдена, чтобы наилучшим образом согласовываться с ними [Hambleton, Swaminathan, 2013; Yen, Fizpatrick, 2006]. В 70–80-х годах прошлого века, когда IRT-модели постепенно приобретали популярность, между сторонниками разных подходов к выбору модели шли интенсивные дебаты [Divgi, 1986]. В настоящее время решение о выборе модели может основываться на типе полученных данных (ранги, классификация или их комбинация), целей моделирования (использовать результаты для оценки респондентов, отразить в модели композитность конструкта и проч.), выдвинутых моделью допущений, соображений удобства статистической обработки данных и интерпретации [Luo, 2021; Robitzsch, 2022].

Когда решение принято, исследователь может сделать два класса выводов (рис. 1). Первый касается шкалирования, начисления баллов за тест, выделения в тесте подшкал, надежности инструмента, целесообразности подсчета общего балла по тесту и проч. [Reise, 2012]. В отличие от этих, сугубо практи-

ческих выводов, выводы второго класса имеют непосредственное отношение к латентному конструкту и соответствующей теоретической модели: согласованная с данными психометрическая модель прямо накладывается на структуру латентного конструкта. Иллюстрацией таких выводов служат формулировки, используемые в статьях и их заголовках, например:

> «Обзор психометрических подходов к раскрытию структуры психиатрических конструктов» [Borsboom et al., 2016]; «Цель исследования заключалась в анализе психометрических свойств испанской версии шкалы тревожности в связи с коронавирусом (CAS) с использованием <...> IRT и конфирматорного факторного анализа <...> Модели <...> показали, что CAS более информативен при высоком уровне тревожности по COVID-19. CAS обладает достаточными психометрическими свойствами для использования в качестве краткой меры тревожности по COVID-19» [Caycho-Rodríguez et al., 2022]; «Из трех протестированных нами моделей неприемлемой была признана та, которая представляла реляционное мышление как одномерный конструкт <...> Оставшиеся две модели представляли реляционное мышление как многомерное <...> Что в конечном итоге отличало эти две модели, так это включение фактора более высокого порядка сверх факторов, представляющих четыре проявления реляционного мышления» [Alexander et al., 2016].

Однако для такого рода утверждений о конструкте требуется, чтобы, во-первых,тест и данные по тесту были связаны с исследуемым конструктом, и только с ним, и, во-вторых, чтобы была известна функция, связывающая тест и конструкт. Но выполняется ли это условие?

Задания теста и конструкт

Задания теста разрабатываются с таким расчетом, чтобы они вызывали поведение, подпадающее под описание латентного конструкта и/или его компонентов [Messick, 1994; Kane, 2016]. Например, если латентный конструкт — это способность к научному мышлению, а теоретическая модель научного мышления включает два его аспекта, а именно умение формулировать гипотезы и знание методов их проверки, то задания теста будут требовать, соответственно, формулировки гипотезы и указания методов их проверки.

Удостовериться в том, что задания вызывают нужное поведение, можно, проведя предварительные «качественные» исследования. Так, можно убедиться, что задание на формулировку гипотезы провоцирует формулировку гипотезы, а не вспоминание научных фактов, например. Кроме того, в «каче-

ственном» исследовании проверяется, понятны ли вопросы, удобно ли расположен стимульный материал, достаточно ли выделенного на решение времени и проч.[5] [Mislevy, Steinberg, Almond, 2002; Messick, 1994].

Однако из полученных в предварительном исследовании свидетельств того, что задания теста вызывают нужное поведение, не следует, что это поведение детерминировано латентным конструктом. И тем более эти свидетельства не проясняют функцию, связывающую особенности того или иного поведения (например, скорость или правильность решения задачи) с уровнем выраженности латентного конструкта. Причин тому две. Во-первых, различия между людьми в выполнении тех или иных действий могут объясняться не искомым специфическим конструктом, а, например, предыдущим опытом, специальными знаниями, особенностями нервной системы, а также взаимовлиянием этих факторов [Costantini et al., 2015]. Во-вторых, типичный дизайн исследования, направленного на сбор свидетельств валидности теста, — корреляционный. Иными словами, никакая критериальная валидизация и полученные в результате статистики в принципе не могут гарантировать тесту статус измеряющего инструмента или хотя бы инструмента, чувствительного к искомому конструкту [Borsboom, Mellenbergh, van Heerden, 2004; Trendler, 2013; 2022; Uher, 2021].

Задача представить баллы по тесту как математическую функцию от изучаемого психологического конструкта ($y = f(x)$) принципиально нерешаема ни процедурно (например, экспертизой), ни технологически (например, *evidence-centered design*). Для ее решения должно быть выполнено невыполнимое требование: необходимо иметь одновременные значения для обоих членов уравнения, тогда как мы можем наблюдать значения только для одного из них — это ответ на задание теста. Эта ситуация с тестами — современная версия психофизической проблемы, известной в истории психологии с тех времен, когда психофизики пытались установить соответствие между феноменами «телесными» и феноменами «душевными» [Johnson, 1945]. Все попытки решить эту проблему приводили и приводят только к «предположительным выводам», которые нельзя проверить эмпирически.

Таким образом, каким бы способом ни был разработан тест, он лишь гипотетически соотносится с искомым латентным конструктом. Поэтому тест не может быть использован в моделировании как источник данных о конструкте.

[5] Разработка теста — гораздо более сложный процесс, чем описано, здесь представлены только принципиальные шаги и направления проверки.

Легитимность выводов психометрического моделирования

Повторимся: достигнутое согласие модели и данных позволяет делать два класса выводов — прикладные и теоретические. Прикладные выводы касаются структуры данных (например, вывод о количестве факторов) и оптимальных вариантов скоринга. Прикладные выводы легитимны, так как именно структура распределения данных и оценивалась. Теоретические выводы касаются латентного конструкта: его психологической структуры и его оценки. Эти выводы нелегитимны (см. рис. 1, пунктирные стрелки).

Моделирование конструкта в психометрике не следует логике репрезентации конструкта, так как структура последнего неизвестна и поэтому репрезентировать его невозможно. Моделирование конструкта не следует и логике обнаружения объекта, так как для плохо изученного конструкта (феномена) не описан класс объектов, признаки которых нужно обнаружить, не создан протокол обнаружения, не сформированы правила вывода и принятия решения об обнаружении. В отличие от моделей в естественных науках, в психометрике правила связывания входящих переменных с «латентными конструктами» подбираются в процессе анализа и независимо от функционирования изучаемой эмпирической системы, поэтому модель не обнаруживает «латентные переменные», но конструирует их способом, наиболее оптимальным для предписанной модели и входящих переменных [Maraun, Halpin, 2008]. Кроме того, не вводятся ограничения адекватности модели: модель предположительно может адекватно описывать «структуру конструкта» для всех людей, независимо от возраста, рода деятельности, социального и культурного контекста, здоровья и т.д. Так что, какой бы модели в конечном итоге ни соответствовали данные тестирования и как бы модель ни отклонялась от данных, мы не получим информации о конструкте просто потому, что неизвестно, с чем связано отклонение модели или согласие данных с моделью — с данными или с моделью. Любые результаты моделирования можно объяснить и неверным представлением о конструкте (например, двухкомпонентная теория неверна), и неверными данными (например, погрешностями теста или выборки).

Моделирование и ad hoc теории в психологии

Прямым следствием такого положения дел является генерация *ad hoc*[6] теорий относительно психологического конструкта (феномена). Их особенность состоит в том, что всякий раз, когда удается привести к согласию какую-то модель с каким-то набором данных, эта модель принимается как теория, объясняющая

[6] ad hoc — теории по случаю, пригодные для отдельного случая.

этот набор данных. При этом, если окажется, что эта модель не воспроизводится на новых данных, она модифицируется с таким расчетом, чтобы лучше подходить для новых данных, и превращается в новую «теорию». При этом старая модель продолжает существовать как объяснение, пригодное для предыдущего случая. То есть если один набор данных согласуется с моделью *X*, а другой — с моделью *Y*, то обе модели и оба теоретических объяснения валидны. В итоге возникает ситуация, когда в отношении одного и того же конструкта может существовать неограниченное число моделей, приведенных в согласие с данными тестов, так что конструкт может быть описан разными структурами, равно пригодными и равно не опровергнутыми [Vessonen, 2021].

Ad hoc теории можно легко обнаружить в рассуждениях о чертах или способностях, когда эти рассуждения опираются на данные моделирования. Например, по мысли авторов идеи конструкта «настойчивость», последняя состоит из «стабильности интересов» и «настойчивости усилий» [Duckworth, Quinn, 2012; Duckworth et al., 2007]. Конструкт исследуется в основном с помощью разных версий шкалы «Настойчивость». Результаты статистического анализа распределения ответов (структуры шкалы) существенно разнятся в разных исследованиях: одномерная, бифакторная или иерархическая структура, зависимы или независимы факторы друг от друга, является ли структура инвариантной и проч. Вследствие прямого накладывания статистической структуры данных теста на латентный конструкт настойчивость описывается соответственно как имеющая бифакторную, иерархическую структуру и т.п. [Tyumeneva, Kardanova, Kuzmina, 2019; Credé, 2018; Tynan, 2021]. Однако, поскольку в психометрических исследованиях черта не исследуется, а конструируется в процессе моделирования на основе тестовых данных, нет никаких перспектив опровергнуть или подтвердить эти теории. Теория, сконструированная на базе психометрической модели, всегда будет соответствовать «своим» данным, а новые данные будут находить «свою» модель.

Заключительные замечания

Наш анализ показал, что то, что психометрики имеют в виду под моделированием, не является моделированием в общенаучном понимании этого слова: ни в смысле репрезентации, ни в смысле обнаружения ненаблюдаемого объекта. Поэтому из двух классов выводов, обычно следующих из результатов психометрического моделирования — о структуре тестовых данных и о структуре латентного конструкта, — последние нельзя считать легитимными. Невозможно построить модель ненаблюдаемого объекта (конструкта), природа и закономер-

ности функционирования которого не установлены, полагаясь на данные, связь которых с конструктом также не установлена.

Вместо того чтобы использовать результаты психометрического анализа данных теста строго по назначению — для шкалирования, скоринга, выявления неинформативных заданий и проч., исследователи делают из них содержательные выводы о психологических характеристиках. Такое использование психометрики не только тормозит развитие психологического знания в целом, но и уводит его в сторону безосновательного и непродуктивного «теоретизирования».

Причины распространения практики прямого переноса выводов психометрического моделирования на психологические феномены заслуживают особого внимания, но не будут здесь обсуждаться. Эта работа ставила целью лишь показать, *что* именно не дает психометрическому моделированию занять нишу полноценного исследовательского инструмента в психологии, каким бы удобным и привычным ни представлялся этот путь.

Благодарности

Исследование реализовано при поддержке факультета социальных наук, Национальный исследовательский университет «Высшая школа экономики».

Автор благодарит обоих рецензентов за содержательную критику и мотивацию к дискуссии.

