Межстрановая сопоставимость результатов тестирования в международных сравнительных исследованиях высшего образования

Д. А. Федерякин

Федерякин Денис Александрович
стажер-исследователь Центра психометрики и измерений в образовании Института образования НИУ ВШЭ. Адрес: 101000, Москва, ул. Мясницкая, 20. E-mail: dafederiakin@hse.ru

Аннотация. В последние 30 лет наблюдается возросший интерес к международным сравнительным исследованиям качества образования, особенно на ступени среднего образования. Главным методологическим вызовом в организации этих исследований является установление межстрановой сопоставимости результатов тестирования. Межстрановая сопоставимость результатов подразумевает, что измерительный инструмент функционирует одинаково во всех сравниваемых странах, несмотря на различия языков и культур. Специфика системы высшего образования осложняет процесс установления межстрановой сопоставимости результатов. В статье рассматривается современное понимание межстрановой сопоставимости результатов тестирования в исследованиях качества образования и способы ее установления. Анализируются специфические черты высшего образования, затрудняющие проведение стандартизированных измерений качества образования и установление межстрановой сопоставимости. Как пример преодоления этих вызовов описаны разработка дизайна и проведение международного сравнительного исследования качества высшего инженерного образования Study of Undergraduate Performance.

Ключевые слова: качество высшего образования, международные сравнительные исследования, межстрановая сопоставимость результатов тестирования, *Study of Undergraduate Performance*.

DOI: 10.17323/1814-9545-2020-2-37-59

Статья поступила в редакцию в октябре 2019 г.

Подходы к развитию систем образования, основанные на теории человеческого капитала, обосновывают экономический интерес к образованию. Для объяснения экономических успехов развитых стран исследователи все чаще обращают-

ся к рассмотрению компетенций населения как общественных и личных инвестиций [Marginson, 2019]. Инвестиционная логика рассмотрения понятия «человеческий капитал» дала мощный толчок развитию образовательной политики по всему миру [Кузьминов, Сорокин, Фрумин, 2019]. Как отмечают исследователи экономики образования, в XX в. произошли беспрецедентное увеличение количества образовательных учреждений и рост охвата населения образованием. При этом если в первой половине XX в. государственные инвестиции направлялись преимущественно в начальное и среднее образование [Meyer, Ramirez, Soysal, 1992], то вторая половина XX в. характеризовалась увеличением количества высших учебных заведений [Cantwell, Marginson, Smolentseva, 2018].

Индикаторы, используемые для измерения эффекта образования с точки зрения качества человеческого капитала, постоянно совершенствовались, и список их увеличивался. В первых исследованиях основателей теории человеческого капитала для измерения компетентности населения применялись такие переменные, как количество лет обучения (например, [Schultz, 1961]). Полученные результаты дали возможность обосновать важность бюджетных решений в области образования и в целом экономического подхода к данной сфере общественных отношений. Однако в более поздних исследованиях было установлено, что само по себе количество времени, потраченного на образование, не дает оснований судить о его результатах (например, [Hanushek, Woessmann, 2008]). Установление этого факта и послужило причиной бума международных сравнительных исследований качества образования (МСИ), начавшегося в 80-х годах XX в. Желание стейкхолдеров проверить обоснованность и окупаемость инвестиций, а также попытки заимствования наиболее удачных практик обусловили необходимость измерения качества образования и межстрановых сравнений его показателей.

Использование в международном контексте измерительных инструментов, тонко настроенных на оценку предметных компетенций, сопряжено со множеством вызовов. Наиболее критичный из них — сопоставимость результатов, особенно в высшем образовании.

Данная работа имеет две цели: во-первых, рассмотрение методологических вопросов установления межстрановой сопоставимости результатов тестирования, а во-вторых, описание методологии МСИ качества высшего образования, которая направлена на минимизацию рисков несопоставимости результатов тестирования и на учет специфики высшего образования. Статья построена следующим образом: в первом разделе рассмотрено современное понимание сопоставимости результатов тестирования; далее сравниваются существующие под-

ходы к установлению межстрановой сопоставимости результатов в МСИ; затем описаны специфические методологические трудности МСИ в высшем образовании; пути преодоления этих трудностей рассматриваются на примере одного из немногих международных исследований качества высшего образования — проекта *Study of Undergraduate Performance*, сосредоточенного на инженерном образовании.

Сопоставимость результатов тестирования

Проблема сопоставимости результатов тестирования возникает всякий раз, когда данные проведенного исследования используются для сравнения разных групп, поэтому проблематика МСИ образования является частным случаем более общей психометрической проблемы. Впервые сопоставимость результатов тестирования между группами стала темой общественного обсуждения в 1984 г., после судебного разбирательства, известного как «Страховая компания *Golden Rule* против *Washburn*»[1]. Страховая компания предъявила департаменту тестовой организации *Educational Testing Service* (ETS) в Иллинойсе претензии, что разработанные ею измерительные инструменты дискриминировали чернокожее население. В 1984 г. было достигнуто внесудебное соглашение между истцом и ответчиком, в котором ETS обязывалась переработать инструменты измерения таким образом, чтобы они не содержали дискриминирующих вопросов.

Межгрупповая сопоставимость результатов тестирования предполагает, что измерительный инструмент функционирует одинаково во всех сравниваемых группах, будь то группы, выделенные по полу, возрасту, родному языку или принадлежности к национальной выборке [Meredith, 1993]. Иными словами, представители всех групп одинаково понимают содержание теоретического конструкта и его отношения с заданиями. Если межгрупповая сопоставимость результатов не была установлена, невозможно судить о том, являются ли обнаруженные различия между группами (или отсутствие таковых) результатом неодинакового функционирования инструмента измерения в выделенных группах или проявлением объективно существующих различий в уровне выраженности целевой характеристики, которую призван измерять инструмент [Borsboom, 2006; Schmitt, Kuljanin, 2008]. Достижение сопоставимости результатов особенно затруднительно в МСИ, поскольку осложняется вмешательством культурных и языковых особенностей национальных выборок.