Литература

1. Айзель Г.В., Гусев Е.М., Насонова О.Н. (2017) Расчеты речного стока на основе модели SWAP для водосборов с недостаточным информационным обеспечением. 2. Использование методов физико-географического подобия и пространственной геостатистики. *Водные ресурсы*, т. 44, № 4, сс. 419–431. https://doi.org/10.7868/S0321059617040022
2. Угланова И.Л., Брун И.В., Васин Г.М. (2018) Методология *Evidence-Centered Design* для измерения комплексных психологических конструктов. *Современная зарубежная психология*, т. 7, № 3, сс. 18–27. https://doi.org/10.17759/jmfp.2018070302
3. Ackerman T.A., Gierl M.J., Walker C.M. (2003) Using Multidimensional Item Response Theory to Evaluate Educational and Psychological Tests. *Educational Measurement: Issues and Practice*, vol. 22, no 3, pp. 37–51. http://dx.doi.org/10.1111/j.1745-3992.2003.tb00136.x
4. Alexander P.A., Dumas D., Grossnickle E.M., List A., Firetto C.M. (2016) Measuring Relational Reasoning. *The Journal of Experimental Education*, vol. 84, no 1, pp. 119–151. http://dx.doi.org/10.1080/00220973.2014.963216
5. Araujo A.L.S.O., Andrade W.L., Guerrero D.D.S., Melo M.R.A. (2019) How Many Abilities Can We Measure in Computational Thinking? A Study on Bebras Challenge. Proceedings of the *50th ACM Technical Symposium on Computer Science Education* (Minneapolis, MN, 2019, 27 February), New York, NY: Machinery, pp. 545–551.

6. Arhonditsis G.B., Stow C.A., Steinberg L.J., Kenney M.A., Lathrop R.C., McBride S.J., Reckhow K.H. (2006) Exploring Ecological Patterns with Structural Equation Modeling and Bayesian Analysis. *Ecological Modelling*, vol. 192, no 3–4, pp. 385–409. https://doi.org/10.1016/j.ecolmodel.2005.07.028
7. Birenbaum M., DeLuca C., Earl L., Heritage M., Klenowski V., Looney A. et al. (2015) International Trends in the Implementation of Assessment for Learning: Implications for Policy and Practice. *Policy Futures in Education*, vol. 13, no 1, pp. 117–140. http://dx.doi.org/10.1177/1478210314566733
8. Birnbaum M.H. (2008) New Paradoxes of Risky Decision Making. *Psychological Review*, vol. 115, no 2, pp. 463–501. https://doi.org/10.1037/0033-295X.115.2.463
9. Borsboom D., Mellenbergh G.J., van Heerden J. (2004) The Concept of Validity. *Psychological Review*, vol. 111, no 4, pp. 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061
10. Borsboom D., Molenaar D. (2015) Psychometrics. *International Encyclopedia of the Social & Behavioral Sciences* (ed. J.D. Wright), Oxford: Elsevier, pp. 418–422. https://doi.org/10.1016/B978-0-08-097086-8.43079-5
11. Borsboom D., Rhemtulla M., Cramer A.O., van der Maas H.L., Scheffer M., Dolan C.V. (2016) Kinds Versus Continua: A Review of Psychometric Approaches to Uncover the Structure of Psychiatric Constructs. *Psychological Medicine*, vol. 46, no 8, pp. 1567–1579. http://dx.doi.org/10.1017/S0033291715001944
12. Buchholz J., Hartig J. (2020) Measurement Invariance Testing in Questionnaires: A Comparison of Three Multigroup-CFA and IRT-Based Approaches. *Psychological Test and Assessment Modeling*, vol. 62, no 1, pp. 29–53.
13. Caycho-Rodríguez T., Vilca L.W., Carbajal-León C., White M., Vivanco-Vidal A., Saroli-Araníbar D. et al. (2022) Coronavirus Anxiety Scale: New Psychometric Evidence for the Spanish Version Based on CFA and IRT Models in a Peruvian Sample. *Death Studies*, vol. 46, no 5, pp. 1090–1099. http://dx.doi.org/10.1080/07481187.2020.1865480
14. Costantini G., Epskamp S., Borsboom D., Perugini M., Mõttus R., Waldorp L.J., Cramer A.O. (2015) State of the aRt Personality Research: A Tutorial on Network Analysis of Personality Data in R. *Journal of Research in Personality*, vol. 54, July, pp. 13–29. https://doi.org/10.1016/j.jrp.2014.07.003
15. Credé M. (2018) What Shall We Do about Grit? A Critical Review of What We Know and What We Don't Know. *Educational Researcher*, vol. 47, no 9, pp. 606–611. http://dx.doi.org/10.3102/0013189X18801322
16. Cronbach L.J., Meehl P.E. (1955) Construct Validity in Psychological Tests. *Psychological Bulletin*, vol. 52, no 4, pp. 281–302. https://doi.org/10.1037/h0040957
17. Dam van J.C., Groenendijk P., Hendriks R.F., Kroes J.G. (2008) Advances of Modeling Water Flow in Variably Saturated Soils with SWAP. *Vadose Zone Journal*, vol. 7, no 2, pp. 640–653. http://dx.doi.org/10.2136/vzj2007.0060
18. Divgi D.R. (1986) Does the Rasch Model Really Work for Multiple Choice Items? Not If You Look Closely. *Journal of Educational Measurement*, vol. 23, no 4, pp. 283–298.
19. Duckworth A.L., Quinn P.D. (2012) Short Grit Scale. *Journal of Personality Assessment*, vol. 91, no 2, pp. 166–174. https://psycnet.apa.org/doi/10.1037/t01598-000
20. Duckworth A.L., Peterson C., Matthews M.D., Kelly D.R. (2007) Grit: Perseverance and Passion for Long-Term Goals. *Journal of Personality and Social Psychology*, vol. 92, no 6, 1087–1101. http://dx.doi.org/10.1037/0022-3514.92.6.1087
21. Dumas D., Dong Y. (2022) Relational Reasoning and Thinking: Theory, Measurement, and Empirical Findings. *International Encyclopedia of Education* (eds R. Tierney, F. Rizvi, K. Ercican), New York, NY: Taylor & Francis. https://doi.org/10.4324/9781138609877-REE179-1

22. Fischer G.H. (1973) The Linear Logistic Test Model as an Instrument in Educational Research. *Acta Psychologica*, vol. 37, no 6, pp. 359–374. http://dx.doi.org/10.1016/0001-6918(73)90003-6
23. Fisher Jr. W.P., Stenner A.J. (2022) Metrology for the Social, Behavioral, and Economic Sciences. *Explanatory Models, Unit Standards, and Personalized Learning in Educational Measurement: Selected Papers by A. Jackson Stenner* (eds W.P. Fisher, P.J. Massengill), Singapore: Springer Nature Singapore, pp. 217–222.
24. Fox J.P. (2005) Multilevel IRT Using Dichotomous and Polytomous Response Data. *British Journal of Mathematical and Statistical Psychology*, vol. 58, no 1, pp. 145–172. http://dx.doi.org/10.1348/000711005X38951
25. Franić S., Borsboom D., Dolan C.V., Boomsma D.I. (2014) The Big Five Personality Traits: Psychological Entities or Statistical Constructs? *Behavior Genetics*, vol. 44, no 6, pp. 591–604. http://dx.doi.org/10.1007/s10519-013-9625-7
26. Franic S., Dolan C.V., Borsboom D., Boomsma D.I. (2012) Structural Equation Modeling in Genetics. *Handbook of Structural Equation Modeling* (ed. R.H. Hoyle), New York, NY: The Guilford, pp. 617–635.
27. Freund P.A., Lohbeck A. (2021) Modeling Self-Determination Theory Motivation Data by Using Unfolding IRT. *European Journal of Psychological Assessment*, vol. 37, no 5, pp. 388–396. http://dx.doi.org/10.1027/1015-5759/a000629
28. Hambleton R.K., Swaminathan H. (2013) *Item Response Theory: Principles and Applications*. Springer Science & Business Media.
29. Hartig J., Höhler J. (2009) Multidimensional IRT Models for the Assessment of Competencies. *Studies in Educational Evaluation*, vol. 35, no 2–3, pp. 57–63. http://dx.doi.org/10.1016/j.stueduc.2009.10.002
30. Hauwaert van S.M., Schimpf C.H., Azevedo F. (2020) The Measurement of Populist Attitudes: Testing Cross-National Scales Using Item Response Theory. *Politics*, vol. 40, no 1, Article no 026339571985930. http://dx.doi.org/10.1177/0263395719859306
31. Johnson H.M. (1945) Are Psychophysical Problems Genuine or Spurious? *The American Journal of Psychology*, vol. 58, no 2, pp. 189–211. https://doi.org/10.2307/1417845
32. Kane M.T. (2016) Explicating Validity. *Assessment in Education: Principles, Policy & Practice*, vol. 23, no 2, pp. 198–211. https://doi.org/10.1080/0969594X.2015.1060192
33. Kunina-Habenicht O., Goldhammer F. (2020) ICT Engagement: A New Construct and Its Assessment in PISA 2015. *Large-Scale Assessments in Education*, vol. 8, no 1, pp. 1–21. http://dx.doi.org/10.1186/s40536-020-00084-z
34. Lange J., Dalege J., Borsboom D., van Kleef G.A., Fischer A.H. (2020) Toward an Integrative Psychometric Model of Emotions. *Perspectives on Psychological Science*, vol. 15, no 2, pp. 444–468. http://dx.doi.org/10.1177/1745691619895057
35. Linden van der W.J., Hambleton R.K. (eds) (2013) *Handbook of Modern Item Response Theory*. Springer Science & Business Media.
36. Luo Y. (2021) A Comparison of Common IRT Model-selection Methods with Mixed-Format Tests. *Measurement: Interdisciplinary Research and Perspectives*, vol. 19, no 4, pp. 199–212. http://dx.doi.org/10.1080/15366367.2021.1878779
37. MacCorquodale K., Meehl P.E. (1948) On a Distinction between Hypothetical Constructs and Intervening Variables. *Psychological Review*, vol. 55, no 2, pp. 95–107. https://doi.org/10.1037/h0056029
38. Maraun M. (2017) The Object Detection Logic of Latent Variable Technologies. *Quality and Quantity*, vol. 51, no 1, pp. 239–259. https://doi.org/10.1007/s11135-015-0303-0
39. Maraun M.D., Gabriel S.M. (2013) Illegitimate Concept Equating in the Partial Fusion of Construct Validation Theory and Latent Variable Modeling. *New Ideas in Psychology*, vol. 31, no 1, pp. 32–42. https://doi.org/10.1016/j.newideapsych.2011.02.006