[1] Golden Rule Insurance Company et al. vs Washburn et al., 419–476 (stipulation for dismissal and order dismissing case, filed in the Circuit Court of the Seventh Judicial Circuit, Sangamon County, IL, 1984).

Между валидностью инструментов, используемых для МСИ, и сопоставимостью результатов, полученных с их помощью, существует прямая связь. Так, один из популярных подходов к валидности определяет это свойство инструментария как сумму доказательств, поддерживающих интерпретацию тестовых баллов [American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014]. Следовательно, если инструменты используются для сравнительного анализа, необходимо подтверждение сопоставимости результатов. Кроме того, показано, что низкая сопоставимость результатов тестирования может привести к неудовлетворительным психометрическим характеристикам инструментов при их анализе на межстрановых выборках [Church et al., 2011].

Подходы к установлению межстрановой сопоставимости

Один из наиболее популярных подходов к обеспечению межстрановой сопоставимости результатов тестирования был предложен европейскими исследователями Ф. ван де Вайвером и Н. Танцером. Они рассматривают межстрановую сопоставимость на трех уровнях [van de Vijver, Tanzer, 2004]: сопоставимость конструкта, сопоставимость метода и сопоставимость на уровне заданий. Сопоставимость на уровне конструкта подразумевает, что структура конструкта абсолютно эквивалентна во всех культурах, в которых проводится исследование. Сопоставимость метода означает эквивалентность процедуры сбора данных, эквивалентность выборок в разных странах и т. д. Сопоставимость на уровне заданий гарантирует, что задания во всех странах — участницах исследования функционируют одинаковым образом на всех уровнях способности, т. е. отсутствует дифференцированное функционирование заданий (*Differential Item Functioning*, DIF). Далее будут рассмотрены все три уровня сопоставимости, которым соответствуют три типа рисков возникновения несопоставимости результатов МСИ.

Если сопоставимость на уровне конструкта не достигнута, то конструкт имеет концептуально разное значение для респондентов из разных стран. Для установления сопоставимости на уровне конструкта необходимы теоретические и эмпирические свидетельства того, что конструкт имеет схожие определение и структуру во всех группах. Для этого применяются две группы процедур анализа: до этапа сбора данных (предварительный анализ) и после сбора данных (пост-хок анализ). На этапе предварительного анализа одна из главных процедур, направленных на установление сопоставимости конструкта, — это экспертный анализ компонентов конструкта, определяющий релевантность каждого из компонентов каждой группе респондентов, для сравнения которых предназначен инструмент [Carmines, Zel-

ler, 1979; Wynd, Schmidt, Schaefer, 2003]. Одна из главных трудностей, связанных с этой процедурой, — это выбор экспертов: они должны быть хорошо осведомлены о взаимодействии конструкта и культурных особенностей стран, а также о проявлениях конструкта внутри каждой из национальных выборок, вызванных этим взаимодействием. После того как эксперты выделят одинаково релевантные для всех выборок аспекты конструкта и связанные с ними поведенческие индикаторы, начинается этап разработки заданий, составляющих инструмент.

Пост-хок процедуры, направленные на установление сопоставимости на уровне конструкта, обычно представляют собой анализ того, как группируются задания (или поведенческие индикаторы), т. е. выяснение того, одинаковы ли по своему содержанию и связи друг с другом ненаблюдаемые переменные, используемые для дифференцирования респондентов. Если размерность инструментов измерения различается между группами — например, если ненаблюдаемая переменная в одной группе, в отличие от других, распадается на две или больше переменных, результаты, полученные в разных странах, будут несопоставимы.

Наиболее частые причины несопоставимости на уровне метода — это неэквивалентность процедур сбора данных; неэквивалентность выборок в разных странах; различие инструкций, предъявленных респондентам; разная степень знакомства респондентов с форматом используемых заданий; различающиеся стратегии и стили ответов [van de Vijver, Tanzer, 2004]. Такие источники неэквивалентности могут стать критическими для сопоставимости результатов в случае, если они не были проконтролированы [Davidov et al., 2014].

Для достижения сопоставимости на уровне метода необходимо, чтобы все процедуры измерения были полностью стандартизированы. Так, в технических отчетах одного из ведущих исследований среднего образования, PISA, большой раздел отводится описанию всех требований к процедурам проведения тестирования [OECD, 2015].

Пост-хок статистический анализ сопоставимости на уровне метода оставался неразработанным на протяжении большей части истории МСИ, поскольку он требует сбора информации о *процессе* выполнения тестовых заданий, а не только о *результатах* их выполнения, которые традиционно были фокусом психометрических исследований на протяжении большей части XX в. Однако развитие технологий компьютерного тестирования позволило исследователям собирать данные о поведении респондентов во время выполнения заданий, и тогда появились публикации с анализом процессов и стратегий выполнения заданий. (Такая информация часто обозначается в научной литературе как коллатеральная [Mislevy, 1988].) Одной

из фундаментальных работ, давших начало этому направлению в психометрике, является статья В. ван дер Линдена, в которой в качестве процедуры контроля сопоставимости используется анализ времени выполнения заданий [van der Linden, 2007]. В дальнейшем появились работы, в которых анализировалось не только время выполнения заданий, но и представления исследователей о последовательности выборов, совершаемых респондентами [Jeon, de Boeck, 2016], а также изменение когнитивных стратегий при выполнении различных заданий (например, [Tijmstra, Bolsinova, Jeon, 2018]). В целом эта область психометрических исследований является сегодня одной из самых бурно развивающихся.

Одной из главных работ по теме сопоставимости отдельных заданий является статья Л. Руссоса и У. Стаута [Roussos, Stout, 1996]. Неэквивалентность заданий обычно означает, что те или иные задания содержат дополнительные ненаблюдаемые характеристики — вторичные конструкты, которые данное задание также измеряет и по которым различаются национальные группы респондентов. Среди причин такой неэквивалентности обычно указывают трудности перевода текста заданий и передачи тонкостей смысла слов на других языках, а также неоднозначность каких-то идей или формулировок, в том числе с негативными или эмоционально насыщенными коннотациями [van de Vijver, Tanzer, 2004].

Для предотвращения несопоставимости результатов тестирования на уровне заданий в процессе разработки инструмента измерения обычно используются различные формализованные процедуры перевода. Например, применяется комбинация прямого и обратного перевода: чередуются перевод и подробное комментирование заданий с одного языка на другой и обратно. Изначально считалось, что такая техника позволяет команде разработчиков уловить тонкие изменения в смыслах заданий (например, [Hambleton, 1993]). Однако этот метод также является одним из самых критикуемых в научной литературе (например, [Brislin, 1970]): реализация прямого и обратного перевода может приводить к большому количеству итераций без значительного улучшения адаптированных версий заданий. В качестве альтернатив этой технике предлагаются такие процедуры, как:

- параллельный перевод с реконсилиацией (несколько переводчиков одновременно переводят и комментируют задания, а затем редактор, являющийся специалистом в исследовании целевого конструкта, сводит разные версии перевода в итоговую [OECD, 2017; 2016]);
- групповой перевод (группа переводчиков встречается очно и согласовывает перевод всех заданий одного за другим [Hambleton, Kanjee, 1995]);

- использование в переводе заданий специалистов-билингвов, которые являются не просто квалифицированными переводчиками, а естественными носителями двух и более языков, жившими в условиях двуязычной среды с раннего детства, и потому обладают развитым чувством языка [Annette, Dannenburg, Janet, 1994];
- а также их множественные комбинации (например, [Lenz et al., 2017]).

Среди наиболее часто применяемых техник пост-хок статистического анализа сопоставимости на уровне заданий — методы анализа дифференцированного функционирования заданий (*Differential Item Functioning*, DIF [Holland, Wainer, 1993]). В ходе такого анализа выясняют, демонстрируют ли задания сопоставимые психометрические характеристики внутри каждой из групп при контроле уровня выраженности целевого конструкта.

В методологической литературе, посвященной обеспечению сопоставимости на уровне заданий, подчеркивается важность следующего за статистическим анализом этапа анализа — интерпретации оцененных статистик (например, [Wang, Shih, Sun, 2012]). В частности, указывается, что если задания демонстрируют несопоставимые психометрические характеристики, но эксперты в целевом конструкте и специфических культурных средах не способны объяснить содержательно причину этих различий, задания не должны рассматриваться как функционирующие несопоставимо.

В случае, если некоторые из разработанных заданий демонстрируют DIF, необходимы специальные процедуры, направленные на нивелирование этого эффекта. Одна из таких процедур — удаление заданий, по-разному функционирующих в разных группах [Magis, Facon, 2013]. Однако тем самым можно нарушить баланс в содержании теста, что несет угрозу содержательной валидности выводов, сделанных по результатам сравнений. С другой стороны, удаление заданий может увеличить ошибку измерения и, как следствие, снизить надежность измерений. Альтернативной процедурой является так называемое расщепление заданий, возможное только в рамках современной теории тестирования [Brodersen et al., 2007]: задание, дифференцированно функционирующее в разных группах, рассматривается как несколько групп-специфичных заданий, при этом допускается существование различающихся для разных групп параметров задания. Этот подход позволяет сохранить баланс содержания измерительного инструмента внутри стран и не увеличивать ошибку измерения, сохраняя психометрические характеристики инструмента на приемлемом уровне.

Другой известный подход к установлению межстрановой сопоставимости результатов тестирования предложили К. Эрсикан

и Ж. Лайон-Томас [Ercikan, Lyons-Thomas, 2013]. Их идея заключается в выделении нескольких групп потенциальных различий, которые уменьшают сопоставимость результатов и валидность выводов, сделанных на основе сравнений:

1) различия выборки;
2) различия конструкта (неэквивалентность психологической реальности, стоящей за исследуемым конструктом, обусловленная различиями в культурных средах);
3) различия инструмента измерения (в первую очередь DIF, языковые различия и вызванные ими различия в презентации информации);
4) различия в процедурах администрирования инструмента;
5) различия в процессе выполнения заданий (в первую очередь разница в стратегиях выполнения заданий).

Очевидно, что между двумя приведенными теоретическими рамками понимания межстрановой сопоставимости результатов можно провести параллели и рассматриваемые в них основные источники проблем в целом схожи. Кроме того, обе теоретические рамки имплицитно подразумевают, что устранение указанных источников различий или причин несопоставимости автоматически приводит к установлению межстрановой сопоставимости результатов тестирования. Рамка, предложенная К. Эрсикан и Ж. Лайонс-Томас, представляется более удобной в использовании для адаптации уже существующих инструментов к применению на новых национальных выборках, поскольку она позволяет элиминировать главные группы источников несопоставимости в порядке этапов разработки инструментов измерения и анализа их результатов. В то же время рамку, предложенную Ф. ван де Вайвером и Н. Танцером, целесообразно использовать в процессе разработки инструментов измерения, с самого начала ориентированных на применение в международном контексте, поскольку она концентрируется на интеграции разных причин несопоставимости данных и их рассмотрении на всех этапах разработки инструментов измерения.

Специфика установления межстрановой сопоставимости результатов в высшем образовании

В современной литературе описаны трудности установления межстрановой сопоставимости результатов тестирования вне зависимости от ступени образования, на которой проводятся МСИ качества образования [Kankaraš, Moors, 2014]. В частности, среди методологических вызовов указываются различия культурных и экономических сред, в которых сравниваются системы образования [Bray, Thomas, 1995], и различия в структуре самих систем образования [Schneider, 2009].