40. Maraun M.D., Halpin P.F. (2008) Manifest and Latent Variates. *Measurement: Interdisciplinary Research and Perspectives*, vol. 6, no 1–2, pp. 113–117. https://doi.org/doi:10.1080/15366360802035596
41. Markus K.A., Borsboom D. (2013) *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning.* New York, NY: Routledge/Taylor & Francis Group. https://doi.org/10.4324/9780203501207
42. Messick S. (1994) The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*, vol. 23, no 2, pp. 13–23. https://doi.org/10.3102/0013189x023002013
43. Michell J. (2013) Constructs, Inferences, and Mental Measurement. *New Ideas in Psychology*, vol. 31, no 1, pp. 13–21. https://doi.org/10.1016/j.newideapsych.2011.02.004
44. Mislevy R.J., Steinberg L.S., Almond R.G. (2002) On the Roles of Task Model Variables in Assessment Design. Generating Items for Cognitive Tests: Theory and Practice (eds S. Irvine, P. Kyllonen), Hillsdale, NY: Erlbaum, pp. 97–128.
45. Nering M.L., Ostini R. (eds) (2010) *Handbook of Polytomous Item Response Theory Models*. New York, NY: Routledge. https://doi.org/10.4324/9780203861264
46. Nima A.A., Cloninger K.M., Persson B.N., Sikström S., Garcia D. (2020) Validation of Subjective Well-Being Measures Using Item Response Theory. *Frontiers in Psychology*, vol. 10, January, Article no 3036. http://dx.doi.org/10.3389/fpsyg.2019.03036
47. Oberkampf W.L., DeLand S.M., Rutherford B.M., Diegert K.V., Alvin K.F. (2002) Error and Uuncertainty in Mmodeling and Ssimulation. *Reliability Engineering & System Safety*, vol. 75, no (3), pp. 333–357. http://dx.doi.org/10.1016/S0951-8320(01)00120-X
48. Ottensen J. (2000) *Mathematical Modelling in Medicine*. Amsterdam: IOS Press.
49. Podolsky A., Kaufman K.R., Cahalan T.D., Aleshinsky S.Y., Chao E.Y. (1990) The Relationship of Strength and Jump Height in Figure Skaters. *The American Journal of Sports Medicine*, vol. 18, no 4, pp. 400–405. https://doi.org/10.1177/036354659001800412
50. Power M.J. (2006) The Structure of Emotion: An Empirical Comparison of Six Models. *Cognition and Emotion,* vol. 20, no 5, pp. 694–713. https://doi.org/10.1080/02699930500367925
51. Pugesek B.H., Tomer A., von Eye A. (2003) *Structural Equation Modeling: Applications in Ecological and Evolutionary Biology*. Cambridge, UK: Cambridge University. https://doi.org/10.1017/CBO9780511542138
52. Qian M., Plucker J.A., Yang X. (2019) Is Creativity Domain Specific or Domain General? Evidence from Multilevel Explanatory Item Response Theory Models. *Thinking Skills and Creativity*, vol. 33, May, Article no 100571. http://dx.doi.org/10.1016/j.tsc.2019.100571
53. Rasch G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danmarks Paedagogiske Institut.
54. Ravand H., Robitzsch A. (2015) Cognitive Diagnostic Modeling Using R. *Practical Assessment, Research, and Evaluation*, vol. 20, no 11. Available at: http://pareonline.net/getvn.asp?v=20&n=11 (accessed 20 August 2023).
55. Reise S.P. (2012) The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research,* vol. 47, no 5, pp. 667–696. https://doi.org/10.1080/00273171.2012.715555
56. Rhodes M., Putkaradze V. (2022) Trajectory Tracing in Figure Skating. *Nonlinear Dynamics*, vol. 110, no 4, pp. 3031–3044. https://doi.org/10.1007/s11071-022-07806-8
57. Riconscente M.M., Mislevy R.J., Corrigan S. (2015) Evidence-Centered Design. *Handbook of Test Development* (eds S. Lane, M.R. Raymond, T.M. Haladyna), New York, NY: Routledge, pp. 40–63. http://dx.doi.org/10.4324/9780203102961.ch3

58. Robitzsch A. (2022) On the Choice of the Item Response Model for Scaling PISA Data: Model Selection Based on Information Criteria and Quantifying Model Uncertainty. *Entropy*, vol. 24, no 6, Article no 760. http://dx.doi.org/10.3390/e24060760
59. Schmittmann V.D., Cramer A.O.J., Waldorp L.J., Epskamp S., Kievit R.A., Borsboom D. (2013) Deconstructing the Construct: A Network Perspective on Psychological Phenomena. *New Ideas in Psychology*, vol. 31, no 1, pp. 43–53. https://doi.org/10.1016/j.newideapsych.2011.02.007
60. Sen S., Cohen A.S. (2019) Applications of Mixture IRT Models: A Literature Review. *Measurement: Interdisciplinary Research and Perspectives*, vol. 17, no 4, pp. 177–191. http://dx.doi.org/10.1080/15366367.2019.1583506
61. Shaw A., Kapnek M., Morelli N.A. (2021) Measuring Creative Self-Efficacy: An Item Response Theory Analysis of the Creative Self-Efficacy Scale. *Frontiers in Psychology*, vol. 12, July, Article no 678033. http://dx.doi.org/10.3389/fpsyg.2021.678033
62. Sijtsma K., Ark van der A. (2020) *Measurement Models for Psychological Attributes: Classical Test Theory, Factor Analysis, Item Response Theory, and Latent Class Models*. Boca Raton, FL: CRC. https://doi.org/10.1201/9780429112447
63. Streckert N., Kurtz L., Kajonius P.J. (2023) Can Your Darkness Be Measured? Analyzing the Full and Brief Version of the Dark Factor of Personality in Swedish. *International Journal of Testing*, vol. 23, no 2, pp. 1–45. http://dx.doi.org/10.1080/15305058.2023.2195659
64. Templin J.L., Henson R.A. (2006) Measurement of Psychological Disorders Using Cognitive Diagnosis Models. *Psychological Methods*, vol. 11, no 3, pp. 287–305. http://dx.doi.org/10.1037/1082-989X.11.3.287
65. Trendler G. (2022) Is Measurement in Psychology an Empirical or a Conceptual Issue? A Comment on David Franz. *Theory & Psychology*, vol. 32, no 1, pp. 164–170. https://doi.org/10.1177/09593543211050025
66. Trendler G. (2013) Measurement in Psychology: A Case of Ignoramus et Ignorabimus? A Rejoinder. *Theory & Psychology*, vol. 23, no 5, pp. 591–615. https://doi.org/10.1177/0959354313490451
67. Tynan M.C. (2021) Deconstructing Grit's Validity: The Case for Revising Grit Measures and Theory. *Multidisciplinary Perspectives on Grit: Contemporary Theories, Assessments, Applications and Critiques* (eds L.E. van Zyl, C. Olckers, L. van der Vaart), Cham: Springer Nature Switzerland, pp. 137–155. http://dx.doi.org/10.1007/978-3-030-57389-8_8
68. Tyumeneva Y., Kardanova E., Kuzmina J. (2019) Grit: Two Related but Independent Constructs Instead of One. Evidence from Item Response Theory. *European Journal of Psychological Assessment*, vol. 35, no 4, pp. 469–478. http://dx.doi.org/10.1027/1015-5759/a000424
69. Uher J. (2021) Quantitative Psychology under Scrutiny: Measurement Requires Not Result-Dependent But Traceable Data Generation. *Personality and Individual Differences*, vol. 170, no 5, Article no 110205. https://doi.org/10.1016/j.paid.2020.110205
70. Vessonen E. (2021) Conceptual Engineering and Operationalism in Psychology. *Synthese*, vol. 199, no 3–4, pp. 10615–10637. https://doi.org/10.1007/s11229-021-03261-x
71. Wagner T.A., Harvey R.J. (2006) Development of a New Critical Thinking Test Using Item Response Theory. *Psychological Assessment,* vol. 18, no 1, pp. 100–105. https://doi.org/10.1037/1040-3590.18.1.100
72. Walton K.E., Roberts B.W., Krueger R.F., Blonigen D.M., Hicks B.M. (2008) Capturing Abnormal Personality with Normal Personality Inventories: An Item Response Theory Approach. *Journal of Personality*, vol. 76, no 6, pp. 1623–1648. http://dx.doi.org/10.1111/j.1467-6494.2008.00533.x

73. Wiggins B.J., Christopherson C.D. (2019) The Replication Crisis in Psychology: An Overview for Theoretical and Philosophical Psychology. *Journal of Theoretical and Philosophical Psychology*, vol. 39, no 4, pp. 202–217. http://dx.doi.org/10.1037/teo0000137
74. Will C.M. (2000) *Einstein's Relativity and Everyday Life*. Available at: http://www.physicscentral.com/writers/writers-00-2.html (accessed 20 August 2023).
75. Wilson M. (2004) *Constructing Measures. An Item Response Modeling Approach*. New York, NY: Routledge.
76. Yen W.M., Fizpatrick A.R. (2006) Item Response Theory. *Educational Measurement* (ed. R.L. Brennan), Westport, CT: American Council on Education and Praeger, pp. 17–64.
77. Zhao H., Alexander P.A., Sun Y. (2021) Relational Reasoning's Contributions to Mathematical Thinking and Performance in Chinese Elementary and Middle-School Students. *Journal of Educational Psychology,* vol. 113, no 2, pp. 279–303. http://dx.doi.org/10.1037/edu0000595

References

Ackerman T.A., Gierl M.J., Walker C.M. (2003) Using Multidimensional Item Response Theory to Evaluate Educational and Psychological Tests. *Educational Measurement: Issues and Practice*, vol. 22, no 3, pp. 37–51. http://dx.doi.org/10.1111/j.1745-3992.2003.tb00136.x

Alexander P.A., Dumas D., Grossnickle E.M., List A., Firetto C.M. (2016) Measuring Relational Reasoning. *The Journal of Experimental Education*, vol. 84, no 1, pp. 119–151. http://dx.doi.org/10.1080/00220973.2014.963216

Araujo A.L.S.O., Andrade W.L., Guerrero D.D.S., Melo M.R.A. (2019) How Many Abilities Can We Measure in Computational Thinking? A Study on Bebras Challenge. Proceedings of the *50th ACM Technical Symposium on Computer Science Education* (Minneapolis, MN, 2019, 27 February), New York, NY: Machinery, pp. 545–551.

Arhonditsis G.B., Stow C.A., Steinberg L.J., Kenney M.A., Lathrop R.C., McBride S.J., Reckhow K.H. (2006) Exploring Ecological Patterns with Structural Equation Modeling and Bayesian Analysis. *Ecological Modelling*, vol. 192, no 3–4, pp. 385–409. https://doi.org/10.1016/j.ecolmodel.2005.07.028

Ayzel G.V., Gusev E.M., Nasonova O.N. (2017) Raschety rechnogo stoka na osnove modeli SWAP dlya vodosborov s nedostatochnym informatsionnym obespecheniem. 2. Ispol´zovanie metodov fiziko-geograficheskogo podpbiya i prostranstvennoy geostatistiki [Runoff Evaluation for Ungauged Watersheds by SWAP Model. 2. Using Methods of Physical and Geographical Similarity and Spatial Geostatistics]. *Water Resources*, vol. 44, no 4, pp. 419–431. https://doi.org/10.7868/S0321059617040022

Birenbaum M., DeLuca C., Earl L., Heritage M., Klenowski V., Looney A. et al. (2015) International Trends in the Implementation of Assessment for Learning: Implications for Policy and Practice. *Policy Futures in Education*, vol. 13, no 1, pp. 117–140. http://dx.doi.org/10.1177/1478210314566733

Birnbaum M.H. (2008) New Paradoxes of Risky Decision Making. *Psychological Review*, vol. 115, no 2, pp. 463–501. https://doi.org/10.1037/0033-295X.115.2.463

Borsboom D., Mellenbergh G.J., van Heerden J. (2004) The Concept of Validity. *Psychological Review*, vol. 111, no 4, pp. 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Borsboom D., Molenaar D. (2015) Psychometrics. *International Encyclopedia of the Social & Behavioral Sciences* (ed. J.D. Wright), Oxford: Elsevier, pp. 418–422. https://doi.org/10.1016/B978-0-08-097086-8.43079-5

Borsboom D., Rhemtulla M., Cramer A.O., van der Maas H.L., Scheffer M., Dolan C.V. (2016) Kinds Versus Continua: A Review of Psychometric Approaches to Un-

cover the Structure of Psychiatric Constructs. *Psychological Medicine*, vol. 46, no 8, pp. 1567–1579. http://dx.doi.org/10.1017/S0033291715001944

Buchholz J., Hartig J. (2020) Measurement Invariance Testing in Questionnaires: A Comparison of Three Multigroup-CFA and IRT-Based Approaches. *Psychological Test and Assessment Modeling*, vol. 62, no 1, pp. 29–53.