Однако система высшего образования имеет ряд черт, которые отличают ее от остальных ступеней образования и обусловливают специфику процедур измерения образовательных достижений. Оценка качества в высшем образовании отличается от оценки в среднем образовании, для которого существует ряд широкомасштабных сравнительных международных исследований (например, PISA и TIMSS), активно развивающих методологию МСИ [OECD, 2017; Martin, Mullis, Hooper, 2016].

Одним из первых проектов, направленных на международное сравнение систем высшего образования, был *Assessment of Learning Outcomes in Higher Education* (AHELO) [Tremblay, 2013]. Именно опыт этого проекта показал, насколько специфично международное сравнительное оценивание в высшем образовании. Критика, с которой столкнулся AHELO, легла в основу практически всех методологических наработок МСИ в высшем образовании.

МСИ в области высшего образования отличаются от МСИ на остальных ступенях системы образования вследствие следующих особенностей высшего образования:

- большое разнообразие содержания образования внутри стран (даже по одним и тем же направлениям подготовки) и глубокие различия в его содержании между странами, в то время как среднее образование в значительной степени унифицировано [Zlatkin-Troitschanskaia et al., 2017]. По этой причине трудно выбрать материал, на основе которого могут быть составлены измерительные инструменты, таким образом, чтобы оценить образовательные достижения студентов, не дискриминируя ни одну из групп. Несоответствие измерительных инструментов данному требованию было одним из основных пунктов критики AHELO [Altbach, 2015];
- высокая селективность системы высшего образования, определяющая предпочтительность лонгитюдных сравнений перед срезовыми. Значительная доля критики проекта AHELO была вызвана недостатком внимания к динамике показателей. Необходимость их учета не столько затрудняет обеспечение межстрановой сопоставимости результатов тестирования, сколько усложняет дизайн МСИ — а следовательно, и установление сопоставимости и выбор содержания для измерительных инструментов:
  – если в одной стране высшее образование более селективно, чем в другой, сравнение результатов этих стран затруднено в силу смещения оценок популяционных параметров, поскольку часть обнаруженных различий будет объясняться именно сложностью поступления в вузы. В таких случаях необходимы не только срезовые сравнения, но и выделение «вклада института» в образовательные до-

стижения студентов, т. е. лонгитюдный дизайн измерения [Jamelske, 2009];
  - сами вузы в национальной выборке могут различаться по селективности. В элитные образовательные институты поступают более талантливые студенты, в результате такого отбора возникают трудности при определении «вклада института» в прогресс студентов. При этом задача измерения вклада института чрезвычайно актуальна в оценке качества высшего образования, поскольку, в отличие от среднего образования, в него вовлекаются не все люди соответствующего возраста [Jamelske, 2009];
  - возможность отчисления студентов, не справляющихся с выполнением образовательных стандартов. На данных 18 стран — участниц Организации экономического сотрудничества и развития показано, что в среднем около 31% студентов начинают, но не заканчивают высшее образование. К тому же этот показатель существенно разнится между странами — от около 54% в США до 10% в Японии. Этот факт диктует необходимость коррекции результатов на отчисление студентов, чтобы избежать смещений в результатах анализа «вклада института» [OECD, 2010];
- подход, центрированный на достижениях студентов, как необходимое условие релевантности оценки образовательных достижений образовательной системе. При этом университеты могут находиться в разных стадиях развития, несравнимых между собой. Так, сравнение только что разработанной образовательной программы с работающей уже несколько лет не выдерживает критики. Процесс «отладки» образовательной программы может занять несколько десятилетий даже в странах — участницах Болонского процесса [Rauhvargers, 2011]. Отдельную и сложную задачу представляет собой оценка навыков мышления высшего порядка [Zlatkin-Troitschanskaia, Pant, Greiff, 2019]. В одних странах и университетах акцент делается именно на развитии этих навыков, другие образовательные системы концентрируются на усвоении студентами профессиональных знаний, и эти различия также затрудняют установление сопоставимости результатов тестирования;
- высокие риски неверного использования результатов исследований. Осознание потенциального воздействия итогов мониторинга на автономию институтов и академическую свободу может приводить к намеренному искажению результатов на разных уровнях реализации исследования. Акцент на этой проблеме сделан в документации проекта AHELO. Мониторинговые исследования, подобные AHELO, не предназначены для построения, например, рейтингов вузов, но попытки неадекватного использования обратной

связи по этим проектам были зарегистрированы [Tremblay, 2013]. В частности, вузы могут пытаться завысить свои результаты, для того чтобы поднять свой рейтинг. Таким образом, неверное использование результатов может привести как к необоснованным и ошибочным выводам, так и к попыткам нарушения процедуры сбора данных [Ibid.];

- мотивированность студентов к участию в исследовании как условие получения надежных результатов тестирования. Если студенты вузов не замотивированы демонстрировать свои навыки, их результаты не могут служить показателем качества системы высшего образования. В МСИ качества образования используются способы минимизации рисков воздействия неправомерных выводов на систему образования (например, избегание индивидуальной оценки, избегание ранжирования образовательных организаций). В результате может возникать ситуация тестирования «с низкими ставками», когда респонденты не замотивированы демонстрировать свои навыки. В таких условиях нельзя рассчитывать на то, что ответы студентов действительно показывают уровень подготовки в вузе, и возникает необходимость использования дополнительных средств мотивирования участников тестирования [Banta, Pike, 2012]. Уровень мотивации участников исследования не только влияет на показатели образовательных достижений, но и может различаться между странами, что усложняет обеспечение межстрановой сопоставимости результатов [Goldhammer et al., 2016].

Чтобы минимизировать риски возникновения межстрановой несопоставимости результатов тестирования в высшем образовании, требуется принять во внимание все перечисленные факторы, влияние которых на других ступенях образования существенно слабее. Таким образом, МСИ качества высшего образования существенно отличаются от исследований других ступеней образования в силу специфики объекта изучения, что накладывает отпечаток на методологию таких исследований.