Caycho-Rodríguez T., Vilca L.W., Carbajal-León C., White M., Vivanco-Vidal A., Saroli-Araníbar D. et al. (2022) Coronavirus Anxiety Scale: New Psychometric Evidence for the Spanish Version Based on CFA and IRT Models in a Peruvian Sample. *Death Studies*, vol. 46, no 5, pp. 1090–1099. http://dx.doi.org/10.1080/07481187.2020.1865480

Costantini G., Epskamp S., Borsboom D., Perugini M., Mõttus R., Waldorp L.J., Cramer A.O. (2015) State of the aRt Personality Research: A Tutorial on Network Analysis of Personality Data in R. *Journal of Research in Personality*, vol. 54, July, pp. 13–29. https://doi.org/10.1016/j.jrp.2014.07.003

Credé M. (2018) What Shall We Do about Grit? A Critical Review of What We Know and What We Don't Know. *Educational Researcher*, vol. 47, no 9, pp. 606–611. http://dx.doi.org/10.3102/0013189X18801322

Cronbach L.J., Meehl P.E. (1955) Construct Validity in Psychological Tests. *Psychological Bulletin*, vol. 52, no 4, pp. 281–302. https://doi.org/10.1037/h0040957

Dam van J.C., Groenendijk P., Hendriks R.F., Kroes J.G. (2008) Advances of Modeling Water Flow in Variably Saturated Soils with SWAP. *Vadose Zone Journal*, vol. 7, no 2, pp. 640–653. http://dx.doi.org/10.2136/vzj2007.0060

Divgi D.R. (1986) Does the Rasch Model Really Work for Multiple Choice Items? Not If You Look Closely. *Journal of Educational Measurement*, vol. 23, no 4, pp. 283–298.

Duckworth A.L., Quinn P.D. (2012) Short Grit Scale. *Journal of Personality Assessment*, vol. 91, no 2, pp. 166–174. https://psycnet.apa.org/doi/10.1037/t01598-000

Duckworth A.L., Peterson C., Matthews M.D., Kelly D.R. (2007) Grit: Perseverance and Passion for Long-Term Goals. *Journal of Personality and Social Psychology*, vol. 92, no 6, 1087–1101. http://dx.doi.org/10.1037/0022-3514.92.6.1087

Dumas D., Dong Y. (2022) Relational Reasoning and Thinking: Theory, Measurement, and Empirical Findings. *International Encyclopedia of Education* (eds R. Tierney, F. Rizvi, K. Ercican), New York, NY: Taylor & Francis. https://doi.org/10.4324/9781138609877-REE179-1

Fischer G.H. (1973) The Linear Logistic Test Model as an Instrument in Educational Research. *Acta Psychologica*, vol. 37, no 6, pp. 359–374. http://dx.doi.org/10.1016/0001-6918(73)90003-6

Fisher Jr. W.P., Stenner A.J. (2022) Metrology for the Social, Behavioral, and Economic Sciences. *Explanatory Models, Unit Standards, and Personalized Learning in Educational Measurement: Selected Papers by A. Jackson Stenner* (eds W.P. Fisher, P.J. Massengill), Singapore: Springer Nature Singapore, pp. 217–222.

Fox J.P. (2005) Multilevel IRT Using Dichotomous and Polytomous Response Data. *British Journal of Mathematical and Statistical Psychology*, vol. 58, no 1, pp. 145–172. http://dx.doi.org/10.1348/000711005X38951

Franić S., Borsboom D., Dolan C.V., Boomsma D.I. (2014) The Big Five Personality Traits: Psychological Entities or Statistical Constructs? *Behavior Genetics*, vol. 44, no 6, pp. 591–604. http://dx.doi.org/10.1007/s10519-013-9625-7

Franic S., Dolan C.V., Borsboom D., Boomsma D.I. (2012) Structural Equation Modeling in Genetics. *Handbook of Structural Equation Modeling* (ed. R.H. Hoyle), New York, NY: The Guilford, pp. 617–635.

Freund P.A., Lohbeck A. (2021) Modeling Self-Determination Theory Motivation Data by Using Unfolding IRT. *European Journal of Psychological Assessment*, vol. 37, no 5, pp. 388–396. http://dx.doi.org/10.1027/1015-5759/a000629

Hambleton R.K., Swaminathan H. (2013) *Item Response Theory: Principles and Applications*. Springer Science & Business Media.

Hartig J., Höhler J. (2009) Multidimensional IRT Models for the Assessment of Competencies. *Studies in Educational Evaluation*, vol. 35, no 2–3, pp. 57–63. http://dx.doi.org/10.1016/j.stueduc.2009.10.002

Hauwaert van S.M., Schimpf C.H., Azevedo F. (2020) The Measurement of Populist Attitudes: Testing Cross-National Scales Using Item Response Theory. *Politics*, vol. 40, no 1, Article no 026339571985930. http://dx.doi.org/10.1177/0263395719859306

Johnson H.M. (1945) Are Psychophysical Problems Genuine or Spurious? *The American Journal of Psychology*, vol. 58, no 2, pp. 189–211. https://doi.org/10.2307/1417845

Kane M.T. (2016) Explicating Validity. *Assessment in Education: Principles, Policy & Practice*, vol. 23, no 2, pp. 198–211. https://doi.org/10.1080/0969594X.2015.1060192

Kunina-Habenicht O., Goldhammer F. (2020) ICT Engagement: A New Construct and Its Assessment in PISA 2015. *Large-Scale Assessments in Education*, vol. 8, no 1, pp. 1–21. http://dx.doi.org/10.1186/s40536-020-00084-z

Lange J., Dalege J., Borsboom D., van Kleef G.A., Fischer A.H. (2020) Toward an Integrative Psychometric Model of Emotions. *Perspectives on Psychological Science*, vol. 15, no 2, pp. 444–468. http://dx.doi.org/10.1177/1745691619895057

Linden van der W.J., Hambleton R.K. (eds) (2013) *Handbook of Modern Item Response Theory*. Springer Science & Business Media.

Luo Y. (2021) A Comparison of Common IRT Model-selection Methods with Mixed-Format Tests. *Measurement: Interdisciplinary Research and Perspectives*, vol. 19, no 4, pp. 199–212. http://dx.doi.org/10.1080/15366367.2021.1878779

MacCorquodale K., Meehl P.E. (1948) On a Distinction between Hypothetical Constructs and Intervening Variables. *Psychological Review*, vol. 55, no 2, pp. 95–107. https://doi.org/10.1037/h0056029

Maraun M. (2017) The Object Detection Logic of Latent Variable Technologies. *Quality and Quantity*, vol. 51, no 1, pp. 239–259. https://doi.org/10.1007/s11135-015-0303-0

Maraun M.D., Gabriel S.M. (2013) Illegitimate Concept Equating in the Partial Fusion of Construct Validation Theory and Latent Variable Modeling. *New Ideas in Psychology*, vol. 31, no 1, pp. 32–42. https://doi.org/10.1016/j.newideapsych.2011.02.006

Maraun M.D., Halpin P.F. (2008) Manifest and Latent Variates. *Measurement: Interdisciplinary Research and Perspectives*, vol. 6, no 1–2, pp. 113–117. https://doi.org/10.1080/15366360802035596

Markus K.A., Borsboom D. (2013) *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning.* New York, NY: Routledge/Taylor & Francis Group. https://doi.org/10.4324/9780203501207

Messick S. (1994) The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*, vol. 23, no 2, pp. 13–23. https://doi.org/10.3102/0013189x023002013

Michell J. (2013) Constructs, Inferences, and Mental Measurement. *New Ideas in Psychology*, vol. 31, no 1, pp. 13–21. https://doi.org/10.1016/j.newideapsych.2011.02.004

Mislevy R.J., Steinberg L.S., Almond R.G. (2002) On the Roles of Task Model Variables in Assessment Design. Generating Items for Cognitive Tests: Theory and Practice (eds S. Irvine, P. Kyllonen), Hillsdale, NY: Erlbaum, pp. 97–128.

Nering M.L., Ostini R. (eds) (2010) *Handbook of Polytomous Item Response Theory Models*. New York, NY: Routledge. https://doi.org/10.4324/9780203861264

Nima A.A., Cloninger K.M., Persson B.N., Sikström S., Garcia D. (2020) Validation of Subjective Well-Being Measures Using Item Response Theory. *Frontiers*

in Psychology, vol. 10, January, Article no 3036. http://dx.doi.org/10.3389/fpsyg.2019.03036

Ottensen J. (2000) *Mathematical Modelling in Medicine*. Amsterdam: IOS Press.

Podolsky A., Kaufman K.R., Cahalan T.D., Aleshinsky S.Y., Chao E.Y. (1990) The Relationship of Strength and Jump Height in Figure Skaters. *The American Journal of Sports Medicine*, vol. 18, no 4, pp. 400–405. https://doi.org/10.1177/036354659001800412

Power M.J. (2006) The Structure of Emotion: An Empirical Comparison of Six Models. *Cognition and Emotion,* vol. 20, no 5, pp. 694–713. https://doi.org/10.1080/02699930500367925

Pugesek B.H., Tomer A., von Eye A. (2003) *Structural Equation Modeling: Applications in Ecological and Evolutionary Biology*. Cambridge, UK: Cambridge University. https://doi.org/10.1017/CBO9780511542138

Qian M., Plucker J.A., Yang X. (2019) Is Creativity Domain Specific or Domain General? Evidence from Multilevel Explanatory Item Response Theory Models. *Thinking Skills and Creativity*, vol. 33, May, Article no 100571. http://dx.doi.org/10.1016/j.tsc.2019.100571

Oberkampf W.L., DeLand S.M., Rutherford B.M., Diegert K.V., Alvin K.F. (2002) Error and Uncertainty in Modeling and Simulation. *Reliability Engineering & System Safety*, vol. 75, no 3, pp. 333–357. http://dx.doi.org/10.1016/S0951-8320(01)00120-X

Rasch G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danmarks Paedagogiske Institut.

Ravand H., Robitzsch A. (2015) Cognitive Diagnostic Modeling Using R. *Practical Assessment, Research, and Evaluation*, vol. 20, no 11. Available at: http://pareonline.net/getvn.asp?v=20&n=11 (accessed 20 August 2023).