Способы установления межстрановой сопоставимости в проекте SUPERtest

Study of Undergraduate PERformance (SUPERtest) — это проект, посвященный межстрановому сравнительному исследованию качества высшего инженерного образования в России, Индии, Китае и США [Loyalka et al., 2019]. Целью проекта является сравнение качества высшего инженерного образования на репрезентативных национальных выборках и выяснение факторов институционального и индивидуального характера, связанных с уровнем образовательных достижений студентов инженерных специальностей. Для этих целей проведен двухэтапный сбор данных на двух когортах, который позволяет измерить компе-

тенции студентов на индивидуальном уровне, проследить их динамику спустя два года и оценить вклад института в развитие студентов (стартовый замер — на 1-м и на 3-м курсах, второй замер — в конце 2-го и 4-го курсов).

Инструментарий проекта включает комплекс методик, направленных на измерение компетенций студентов в дисциплинах фундаментального и профессионального циклов, навыков мышления высшего порядка (логического мышления, критического мышления и креативности), а также серию анкет для студентов, преподавателей и администраторов вузов, направленных на сбор большого количества контекстной информации. Исследование проводилось в компьютеризированной форме, что позволило не только оптимизировать процесс сбора информации, но и получить сведения о поведении респондентов во время тестирования.

Процедуры разработки инструментов измерения в проекте *SUPERtest* были направлены на минимизацию рисков в процессе установления межстрановой сопоставимости результатов с использованием подхода Ф. ван де Вайвера и Н. Танцера [van de Vijver, Tanzer, 2004]. Три уровня анализа сопоставимости, предусмотренные этим подходом, не являются этапами разработки инструментов измерения. Более того, процесс разработки инструментов измерения, основанный на этом подходе, требует интеграции всех трех уровней анализа сопоставимости на каждом этапе разработки. В процессе подготовки проекта *SUPERtest* некоторые этапы разработки инструментов измерения обеспечивали несколько уровней сопоставимости результатов. Поскольку целевые группы студентов узко определены задачами проекта, сопоставимость выборок не является проблемой для данного проекта.

На первом этапе был проведен анализ содержания и конструктной валидности с использованием экспертов в национальных системах образования, которые оценивали области содержания образования, субобласти и сами тестовые задания для каждой дисциплины. Этот этап был направлен на обеспечение сопоставимости конструкта между странами. Эксперты были рекрутированы из ряда элитных и неэлитных инженерных программ в Китае, России, Индии и США. Для отбора элементов содержания инструментов измерения использовались национальные эксперты в области высшего инженерного образования. Они выделили содержательные области знаний, которые проходят студенты в каждой из стран-участниц согласно образовательным стандартам. Далее отобранные элементы содержания переводились на языки стран-участниц командой, состоящей из профессиональных переводчиков и специалистов в предмете изучения, говорящих на соответствующих языках. После этого все национальные эксперты в системах образова-

ния рейтинговали все перечисленные области элементов потенциального содержания инструментов с целью выделения тем, наиболее важных для дальнейшей профессиональной деятельности выпускников. Результаты этих рейтингов обрабатывались с помощью многофасетного Раш-моделирования с целью выделить наиболее релевантные области без учета индивидуальных эффектов и искажений каждого из экспертов (например, [Zhu, Ennis, Chen, 1998]).

После того как были объективно выделены наиболее важные области профессиональных компетенций, начинался этап отбора заданий. На этом этапе необходимо было обеспечить сопоставимость на уровне заданий. Силами всех экспертов был собран большой пул заданий: в него вошли все задания, которые любой из экспертов счел необходимым предложить. Далее этот пул заданий поэтапно просеивали через серию критериев и оценочных процедур. На первом этапе осуществлялась экспертная оценка разработанного пула заданий — такую же оценку проходили все предложенные национальными экспертами элементы содержания. Критерии, по которым эксперты оценивали задания, включали предполагаемую трудность задания; количество времени, которое потратит средний студент на решение этого задания; уровень сложности когнитивных операций, необходимых для правильного выполнения этого задания, и т. д. Такое оценивание позволило подготовить для первой апробации задания, которые измеряют наиболее важные и одинаково релевантные во всех странах элементы содержания.

Далее следовал этап апробации отобранных заданий, на котором принимались меры к обеспечению одновременно всех уровней межстрановой сопоставимости. Сначала в каждой из стран — участниц проекта была проведена серия когнитивных лабораторий (интервью с рассуждениями вслух), которые помогли понять, как респонденты из целевой популяции воспринимают и обрабатывают информацию из заданий. По результатам этих интервью были подготовлены и переведены на иностранные языки комментарии респондентов из всех стран-участниц, которые позволили выделить задания, которые оказались наиболее неоднозначными и неясными. В частности, на этом этапе учитывались национально специфические трудности в понимании материала. Затем была проведена серия небольших апробаций с последующими фокус-группами, в которых респондентов просили обсуждать тестовые материалы и свое восприятие заданий. На этом этапе анализировалось не столько содержание инструментов, сколько способы презентации и организации этого содержания, а также рекомендации респондентов по организации процедуры тестирования. Эта часть апробационных исследований была направлена на установление сопоставимости метода, а именно:

- на изучение стратегий решения заданий;
- оценку того, насколько знакомы респондентам данные форматы заданий;
- анализ способов стандартизации процедуры тестирования, которые были бы приемлемы для респондентов из всех стран-участниц.

После этого процедура администрирования теста была полностью стандартизирована и согласована с представителями всех стран — участниц проекта, чтобы обеспечить возможность объективного анализа характеристик заданий.

Следующим этапом разработки инструментов стали полномасштабные апробации, в которых анализировались психометрические характеристики заданий. На этом этапе контролируется сопоставимость конструкта (с помощью методов психометрического анализа количества ненаблюдаемых характеристик, измеряемых инструментом), а также сопоставимость заданий (DIF-анализ). На каждом этапе пул заданий сокращался за счет тех заданий, которые демонстрировали наибольшие проблемы: например, наиболее сложные формулировки (в том числе те, которые характеризовались высокой неоднозначностью перевода и вариативностью терминологии), затруднения в организации иллюстраций. Кроме того, серия полномасштабных апробаций позволила уточнить инструкции, которые зачитывались респондентам и администраторам тестирования, что также способствует сопоставимости метода и повышению мотивации студентов [Liu, Rios, Borden, 2015].