Reise S.P. (2012) The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research,* vol. 47, no 5, pp. 667–696. https://doi.org/10.1080/00273171.2012.715555

Riconscente M.M., Mislevy R.J., Corrigan S. (2015) Evidence-Centered Design. *Handbook of Test Development* (eds S. Lane, M.R. Raymond, T.M. Haladyna), New York, NY: Routledge, pp. 40–63. http://dx.doi.org/10.4324/9780203102961.ch3

Robitzsch A. (2022) On the Choice of the Item Response Model for Scaling PISA Data: Model Selection Based on Information Criteria and Quantifying Model Uncertainty. *Entropy*, vol. 24, no 6, Article no 760. http://dx.doi.org/10.3390/e24060760

Rhodes M., Putkaradze V. (2022) Trajectory Tracing in Figure Skating. *Nonlinear Dynamics*, vol. 110, no 4, pp. 3031–3044. https://doi.org/10.1007/s11071-022-07806-8

Schmittmann V.D., Cramer A.O.J., Waldorp L.J., Epskamp S., Kievit R.A., Borsboom D. (2013) Deconstructing the Construct: A Network Perspective on Psychological Phenomena. *New Ideas in Psychology*, vol. 31, no 1, pp. 43–53. https://doi.org/10.1016/j.newideapsych.2011.02.007

Sen S., Cohen A.S. (2019) Applications of Mixture IRT Models: A Literature Review. *Measurement: Interdisciplinary Research and Perspectives*, vol. 17, no 4, pp. 177–191. http://dx.doi.org/10.1080/15366367.2019.1583506

Shaw A., Kapnek M., Morelli N.A. (2021) Measuring Creative Self-Efficacy: An Item Response Theory Analysis of the Creative Self-Efficacy Scale. *Frontiers in Psychology*, vol. 12, July, Article no 678033. http://dx.doi.org/10.3389/fpsyg.2021.678033

Sijtsma K., Ark van der A. (2020) *Measurement Models for Psychological Attributes: Classical Test Theory, Factor Analysis, Item Response Theory, and Latent Class Models*. Boca Raton, FL: CRC. https://doi.org/10.1201/9780429112447

Streckert N., Kurtz L., Kajonius P.J. (2023) Can Your Darkness Be Measured? Analyzing the Full and Brief Version of the Dark Factor of Personality in Swed-

ish. *International Journal of Testing*, vol. 23, no 2, pp. 1–45. http://dx.doi.org/10.1080/15305058.2023.2195659

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

Trendler G. (2022) Is Measurement in Psychology an Empirical or a Conceptual Issue? A Comment on David Franz. *Theory & Psychology*, vol. 32, no 1, pp. 164–170. https://doi.org/10.1177/09593543211050025

Trendler G. (2013) Measurement in Psychology: A Case of Ignoramus et Ignorabimus? A Rejoinder. *Theory & Psychology*, vol. 23, no 5, pp. 591–615. https://doi.org/10.1177/0959354313490451

Tynan M.C. (2021) Deconstructing Grit's Validity: The Case for Revising Grit Measures and Theory. *Multidisciplinary Perspectives on Grit: Contemporary Theories, Assessments, Applications and Critiques* (eds L.E. van Zyl, C. Olckers, L. van der Vaart), Cham: Springer Nature Switzerland, pp. 137–155. http://dx.doi.org/10.1007/978-3-030-57389-8_8

Tyumeneva Y., Kardanova E., Kuzmina J. (2019) Grit: Two Related but Independent Constructs Instead of One. Evidence from Item Response Theory. *European Journal of Psychological Assessment*, vol. 35, no 4, pp. 469–478. http://dx.doi.org/10.1027/1015-5759/a000424

Uglanova I.L., Brun I.V., Vasin G.M. (2018) Metodologiya Evidence-Centered Design dlya izmereniya kompleksnykh psikhologicheskikh konstruktov [Evidence-Centered Design Method for Measuring Complex Psychological Constructs]. *Journal of Modern Foreign* Psychology, vol. 7, no 3, pp. 18–27. https://doi.org/10.17759/jmfp.2018070302

Uher J. (2021) Quantitative Psychology under Scrutiny: Measurement Requires Not Result-Dependent But Traceable Data Generation. *Personality and Individual Differences*, vol. 170, no 5, Article no 110205. https://doi.org/10.1016/j.paid.2020.110205

Vessonen E. (2021) Conceptual Engineering and Operationalism in Psychology. *Synthese*, vol. 199, no 3–4, pp. 10615–10637. https://doi.org/10.1007/s11229-021-03261-x

Wagner T.A., Harvey R.J. (2006) Development of a New Critical Thinking Test Using Item Response Theory. *Psychological Assessment,* vol. 18, no 1, pp. 100–105. https://doi.org/10.1037/1040-3590.18.1.100

Walton K.E., Roberts B.W., Krueger R.F., Blonigen D.M., Hicks B.M. (2008) Capturing Abnormal Personality with Normal Personality Inventories: An Item Response Theory Approach. *Journal of Personality*, vol. 76, no 6, pp. 1623–1648. http://dx.doi.org/10.1111/j.1467-6494.2008.00533.x

Wiggins B.J., Christopherson C.D. (2019) The Replication Crisis in Psychology: An Overview for Theoretical and Philosophical Psychology. *Journal of Theoretical and Philosophical Psychology*, vol. 39, no 4, pp. 202–217. http://dx.doi.org/10.1037/teo0000137

Will C.M. (2000) *Einstein's Relativity and Everyday Life.* Available at: http://www.physicscentral.com/writers/writers-00-2.html (accessed 20 August 2023).

Wilson M. (2004) *Constructing Measures. An Item Response Modeling Approach*. New York, NY: Routledge.

Yen W.M., Fizpatrick A.R. (2006) Item Response Theory. *Educational Measurement* (ed. R.L. Brennan), Westport, CT: American Council on Education and Praeger, pp. 17–64.

Zhao H., Alexander P.A., Sun Y. (2021) Relational Reasoning's Contributions to Mathematical Thinking and Performance in Chinese Elementary and Middle-School Students. *Journal of Educational Psychology,* vol. 113, no 2, pp. 279–303. http://dx.doi.org/10.1037/edu0000595

Вычислительная психометрика: ближайшее будущее или уже реальность

Рецензия на книгу "Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment"[1]

Ксения Тарасова, Дарья Грачева

Статья поступила в редакцию в мае 2023 г.

Тарасова Ксения Вадимовна — кандидат педагогических наук, заместитель руководителя лаборатории измерения новых конструктов и дизайна тестов Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики». Адрес: 101000, Москва, Потаповский пер., 16, стр. 10. E-mail: ktarasova@hse.ru. ORCID: https://orcid.org/0000-0002-3915-3165 (контактное лицо для переписки)

Грачева Дарья Александровна — младший научный сотрудник Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики». E-mail: dgracheva@hse.ru. ORCID: https://orcid.org/0000-0002-4646-7349

Аннотация

Рецензируемое издание продолжает серию книг о методологии образовательного тестирования и оценивания, написанных ведущими психометриками и исследователями в области оценивания в образовании. Вычислительная психометрика рассматривается как сочетание методов вычислительных наук и психометрических принципов измерений для анализа данных, полученных в результате тестирования с использованием технологически усовершенствованных форматов теста. В первой части книги обсуждаются изменения, которые произошли в обучении и образовательном оценивании под влиянием цифровых технологий. Во второй части представлен обзор методов вычислительной психометрики: от традиционных психометрических моделей до технологий машинного обучения.

Материалы книги могут быть полезны студентам и исследователям в области психометрики, которые занимаются разработкой, проектированием, анализом систем обучения и измерениями с использованием сложных тестовых форматов и данных. Сильной стороной книги является электронное приложение, содержащие код среды программирования *R* или *Python* для методологических глав.

Ключевые слова

вычислительная психометрика, машинное обучение, образовательное оценивание, Evidence-Centered Design

[1] Davier von A.A., Mislevy R.J., Hao J. (eds) (2022) Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment. With Examples in R and Python. Cham: Springer Nature.

Для цитирования Тарасова К.В., Грачева Д.А. (2023) Вычислительная психометрика: ближайшее будущее или уже реальность. Рецензия на книгу “Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment“. *Вопросы образования / Educational Studies Moscow*, № 3, сс. 221–230. https://doi.org/10.17323/vo-2023-17938

Computational Psychometrics: Near Future or Reality

Review of the book “Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment”

Ksenia Tarasova, Daria Gracheva

Ksenia V. Tarasova — Candidate of Sciences in Education, Deputy Head of the Laboratory for Measuring New Constructs and Test Design, Centre for Psychometrics and Measurement in Education, National Research University Higher School of Economics. Address: Bld. 10, 16 Potapovsky Ln, 101000 Moscow, Russian Federation. E-mail: ktarasova@hse.ru. ORCID: https://orcid.org/0000-0002-3915-3165 (corresponding author)

Daria A. Gracheva — Junior Research Fellow at the Center for Psychometrics and Measurements in Education, Institute of Education, National Research University Higher School of Economics. E-mail: dgracheva@hse.ru. ORCID: https://orcid.org/0000-0002-4646-7349

Abstract This book continues a series of books on the methodology of educational testing and assessment written by leading psychometricians and researchers in the field of educational assessment. Computational psychometrics is defined as the combination of computer science methods and psychometric measurement principles for analysing data obtained as a result of testing using technologically advanced test formats. The first part of the book discusses the changes that have occurred in teaching and educational assessment under the influence of digital technologies. The second part provides an overview of computational psychometric methods: from traditional psychometric models to machine learning technologies. The material in the book can be useful to students and researchers in the field of psychometrics who are involved in the development, design and analysis of learning systems and measurements using complex test formats and data. The strength of the book is an electronic application containing the code of the R or Python programming environment for the methodological chapters.

Keywords computational psychometrics, machine learning, educational assessment, evidence-centered design

For citing Tarasova K.V., Gracheva D.A. (2023) Vychislitel´naya psikhometrika: blizhajshee budushchee ili uzhe real´nost´. Retsenziya na knigu “Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment” [Computational Psychometrics: Near Future or Reality. Review of the book “Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment”]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 3, pp. 221–230. https://doi.org/10.17323/vo-2023-17938

Психометрика возникла более века назад, когда для обоснования выводов о психологических свойствах и процессах стали применять количественные методы обработки данных, полученных в ходе наблюдения. В широком смысле содержание психометрики с тех пор не изменилось, несмотря на прогресс статистических теорий, моделей и методов.

Современные условия, в частности большой объем данных, получаемых в процессе тестирования, ставят перед исследователями в области психометрики новые задачи по их обработке, анализу и интерпретации. Среди актуальных вызовов — повышение точности и валидности измерений, выявление паттернов и стилей ответов, определение новых конструктов и предоставление качественной обратной связи участникам тестирования и всем заинтересованным сторонам.

После нескольких лет обсуждений специфики новых методологий в тестировании исследователи обратили внимание на то, что они, несмотря на все различия, имеют одну общую черту — использование вычислительных моделей. Так родился лаконичный термин — вычислительная психометрика, введенный Алиной фон Давье на международной конференции по машинному обучению [Davier von, 2015].

Вычислительная психометрика стремится интегрировать методы науки о данных с методами машинного обучения в психометрике, придерживаясь установленных принципов измерений в психологии и образовании.

В книге представлены основы вычислительной психометрики: в первой части обсуждаются изменения в образовательном ландшафте, обусловившие переход к следующему поколению методов обучения и оценивания, а также отдельные технологии, способные трансформировать образовательный опыт; во второй части описаны конкретные методы как актуальные иллюстрации возможностей вычислительной психометрики.

Обучение и оценивание в новых условиях

В написанном Р. Мислеви разделе книги, посвященном методологии вычислительной психометрики, представлен ситуативный, социокогнитивный взгляд на природу человеческих способностей, на их развитие и на способы их использования человеком для взаимодействия с социальным и физическим миром [Greeno, 1998]. Эта исследовательская перспектива создает запрос на широкий спектр типов оценки, обычно включающих взаимодействие со средой, часто в цифровых контекстах, а иногда требующих сотрудничества между людьми. Именно на основе оценки действий испытуемых в таких средах можно получить прямые свидетельства способностей, необходимых для формирования исследовательских навыков, умения решать проблемы, коммуникативных компетенций и т.п.

Вычислительная психометрика не просто предлагает совокупность методов, но может служить основой виртуальной среды обучения и оценки, такой как игры, симуляции, совместная работа и дополненные среды. В этих средах возможно получение большого количества данных сложной структуры.

В разделе, посвященном специфике вычислительной психометрики, А. фон Давье и ее соавторы подчеркивают, что сбор данных в цифровых средах должен быть целенаправленным в максимально возможной степени. Поэтому важно пересмотреть весь процесс разработки инструмента оценивания. Инструменты, ориентированные на доказательную базу или доказательный дизайн (*evidence-centered design*, ECD) [Mislevy, Almond, Lukas, 2003; Arieli-Attali et al., 2019], создаются с таким расчетом, чтобы обеспечить возможность фиксации и учета каждого элемента процесса выработки ответа или решения. В частности, дизайн таких инструментов включает концепции модели учащегося, модели задания и модели доказательств. Модель учащегося определяет конструкты, о которых мы хотим сделать выводы. Модель задания предусматривает виды деятельности, которые позволят получить доказательства: это могут быть вопросы с множественным выбором, симуляция, игра или любая другая деятельность, в которой может участвовать испытуемый. В модели доказательств особое внимание уделяется их идентификации (извлечение ключевых элементов из продукта деятельности, созданного учащимся в результате тестирования) и агрегированию (аккумуляция информации в балл с помощью статистической модели). При анализе данных акцент смещается на то, как информация, полученная в результате выполнения задачи, обобщается для составления суждений о навыках учащихся, которые могут проявиться в разных контекстах. Эта перспектива особенно актуальна для оценивания высокоуровневых навыков.

Тестирование с использованием цифровых технологий позволяет разрабатывать насыщенные и технологически усовершенствованные типы заданий, которые лучше отражают ситуации из реальной жизни и позволяют измерять сложные навыки, обеспечивая тем самым более глубокое оценивание способностей учащихся. Эволюция от технологически усовершенствованных заданий к технологически усовершенствованному оцениванию происходит в несколько этапов [DiCerbo, Behrens, 2012]. На первом этапе технологии используются для создания новых типов элементов, механик и улучшения обратной связи; на втором задание строится на основе центрального контекста — проблемы, которую необходимо решить (симуляционные задания); на третьем этапе создаются естественные цифровые среды, разрабатываемые с учетом аргумента оце-

нивания (использование скрытого оценивания в игре) [Shute, Ventura, 2013]; четвертый этап представляет собой экосистему, в которой накапливается информация из разных естественных цифровых сред, подобных созданным на третьем уровне.

Отдельный раздел в книге посвящен виртуальным тестовым средам (*virtual performance-based assessments*, VPBAs), в которых участники тестирования взаимодействуют с системами, иногда включающими других людей или агентов, чтобы представить доказательства своих знаний, умений или других качеств. Сбор цифровых данных позволяет получить подробную информацию о действиях испытуемых и развитии ситуаций, в которых эти действия происходят. При этом перед исследователем стоит задача сконструировать VPBA, способную инициировать проявление целевых способностей участников тестирования, дать им возможность продемонстрировать эти способности, зафиксировать соответствующие аспекты выполнения заданий, выявить значимые закономерности, которые являются доказательством целевых способностей, и предоставить основу для синтеза доказательств и описания их свойств.

К категории виртуальных тестовых сред современные исследователи относят: задания сценарного типа (*scenario-based assessments*), симуляционные задания *(simulation-based assessments),* оценивание с использованием заданий игрового типа (*game-based assessments),* оценивание совместной работы *(collaborative assessments).*

Оценка с помощью заданий сценарного типа актуализирует опыт испытуемого повествовательным контекстом с целью добавления слоя смысла к мотивам, мыслям и действиям. В самом простом варианте непрерывный контекст (развитие сюжета) вводится для последовательности заданий, которые часто имеют традиционный формат, например множественный выбор или краткий ответ. Более сложные задания сценарного типа могут содержать элементы симуляции и игр.

В симуляционных заданиях воспроизводятся обстоятельства, характерные для тех или иных видов реальной деятельности: например, проведение научного эксперимента, оказание медицинской помощи или полет на реактивном истребителе. Такой дизайн тестирования позволяет оценить способности индивидов применительно к конкретной сфере занятости [Davey et al., 2015]. Симуляционные задания выполняют очень важную функцию: они дают возможность людям продемонстрировать свои знания, умения и навыки в таких ситуациях, которые могут быть слишком опасными для того, чтобы проводить испытания в реальных условиях, или организация таких испытаний обходится слишком дорого либо затратна по времени [Snir, Smith, Grosslight, 1993].

Оценивание с использованием заданий игрового типа или основанное на игре объединяет принципы игрового дизайна с принципами дизайна заданий для измерения знаний, умений и навыков. Такие задания могут повысить мотивацию тестируемых за счет своей увлекательности. При разработке заданий игрового типа важно найти баланс между игровыми элементами и характеристиками, такими как увлекательность и вовлеченность, с одной стороны, и элементами оценивания — с другой, т.е. обеспечить соответствие игры психометрическим требованиям, такими как валидность, надежность и справедливость. Только так можно достичь заданной цели и предоставить испытуемому возможность продемонстрировать свои способности [Kim, Shute, 2015]. Создаваемые разработчиками игр сложные интерактивные среды подходят для исследования таких комплексных конструктов, как системное мышление и межкультурная компетентность.

Оценивание совместной работы — востребованная тестовая среда. Поскольку в современном мире планирование деятельности, принятие решений и урегулирование проблем по большей части осуществляется в составе команд, навыки, необходимые для достижения успеха в групповой работе, становятся важным критерием оценки персонала. В рамках оценивания совместной работы тестируемый получает возможность проявить свои умения эффективно работать или общаться с другими людьми. Для этого создается среда, которая обеспечивает его общение и сотрудничество с одним или несколькими лицами. Это общение может быть асинхронным, как, например, потоковое онлайн-обсуждение в системе управления обучением или на массовых открытых онлайн-курсах (MOOK) [Rosé, Ferschke, 2016; Wise, Chiu, 2011], или синхронным, когда люди общаются в режиме реального времени через чат или аудио- и видеоканалы [Andrews-Todd, Forsyth, 2020; Bower, Lee, Dalgarno, 2017]. Возможны среды, поддерживающие общение между человеком и одним или несколькими компьютерными агентами [Biswas et al., 2010; Graesser et al., 2012], взаимодействие двух людей с искусственными агентами [Liu et al., 2015]. В виртуальной среде можно фиксировать все действия и рассуждения членов команды, а не только ответы, которые они дают. По результатам такого оценивания можно судить не только о стратегиях, которые члены команды используют во время работы для решения проблемы, но и о характере их взаимодействия.

В последние годы новые возможности проектирования, сбора и анализа данных используются не только для создания цифровых сред оценки, но и для персонализации обучения на основе данных об учебном опыте студентов. Один из разделов книги посвящен обзору моделей адаптивного обу-

чения и методов интеллектуального анализа данных. Системы онлайн-обучения, использующие данные о характеристиках и успеваемости учащихся для подстраивания учебных программ к их потребностям, становятся адаптивными образовательными средами. Они учитывают широкий спектр характеристик студентов и конструктов, влияющих на процесс обучения, — их знания, эмоции, поведенческую активность и мотивацию. Успешные адаптивные системы обучения могут использовать множество данных о деятельности учащегося в режиме реального времени, чтобы делать выводы, на основе которых учащемуся автоматически предоставляется необходимая для достижения успеха индивидуализированная поддержка.

Методы вычислительной психометрики

Во второй части книги рассматриваются методы вычислительной психометрики и возможности их применения в новой реальности: для сложных тестовых форматов и данных.

Вводный раздел посвящен фундаментальным идеям и моделям психометрики. В широком смысле модели психометрики связывают наблюдаемые данные тестирования с ненаблюдаемыми (латентными) характеристиками респондентов. В качестве наиболее распространенных сегодня психометрических моделей Р. Мислеви и М. Большова упоминают классическую теорию тестирования и ее расширения, в частности теорию генерализации, непрерывные и дискретные модели современной теории тестирования, а также эксплораторный и конфирматорный факторный анализ.

Развитие цифровых технологий и появление интерактивных заданий, реализуемых в насыщенной тестовой среде, предъявляет дополнительные требования к устоявшимся психометрическим практикам. В частности, на первый план выходят проблемы локальной зависимости заданий, многомерности оцениваемых конструктов, больших объемов данных о процессе выполнения заданий (*process data*) и оценки навыков высокого порядка.

Перспективным для оценки параметров моделей в новых реалиях авторам представляется байесовский подход (*Bayesian inference*) с применением алгоритма Монте-Карло по схеме цепи Маркова (*Markov chain Monte Carlo*, MCMC), ему посвящен отдельный раздел книги. Авторы также высоко оценивают возможности в современных условиях таких психометрических практик, как моделирование иерархических структур данных (*hierarchical models*), компьютерное адаптивное тестирование на основе IRT (*computer adaptive testing*, CAT IRT), внедрение коллатеральной информации о респондентах и заданиях с опорой на когнитивные теории.

Одним из активно развивающихся направлений сегодня является сетевая психометрика [Marsman et al., 2018; Epskamp et al., 2018], которая предлагает альтернативное понимание психологических конструктов. В сетевой психометрике связи между наблюдаемыми переменными, входящими в сеть (конструкт), считаются корреляционными и при отсутствии общей причины для их значений, в то время как в традиционных психометрических моделях корреляционные отношения принято объяснять наличием общей причины — латентной переменной. Сетевые модели представляют структуру психологических конструктов в виде сложных взаимосвязей между психологическими, биологическими, социологическими и другими переменными без дополнительных допущений о причинности в отношениях между наблюдаемыми переменными, которыми оперирует традиционная психометрика.

Переход к цифровым инструментам оценивания сопровождается увеличением объема данных о респонденте. Традиционная психометрика основывается на конкретных и структурированных данных, например на баллах за выполнение заданий, в которых респондент поставлен перед ограниченным выбором из нескольких вариантов ответа. Вычислительная психометрика часто имеет дело с данными «на микроуровне» — с детализированными и неструктурированными сведениями о действиях тестируемых в процессе выполнения задания, например о движениях мыши, нажатии на клавиши, кликах. В состав этих данных могут входить нейрофизиологические показатели, например результаты фиксации взгляда.

Задача вычислительной психометрики состоит в том, чтобы на основе микроданных сформулировать содержательный вывод о респонденте. Эффективными инструментами для обнаружения закономерностей в микроданных служат подходы, сложившиеся в рамках интеллектуального анализа данных (*data mining*), учебной аналитики (*learning analytics*), машинного обучения. В рецензируемом издании описаны два основных направления машинного обучения — обучение с учителем (*supervised machine learning*) и обучение без учителя (*unsupervised machine learning*). Данные о процессе, т.е. о последовательности действий внутри задания, также могут быть проанализированы с использованием моделей временных рядов или социальных сетей (*social network analysis*).

Книга завершается обзором использования моделей обработки естественного языка (*natural language processing*, NLP) в области оценивания. Авторы рассматривают задачи интеллектуального анализа текстов (*text mining*) и автоматического скоринга заданий открытого типа (*automatic scoring*) с коротким и длинным (эссе) конструированным ответом.