После этого были составлены национально репрезентативные выборки в каждой из стран-участниц. Выборки формировались с учетом кластеризации респондентов в целях снижения стоимости исследования. Рандомизация в процессе организации выборки позволяет обеспечить сопоставимость результатов тестирования на уровне метода.

После сбора данных применялись пост-хок статистические анализы для обеспечения межстрановой сопоставимости результатов тестирования. Статистические процедуры в проекте *SUPERtest* лежат в парадигме современной теории тестирования и зависят от конкретного измерительного инструмента. Для проверки межстрановой сопоставимости на уровне всего конструкта используются анализ размерности одномерных инструментов измерения [Kardanova et al., 2016] и моделирование бифакторных структур [Wang, Wilson, 2005] для композитных инструментов измерения [Dumas, Alexander, 2016]. Для проверки межстрановой сопоставимости на уровне метода применяются древовидные модели современной теории тестирования [Jeon, de Boeck, 2016] с учетом времени ответа на задания [Molenaar, Tuerlinckx, van der Maas, 2015]. Для проверки сопоставимости

на уровне отдельных заданий используются наиболее хорошо изученные и известные методы DIF-анализа [Rogers, Swaminathan, 1993]. Описание этих процедур лежит за пределами данной работы, поскольку они представляют собой применение только некоторых статистических техник.

Описанная методология позволила разработать инструменты измерения, которые дают возможность сравнивать уровни развития компетенций у студентов, получающих высшее инженерное образование в разных странах. Кроме того, с помощью последующих статистических анализов, использующихся в проекте, оценивается межстрановая сопоставимость уже собранных данных.

Заключение

Главными драйверами методологии МСИ в области образования, безусловно, являются исследования, сосредоточенные на ступени среднего образования, — такие, как PISA и TIMSS. Эти исследования заложили основу методологии реализации МСИ, определив традиционное понимание их дизайна и целей. В последние годы наблюдается растущая потребность в подобных исследованиях на уровне высшего образования. Однако попытки реализации идеи МСИ применительно к высшему образованию были до сих пор не вполне успешными.

Опыт проекта AHELO дал толчок развитию нескольких других проектов по оценке в высшем образовании (например, [Zlatkin-Troitschanskaia et al., 2017; Shavelson, Zlatkin-Troitschanskaia, Mariño, 2018; Aloisi, Callaghan, 2018]), но при этом показал, что использование традиционных подходов к разработке измерительного инструментария является неудачной стратегией для МСИ на уровне высшего образования. В силу большого разнообразия образовательных программ крайне затруднительно предоставить интерпретируемые тестовые баллы студентов из разных университетов даже внутри одной страны, не говоря уже о шкалах для межстрановых сравнений. Методологические вызовы, специфические для высшего образования, затрудняют проведение подобных проектов в области высшего образования.

Опыт последующих проектов, в частности проекта *SUPERtest*, убеждает в возможности проводить МСИ и на ступени высшего образования. Для этого необходимы существенные модификации дизайна исследования, в первую очередь направленные на обеспечение межстрановой сопоставимости результатов тестирования. Такие модификации дизайна должны быть основаны на одном из подходов к установлению межстрановой сопоставимости, которые систематизируют причины несопоставимости результатов тестирования и предоставляют согласованную теоретическую рамку для их понимания и ми-

нимизации их воздействия. В исследовании *SUPERtest* используется подход Ф. ван де Вайвера и Н. Танцера [van de Vijver, Tanzer, 2004]. Он оптимален для создания инструментов измерения «с нуля», поскольку естественным образом вынуждает разработчиков проектировать инструменты измерения, принимая во внимание все уровни сопоставимости: уровень конструкта (его структуры и понимания), уровень метода (процедуры сбора данных) и уровень заданий (психологический смысл каждого отдельного используемого индикатора). Использование этого подхода позволило разработать методологию МСИ, изначально ориентированную на создание инструментария, пригодного для межстрановых сравнений. Описанная методология разработки инструментов измерения является в высокой степени универсальной и может быть применена в других МСИ качества образования.

Литература

1. Кузьминов Я., Сорокин П., Фрумин И. (2019) Общие и специальные навыки как компоненты человеческого капитала: новые вызовы для теории и практики образования // Форсайт. № 13 (S2). С. 19–41.
2. Aloisi C., Callaghan A. (2018) Threats to the Validity of the Collegiate Learning Assessment (CLA+) as a Measure of Critical Thinking Skills and Implications for Learning Gain // Higher Education Pedagogies. Vol. 3. No 1. P. 57–82.
3. Altbach P. G. (2015) AHELO: The Myth of Measurement and Comparability // International Higher Education. No 82. P. 2–3.
4. American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014) Standards for Educational and Psychological Testing. Washington, DC.
5. Annette D. G., Dannenburg L., Janet G. (1994) Forward and Backward Word Translation by Bilinguals // Journal of Memory and Language. Vol. 33. No 5. P. 600–629.
6. Banta T., Pike G. (2012) Making the Case Against — One More Time // Occasional Paper of National Institute for Learning Outcomes Assessment. Vol. 15. P. 24–30.
7. Borsboom D. (2006) When Does Measurement Invariance Matter? // Medical Care. Vol. 44. No 11. P. S176–S181.
8. Bray M., Thomas R. M. (1995) Levels of Comparison in Educational Studies: Different Insights from Different Literatures and the Value of Multilevel Analyses // Harvard Educational Review. Vol. 65. No 3. P. 472–490.
9. Brislin R. W. (1970) Back-Translation for Cross-Cultural Research // Journal of Cross-Cultural Psychology. Vol. 1. No 3. P. 185–216.
10. Brodersen J., Meads D., Kreiner S., Thorsen H., Doward L., McKenna S. (2007) Methodological Aspects of Differential Item Functioning in the Rasch Model // Journal of Medical Economics. Vol. 10. No 3. P. 309–324.
11. Cantwell B., Marginson S., Smolentseva A. (eds) (2018) High Participation Systems of Higher Education. Oxford: Oxford University.
12. Carmines E. G., Zeller R. A. (1979) Reliability and Validity Assessment. Vol. 17. Thousand Oaks, CA: Sage.
13. Church A. T., Alvarez J. M., Mai N. T., French B. F., Katigbak M. S., Ortiz F. A. (2011) Are Cross-Cultural Comparisons of Personality Profiles Meaningful? Differential Item and Facet Functioning in the Revised NEO Per-

sonality Inventory // Journal of Personality and Social Psychology. Vol. 101. No 5. P. 1068–1089.
14. Davidov E., Meuleman B., Cieciuch J., Schmidt P., Billiet J. (2014) Measurement Equivalence in Cross-National Research // Annual Review of Sociology. Vol. 40. P. 55–75.
15. Dumas D., Alexander P. A. (2016) Calibration of the Test of Relational Reasoning // Psychological Assessment. Vol. 28. No 10. P. 1303–1319.
16. Ercikan K., Lyons-Thomas J. (2013) Adapting Tests for Use in Other Languages and Cultures // K. F. Geisinger (ed.) APA Handbook of Testing and Assessment in Psychology. Washington: American Psychological Association. Vol. 3. P. 545–569.
17. Goldhammer F., Martens T., Christoph G., Lüdtke O. (2016) Test-Taking Engagement in PIAAC. Paris: OECD.
18. Hambleton R. K. (1993) Translating Achievement Tests for Use in Cross-National Studies // European Journal of Psychological Assessment. Vol. 9. No 1. P. 57–68.
19. Hambleton R. K., Kanjee A. (1995) Increasing the Validity of Cross-Cultural Assessments: Use of Improved Methods for Test Adaptations // European Journal of Psychological Assessment. Vol. 11. No 3. P. 147–157.
20. Hanushek E. A., Woessmann L. (2008) The Role of Cognitive Skills in Economic Development // Journal of Economic Literature. Vol. 46. No 3. P. 607–668.
21. Holland P. W., Wainer H. (1993) Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum Associates.
22. Jamelske E. (2009) Measuring the Impact of a University First-Year Experience Program on Student GPA and Retention // Higher Education. Vol. 57. No 3. P. 373–391.
23. Jeon M., De Boeck P. (2016) A Generalized Item Response Tree Model for Psychological Assessments // Behavior Research Methods. Vol. 48. No 3. P. 1070–1085.
24. Kankaraš M., Moors G. (2014) Analysis of Cross-Cultural Comparability of PISA 2009 Scores //Journal of Cross-Cultural Psychology. Vol. 45. No 3. P. 381–399.
25. Kardanova E., Loyalka P., Chirikov I. et al. (2016) Developing Instruments to Assess and Compare the Quality of Engineering Education: The Case of China and Russia // Assessment & Evaluation in Higher Education. Vol. 41. No 5. P. 770–786.
26. Lenz S. A., Soler I. G., Dell'Aquilla J., Uribe P. M. (2017) Translation and Cross-Cultural Adaptation of Assessments for Use in Counseling Research // Measurement and Evaluation in Counseling and Development. Vol. 50. No 4. P. 224–231.
27. Liu O. L., Rios J. A., Borden V. (2015) The Effects of Motivational Instruction on College Students' Performance on Low-Stakes Assessment // Educational Assessment. Vol. 20. No 2. P. 79–94.
28. Loyalka P., Liu O. L., Li G. et al. (2019) Computer Science Skills Across China, India, Russia, and the United States // Proceedings of the National Academy of Sciences. Vol. 116. No 14. P. 6732–6736.
29. Magis D., Facon B. (2013) Item Purification Does Not Always Improve DIF Detection: A Counterexample with Angoff's Delta Plot // Educational and Psychological Measurement. Vol. 73. No 2. P. 293–311.
30. Marginson S. (2019) Limitations of Human Capital Theory // Studies in Higher Education. Vol. 44. No 2. P. 287–301.
31. Martin M. O., Mullis I. V. S., Hooper M. (eds) (2016) Methods and Procedures in TIMSS2015. http://timssandpirls.bc.edu/publications/timss/2015-methods.html