Заключение

Результатом использования цифровых форматов оценивания становится большой объем данных о поведении респондента в тестовой среде. Для анализа таких данных требуются специальные методы, расширяющие возможности традиционных психометрических моделей. Следуя этому тренду, *EdTech*-компании начали применять к результатам тестирований продвинутые методы наук о данных, однако бизнес-сообщество уделяет мало внимания фундаментальным принципам измерений — валидности, надежности, справедливости оценивания, что отражается на качестве выводов о тестируемых. Вычислительная психометрика сочетает применение продвинутых методов вычислительных наук с соблюдением психометрических принципов измерений при анализе сложных данных. Рецензируемая книга дает представление об основных принципах измерений, которые должны соблюдаться на протяжении всего тестирования — от разработки заданий до представления результатов. Один из разделов книги посвящен обзору методов вычислительной психометрики с примерами кода среды программирования *R* или *Python*. Многие из этих методов еще не вошли в повседневный инструментарий психометрика, а навыки, необходимые для моделирования сложных данных, как правило, недостаточно освещаются в большинстве образовательных программ по измерениям. Таким образом, книга знакомит широкий круг читателей с идеями и методами вычислительной психометрики, которые способны обогатить образовательное оценивание и сделать более точными выводы в отношении каждого респондента и системы обучения и образования в целом.

Referenses

Andrews-Todd J., Forsyth C.M. (2020) Exploring Social and Cognitive Dimensions of Collaborative Problem Solving in an Open Online Simulation-Based Task. *Computers in Human Behavior*, vol. 104, January, Article no 105759. https://doi.org/10.1016/j.chb.2018.10.025

Arieli-Attali M., Ward S., Thomas J., Deonovic B., von Davier A.A. (2019) The Expanded Evidence-Centered Design (E-ECD) for Learning and Assessment Systems: A Framework for Incorporating Learning Goals and Processes within Assessment Design. *Frontiers in Psychology*, vol. 10, April, Article no 853. https://doi.org/10.3389/fpsyg.2019.00853

Biswas G., Jeong H., Kinnebrew J.S., Sulcer B., Roscoe R. (2010) Measuring Self-Regulated Learning Skills through Social Interactions in a Teachable Agent Environment. *Research and Practice in Technology Enhanced Learning*, vol. 5, no 2, pp. 123–152. http://dx.doi.org/10.1142/S1793206810000839

Bower M., Lee M.J., Dalgarno B. (2017) Collaborative Learning across Physical and Virtual Worlds: Factors Supporting and Constraining Learners in a Blended Reality Environment. *British Journal of Educational Technology*, vol. 48, no 2, pp. 407–430. http://dx.doi.org/10.1111/bjet.12435

Davey T., Ferrara S., Shavelson R., Holland P., Webb N., Wise L. (2015) *Psychometric Considerations for the Next Generation of Performance Assessment. Report of the Center for K-12 Assessment & Performance Management.* Available at: https://

www.ets.org/Media/Research/pdf/psychometric_considerations_white_paper.pdf (accessed 20 August 2023).

Davier von A.A. (2015) *Virtual and Collaborative Assessments: Examples, Implications, and Challenges for Educational Measurement*. Paper presented at the 32nd International Conference on Machine Learning (ICML 2015) (Lille, France, 2015, July 6–11).

DiCerbo K.E., Behrens J.T. (2012) Implications of the Digital Ocean on Current and Future Assessment. *Computers and Their Impact on State Assessment: Recent History and Predictions for the Future* (eds R. Lissitz, H. Jiao), Charlotte, NC: Information Age Publishing, pp. 273–306.

Epskamp S., Maris G., Waldorp L.J., Borsboom D. (2018) Network Psychometrics. *The Wiley Handbook of Psychometric Testing* (eds P. Irwing, D. Hughes, T. Booth), New York, NY: Elsevier.

Graesser A.C., D'Mello S., Hu X., Cai Z., Olney A., Morgan B. (2012) AutoTutor. *Applied Natural Language Processing: Identification, Investigation and Resolution* (eds P. McCarthy, C. Boonthum-Denecke), Hershey, PA: IGI Global, pp. 169–187. http://dx.doi.org/10.4018/978-1-60960-741-8

Greeno J.G. (1998) The Situativity of Knowing, Learning, and Research. *American Psychologist*, vol. 53, no 1, pp. 5–26. https://doi.org/10.1037/0003-066X.53.1.5

Kim Y.J., Shute V.J. (2015) The Interplay of Game Elements with Psychometric Qualities, Learning, and Enjoyment in Game-Based Assessment. *Computers & Education*, vol. 87, no 2, pp. 340–356. http://dx.doi.org/10.1016/j.compedu.2015.07.009

Liu L., von Davier A.A., Hao J., Kyllonen P., Zapata-Rivera J.-D. (2015) A Tough Nut to Crack: Measuring Collaborative Problem Solving. *Handbook of Research on Computational Tools for Real-World Skill Development* (eds Y. Rosen, S. Ferrara, M. Mosharraf), Hershey, PA: IGI-Global, pp. 344–359.

Marsman M., Borsboom D., Kruis J., Epskamp S., van Bork R., Waldorp L. et al. (2018) An Introduction to Network Psychometrics: Relating Ising Network Models to Item Response Theory Models. *Multivariate Behavioral Research*, vol. 53, no 1, pp. 15–35. http://dx.doi.org/10.1080/00273171.2017.1379379

Mislevy R.J., Almond R.G., Lukas J.F. (2003) *A Brief Introduction to Evidence-Centered Design. Center for the Study of Evaluation Report no 632*. Los Angeles, CA: University of California. http://dx.doi.org/10.1002/j.2333-8504.2003.tb01908.x

Rosé C.P., Ferschke O. (2016) Technology Support for Discussion Based Learning: From Computer Supported Collaborative Learning to the Future of Massive Open Online Courses. *International Journal of Artificial Intelligence in Education*, vol. 26, no 2, pp. 660–678. http://dx.doi.org/10.1007/s40593-016-0107-y

Shute V.J., Ventura M. (2013) *Stealth Assessment: Measuring and Supporting Learning in Video Games*. Cambridge, MA: MIT. https://doi.org/10.7551/mitpress%2F9589.001.0001

Snir J., Smith C., Grosslight L. (1993) Conceptually Enhanced Simulations: A Computer Tool for Science Teaching. *Journal of Science Education and Technology*, vol. 2, no 2, pp. 373–388. https://doi.org/10.1007/BF00694526

Wise A.F., Chiu M.M. (2011) Analyzing Temporal Patterns of Knowledge Construction in a Role-Based Online Discussion. *International Journal of Computer-Supported Collaborative Learning*, vol. 6, no 3, pp. 445–470. http://dx.doi.org/10.1007/s11412-011-9120-1

Сила вероятности в психометрике

Рецензия на книгу "Bayesian Psychometric Modelling"[1]

Ирина Угланова

Статья поступила в редакцию в мае 2023 г.

Угланова Ирина Львовна — научный сотрудник Лаборатории измерения новых конструктов и дизайна тестов Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики». Адрес: 101000, Москва, Потаповский пер., 16, стр. 10. E-mail: iluglanova@gmail.com. ORCID: https://orcid.org/0000-0001-9117-5997

Аннотация

Возникновение и развитие байесовской психометрики — закономерный результат стремления психометрики уловить и уменьшить ошибку измерения. В рецензируемой книге впервые представлено систематическое описание байесовского подхода в психометрических исследованиях. Книга состоит из двух частей: в первой изложены основные положения байесовского подхода (*Foundations*), во второй — их применение в психометрическом моделировании (*Psychometrics*). Автор этой рецензии считает, что издание будет полезно тем, кто работает в частотном подходе и хотел бы узнать про байесовский. При этом тем, кто не уверен в надежности своего математического бэкграунда, автор рецензии рекомендует обращаться к дополнительным источникам — не оснащенным таким детальным математическим сопровождением. Книга не переведена на русский язык.

Ключевые слова

психометрика, байесовский подход, байесовская статистика, рецензия

Для цитирования

Угланова И.Л. (2023) Сила вероятности в психометрике. Рецензия на книгу "Bayesian Psychometric Modelling". *Вопросы образования / Educational Studies Moscow*, № 3, сс. 231–236. https://doi.org/10.17323/vo-2023-17952

Power of Probability in Psychometrics

Review of the book "Bayesian Psychometric Modeling"

Irina Uglanova

Irina L. Uglanova — Researcher at the Laboratory of Measurement of New Constructs and Test Design, Centre for Psychometrics and Measurement in Education, Institute of Education, National Research University Higher School of Economics. Address: Bld. 10, 16 Potapovsky Ln, 101000 Moscow, Russian Federation. E-mail: iluglanova@gmail.com. ORCID: https://orcid.org/0000-0001-9117-5997

[1] Levy R., Mislevy R.J. (2016) Bayesian Psychometric Modeling. Boca Raton, FL: Chapman and Hall/CRC. https://doi.org/10.1201/9781315374604

Abstract The emergence and development of Bayesian psychometrics is a result of psychometrics' desire to reduce measurement error. This book is the first to present a systematic description of the Bayesian approach in psychometric research. The book consists of two parts: the first one presents the main principles of the Bayesian approach (Foundations), the second one includes their application in psychometric modeling (Psychometrics). The reviewer believes that the publication will be useful for those who used to work in the frequentist approach and would like to learn about the Bayesian approach. At the same time, she recommends those who are not sure of the quality of their mathematical background to additionally turn to other sources that are not equipped with such a detailed mathematical description. The book has not been translated into Russian.

Keywords psychometrics, Bayesian approach, Bayesian statistics, review

For citing Uglanova I.L. (2023) Sila veroyatnosti v psikhometrike. Retsenziya na knigu "Bayesian Psychometric Modelling" [Power of Probability in Psychometrics. Review of the book "Bayesian Psychometric Modeling"]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 3, pp. 231–236. https://doi.org/10.17323/vo-2023-17952

«Значительная часть этой книги посвящена тому, как применение вероятностного подхода к психометрическому моделированию позволяет не только уловить разрыв между [латентными] способностями и [наблюдаемыми] результатами, но и преодолеть его»[2]

Для современной психометрики интерпретировать результаты работы с данными в терминах вероятности — привычный и устоявшийся подход. Он проявляется и в том, как описывается качество теста, и в формулировках обратной связи для участников тестирования. В психометрических исследованиях использование вероятностного подхода идет рука об руку с необходимостью учитывать ошибку измерения, т.е. разрыв между наблюдаемым поведением и той характеристикой, которую это наблюдаемое поведение должно было бы отражать. Не будет сильным упрощением сказать, что развитие психометрики — это осмысление ошибки измерения и усилия, направленные на ее сокращение.

С этой точки зрения возникновение и развитие байесовской психометрики оказывается ожидаемым, неизбежным и очень перспективным. Байесовское психометрическое моделирование объединяет принципы байесовской статистики с психометрическими моделями. Байесовская статистика уже активно

[2] "Much of this book is devoted to illuminating how adopting a probabilistic approach to psychometric modeling aids not only in representing the disconnect between [latent] proficiency and [observed] performance, but also in bridging that divide" (Levy R., Mislevy R.J. (2016) Bayesian Psychometric Modeling. Boca Raton, FL: Chapman and Hall/CRC. P. 9).

используется в машинном обучении и подарила нам возможность пользоваться «умными», т.е. самообучающимися, нейросетями. Использование байесовского подхода в психологии и образовании только набирает популярность [König, van de Schoot, 2018].

Байесовский подход часто рассматривается в противовес более привычному — частотному. Между ними существуют три базовых различия. Во-первых, в частотном подходе мы делаем вывод о вероятности наступления события только на основе собранных данных. Казалось бы, как может быть по-другому? Байесовский подход позволяет использовать не только уже существующие данные, но и наши предположения о том, как могут вести себя данные в будущем или что характерно для оцениваемого параметра в целом, вне зависимости от того, какая выборка нам попалась. Иными словами, байесовский подход позволяет использовать априорную информацию.

Во-вторых, в частотном подходе мы оцениваем интересующий нас параметр как некое конкретное значение, возможно, с ошибкой оценки параметра (*confidence interval*). А в байесовском подходе интересующий исследователя параметр рассматривается как случайная величина, имеющая свое (апостериорное) распределение вероятности. В этом распределении можно выделить некое значение, например среднее, и определенный интервал для дальнейшей интерпретации (*credibility interval*).

В-третьих, существование распределения вероятности в байесовском подходе позволяет отойти от оценки статистической значимости и проверять гипотезу на основе апостериорного распределения вероятности.

Применение байесовской статистики распространялось в исследовательском поле по мере роста компьютерных мощностей: ограничениями в повсеместном использовании байесовского подхода выступают продолжительность расчетов и необходимость наличия больших вычислительных мощностей. В 1995 г. вышло первое издание книги Э. Гелмана и его коллег *"Bayesian Data Analysis"*, а в 2013 г. опубликовано ее третье расширенное издание [Gelman et al., 2013]. В ней приведены концептуальные постулаты байесовского подхода, его математические обоснования, а также примеры использования в разных областях[3]. Однако Э. Гелман с коллегами не рассматривают вопросы, специфические для психометрических исследований.

В 2010 г. увидела свет книга Ж.-П. Фокса *"Bayesian Item Response Modeling. Statistics for Social and Behavioral Sciences"* [Fox, 2010]. В ней рассмотрены базовые и передовые психометри-

[3] http://www.stat.columbia.edu/~gelman/book/ Текст книги и сопроводительные материалы доступны по ссылке для некоммерческого использования.

ческие модели современной теории тестирования в преломлении к байесовской парадигме анализа данных[4]. Существуют материалы, специфичные и для других психометрических подходов, например для структурного моделирования[5] [Muthén, Asparouhov, 2012].

В рецензируемой книге Р. Леви и Р. Мислеви *"Bayesian Psychometric Modeling"* впервые представлено систематическое описание байесовского подхода в психометрических исследованиях. Издание содержит основные положения байесовской статистики как на уровне концептуального осмысления вероятностного вывода, так и на уровне математических формул. При этом баланс между концептуальными рассуждениями, иллюстративными примерами и строгим статистическим описанием сохраняется для всех представленных методов психометрического анализа.

Книга состоит из двух частей: основные положения байесовского подхода (*Foundations*) и их применение в психометрическом моделировании на примере разных методов (*Psychometrics*).

Первая часть включает шесть крупных разделов. Первые три раздела посвящены введению в оценивание и байесовскую вероятность. В частности, описана методология доказательной аргументации в разработке тестов (*evidence-centered design*) и ее связь с вероятностным моделированием. Кроме того, рассматриваются особенности байесовского вывода, ключевые формы распределений, а также возможные проблемы и подходы к их решению, например при использовании априорных вероятностей.

Четвертый, пятый и шестой разделы первой части книги посвящены подготовке к работе с реальными данными. Так, рассматривается моделирование переменной с нормальным распределением и описывается использующийся при этом метод оценки параметров — алгоритм Монте-Карло по схеме марковской цепи (*Markov Chain Monte Carlo*, MCMC). Кроме этого, подробно представлено байесовское моделирование линейной регрессии с непрерывной зависимой переменной, которое служит базой для более сложных психометрических моделей.

Вторая часть книги посвящена разным психометрическим моделям в рамках байесовского подхода. Авторы начинают с описания байесовского моделирования классической теории тестирования (*classical test theory*), а затем переходят к представлению байесовского конфирматорного факторного анализа

[4] https://www.jean-paulfox.com/ сопроводительные материалы доступны по ссылке для некоммерческого использования.

[5] https://mc-stan.org/users/documentation/case-studies/sem.html

(*confirmatory factor analysis*) и современной теории тестирования (*item response theory*), включая анализ заданий с дихотомическим и политомическим начислением баллов, многомерные модели. Далее авторы рассматривают моделирование латентных классов (*latent class analysis*) с примером для дихотомических латентных и наблюдаемых переменных, а также довольно молодой метод анализа данных в психологии и образовании — моделирование байесовскими сетями (*Bayesian networks*).

Большое внимание в книге уделяется анализу качества модели. Этой теме посвящена отдельная глава (*Model Evaluation*), а также разделы (*Model-Data Fit*) в каждой главе, описывающей психометрические модели. Авторы рассмотрели и еще одну важную проблему психометрики — работу с пропущенными данными (*missing data modeling*). Байесовский подход служит опорой в ее решении не только в психометрике, но и при анализе данных в социальных науках в целом.

Сильная сторона книги — ее направленность на практическое применение. Авторы подробно описывают, как реализовать байесовские психометрические модели с помощью бесплатного программного пакета WinBUGS[6]. Этот пакет и схожая с ним система OpenBUGS предоставляют мощный аппарат для получения апостериорных вероятностей, однако при их использовании возможности последующего анализа данных ограничены. Более удобен пакет R2WinBUGS [Sturtz, Ligges, Gelman, 2005]: он позволяет применять WinBUGS через привычный интерфейс программы R[7], а также использовать возможности R для последующего анализа данных.

Книгу Р. Леви и Р. Мислеви *"Bayesian Psychometric Modeling"* можно использовать как учебник: в конце каждой главы приведены несколько упражнений разного уровня трудности.

Одним из потенциальных недостатков книги выступает уровень ее математической строгости. Авторы предполагают, что читатель обладает базовыми знаниями в теории вероятности и статистике. Кроме того, некоторые из наиболее сложных тем могут быть трудными для читателей, не знакомых с байесовской статистикой.

В кратком отклике на рецензируемую книгу Д. Хатчисон [Hutchison, 2018] отмечает, что она подходит для тех, кто работает в частотном подходе и хотел бы узнать про байесовский, а также для студентов аспирантуры. Рекомендация в целом верная, но требует уточнения. Во-первых, профессионалам, не уверенным в надежности своего математического бэкграунда

[6] Spiegelhalter D., Thomas A., Best N., Lunn D. (2003) WinBUGS User Manual.
[7] R Core Team (2022) R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/

(а таких среди исследователей в психологии и образовании немало), для более легкого старта стоит начать с менее математически нагруженных источников, например [Schoot van de et al., 2014]. Во-вторых, тем, кто заинтересован в освоении новой для него психометрической модели, например латентного классового анализа, целесообразно сначала освоить частотный вариант оценки параметров и после этого перейти к байесовскому.

"Bayesian Psychometric Modeling" Р. Леви и Р. Мислеви — это отличный ресурс для всех, кто заинтересован в изучении байесовской статистики и ее применении в психометрике. Книга содержит исчерпывающее описание этого подхода и включает многочисленные примеры его практического применения. Она, возможно, окажется сложной для читателей без сильной математической подготовки, но погружение в нее принесет пользу каждому, кто заинтересован в передовом анализе данных в психологии и образовании.

Благодарности Спасибо сотрудникам Центра психометрики и измерений в образовании за замечательный подарок на день рождения, который позволил мне стать ближе к байесовскому подходу, — за бумажный экземпляр книги *"Bayesian Psychometric Modeling"*.

References

Fox J.-P. (2010) *Bayesian Item Response Modeling. Theory and Applications*. New York, NY: Springer. https://doi.org/10.1007/978-1-4419-0742-4_5

Gelman A., Carlin J.B., Stern H.S., Dunson D.B., Vehtari A., Rubin D.B. (2013) *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Hutchison D. (2018) Bayesian Psychometric Modelling. *Journal of the Royal Statistical Society Series A*, vol. 181, no 2, pp. 550–550. https://doi.org/10.1111/rssa.12344

König C., van de Schoot R. (2018) Bayesian Statistics in Educational Research: A Look at the Current State of Affairs. *Educational Review*, vol. 70, no 4, pp. 486–509. http://dx.doi.org/10.1080/00131911.2017.1350636

Muthén B., Asparouhov T. (2012) Bayesian Structural Equation Modeling: A More Flexible Representation of Substantive Theory. *Psychological Methods*, vol. 17, no 3, pp. 313–335. http://dx.doi.org/10.1037/a0026802

Schoot van de R., Kaplan D., Denissen J., Asendorpf J.B., Neyer F.J., van Aken M.A. (2014) A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research. *Child Development*, vol. 85, no 3, pp. 842–860. http://dx.doi.org/10.1111/cdev.12169

Sturtz S., Ligges U., Gelman A. (2005) R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software*, vol. 12, no 3, pp. 1–16. http://dx.doi.org/10.18637/jss.v012.i03

Уважаемые коллеги,

приносим свои извинения в связи с обнаруженной ошибкой в публикации. В статье

Бочавер А.А., Михайлова О.Р. (2023). Выгорание школьников: адаптация опросника на российской выборке. *Вопросы образования / Educational Studies Moscow*, № 2, сс. 70–100. https://doi.org/10.17323/1814-9545-2023-2-70-100

в тексте опросника, размещенном в приложении, несколько утверждений приведены в промежуточном варианте перевода, а не в окончательном виде, который использовался при подготовке статьи. Приводим корректный вариант формулировок всех пунктов опросника полностью. Пожалуйста, используйте его, проектируя ваши исследования.

	Вопросы	Полностью не согласен	Не согласен	Скорее не согласен	Скорее согласен	Согласен	Полностью согласен
1	Я чувствую себя перегруженным(ой) учебой	1	2	3	4	5	6
2	Мне не хватает желания учиться в школе, я часто думаю о том, чтобы бросить учебу	1	2	3	4	5	6
3	Я часто чувствую, что недотягиваю в учебе	1	2	3	4	5	6
4	Проблемы в школе часто нарушают мой сон	1	2	3	4	5	6
5	Я чувствую, что у меня исчезает интерес к учебе в школе	1	2	3	4	5	6
6	Я постоянно задаюсь вопросом, имеет ли моя учеба в школе какой-то смысл	1	2	3	4	5	6
7	Мысли об учебе беспокоят меня даже в свободное от учебы время	1	2	3	4	5	6
8	Раньше у меня были более высокие ожидания от моей учебы в школе, чем сейчас	1	2	3	4	5	6
9	Мои отношения с родителями или друзьями портятся из-за того, что происходит в школе	1	2	3	4	5	6

Ключи:
Шкала истощения: утверждения 1, 4, 7, 9.
Шкала цинизма: утверждения 2, 5, 6.
Шкала чувства несоответствия: утверждения 3, 8.

Адрес редакции
Россия, 101000 Москва,
ул. Мясницкая, д. 20, НИУ ВШЭ
Телефон: (495) 772 95 90 *15511, *15512
E-mail: edu.journal@hse.ru
Сайт: http://vo.hse.ru

Адрес издателя и распространителя
Россия, 101000 Москва,
ул. Мясницкая, д. 20, НИУ ВШЭ
Издательский дом ВШЭ
Телефон/факс: (495) 772 95 90 *15298
E-mail: id.hse@mail.ru

Тираж 300 экз. Заказ №
Отпечатано в ООО "Фотоэксперт",
109316, Москва, Волгоградский проспект, д. 42