32. Meredith W. (1993) Measurement Invariance, Factor Analysis and Factorial Invariance // Psychometrika. Vol. 58. No 4. P. 525–543.
33. Meyer J. W., Ramirez F. O., Soysal Y. N. (1992) World Expansion of Mass Education, 1870–1980 // Sociology of Education. Vol. 65. No 2. P. 128–149.
34. Mislevy R. J. (1988) Exploiting Collateral Information in the Estimation of Item Parameters // ETS Research Report Series. Vol. 2. P. 1–31.
35. 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. P. 56–74.
36. OECD (2017) PISA 2015 Technical Report. https://www.oecd.org/pisa/data/2015-technical-report/
37. OECD (2016) PISA 2018 Translation and Adaptation Guidelines. https://www.oecd.org/pisa/pisaproducts/PISA-2018-translation-and-adaptation-guidelines.pdf
38. OECD (2015) PISA 2018 Technical Standards. https://www.oecd.org/pisa/pisaproducts/PISA-2018-Technical-Standards.pdf
39. OECD (2010) Education at a Glance 2010: OECD Indicators. www.oecd.org/edu/eag2010
40. Rauhvargers A. (2011) Global University Rankings and Their Impact: EUA Report on Rankings 2011. Brussels: European University Association.
41. Rogers H. J., Swaminathan H. (1993) A Comparison of Logistic Regression and Mantel-Haenszel Procedures for Detecting Differential Item Functioning // Applied Psychological Measurement. Vol. 17. No 2. P. 105–116.
42. Roussos L., Stout W. (1996) A Multidimensionality-Based DIF Analysis Paradigm // Applied Psychological Measurement. Vol. 20. No 4. P. 355–371.
43. Schmitt N., Kuljanin G. (2008) Measurement Invariance: Review of Practice and Implications // Human Resource Management Review. Vol. 18. No 4. P. 210–222.
44. Schneider S. L. (2009) Confusing Credentials: The Cross-Nationally Comparable Measurement of Educational Attainment (PhD Thesis). Oxford: Oxford University.
45. Schultz T. W. (1961) Investment in Human Capital // The American Economic Review. Vol. 51. No 1. P. 1–17.
46. Shavelson R. J., Zlatkin-Troitschanskaia O., Mariño J. P. (2018) International Performance Assessment of Learning in Higher Education (iPAL): Research and Development // O. Zlatkin-Troitschanskaia, M. Toepper, H. A. Pant, C. Lautenbach, C. Kuhn (eds) Assessment of Learning Outcomes in Higher Education. Springer. P. 193–214.
47. Tijmstra J., Bolsinova M. A., Jeon M. (2018) Generalized Mixture IRT Models with Different Item-Response Structures: A Case Study Using Likert-Scale Data // Behavior Research Methods. Vol. 50. No 6. P. 2325–2344.
48. Tremblay K. (2013) OECD Assessment of Higher Education Learning Outcomes (AHELO): Rationale, Challenges and Initial Insights from the Feasibility Study // S. Blömeke, O. Zlatkin-Troitschanskaia, Ch. Kuhn, Ju. Fege (eds) Modeling and Measuring Competencies in Higher Education. Rotterdam: Brill Sense. P. 113–126.
49. Van de Vijver F., Tanzer N. K. (2004) Bias and Equivalence in Cross-Cultural Assessment: An Overview // European Review of Applied Psychology. Vol. 54. No 2. P. 119–135.
50. Van der Linden W. J. (2007) A Hierarchical Framework for Modeling Speed and Accuracy on Test Items // Psychometrika. Vol. 72. No 3. P. 287–308.

51. Wang W. C., Wilson M. (2005) The Rasch Testlet Model // Applied Psychological Measurement. Vol. 29. No 2. P. 126–149.
52. Wang W. C., Shih C. L., Sun G. W. (2012) The DIF-free-then-DIF Strategy for the Assessment of Differential Item Functioning // Educational and Psychological Measurement. Vol. 72. No 4. P. 687–708.
53. Wynd C. A., Schmidt B., Schaefer M. A. (2003) Two Quantitative Approaches for Estimating Content Validity // Western Journal of Nursing Research. Vol. 25. No 5. P. 508–518.
54. Zhu W., Ennis C. D., Chen A. (1998) Many-Faceted Rasch Modeling Expert Judgment in Test Development // Measurement in Physical Education and Exercise Science. Vol. 2. No 1. P. 21–39.
55. Zlatkin-Troitschanskaia O., Pant H. A., Greiff S. (2019) Assessing Generic and Domain-Specific Academic Competencies in Higher Education // Zeitschrift für Pädagogische Psychologie. Vol. 33. No 2. P. 91–93. ·
56. Zlatkin-Troitschanskaia O., Pant H. A., Lautenbach C., Molerov D., Toepper M., Brückner S. (2017) Modeling and Measuring Competencies in Higher Education. Approaches to Challenges in Higher Education Policy and Practice. Wiesbaden: Springer.

Cross-National Comparability of Assessment in Higher Education

Author **Denis Federiakin**
Intern Researcher, Center for Psychometrics and Measurements in Education, Institute of Education, National Research University Higher School of Economics. Address: 20 Myasnitskaya Str., 101000 Moscow, Russian Federation. E-mail: dafederiakin@hse.ru

Abstract The last three decades have seen an increase in researchers' interest in international comparative assessments of educational outcomes, particularly at the level of secondary schools. Achieving cross-national comparability is the main methodological challenge in the design of such studies. Cross-national comparability of test scores implies that the measure operates similarly across all the participating countries, regardless of their linguistic and cultural differences. The process of achieving cross-national comparability in higher education is more complicated due to specific features of higher education. This article explores the modern understanding of cross-national comparability of student assessment results and the possible ways of achieving it. It analyzes the specific aspects of higher education that complicate standardized measurement of educational outcomes and trivial achievement of cross-national comparability. The process of designing and conducting the Study of Undergraduate Performance—an international comparative research project aimed to assess and compare higher engineering education across nations—is described as an example of overcoming those challenges.

Keywords quality of higher education, international comparative assessments, cross-cultural comparability of test scores, Study of Undergraduate Performance.

References
Aloisi C., Callaghan A. (2018) Threats to the Validity of the Collegiate Learning Assessment (CLA+) as a Measure of Critical Thinking Skills and Implications for Learning Gain. *Higher Education Pedagogies*, vol. 3, no 1, pp. 57–82.

Altbach P. G. (2015) AHELO: The Myth of Measurement and Comparability. *International Higher Education*, no 82, pp. 2–3.

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

Annette D. G., Dannenburg L., Janet G. (1994) Forward and Backward Word Translation by Bilinguals. *Journal of Memory and Language*, vol. 33, no 5, pp. 600–629.

Banta T., Pike G. (2012) Making the Case Against—One More Time. *Occasional Paper of National Institute for Learning Outcomes Assessment*, vol. 15, pp. 24–30.

Borsboom D. (2006) When Does Measurement Invariance Matter? *Medical Care*, vol. 44, no 11, pp. S176–S181.

Bray M., Thomas R. M. (1995) Levels of Comparison in Educational Studies: Different Insights from Different Literatures and the Value of Multilevel Analyses. *Harvard Educational Review*, vol. 65, no 3, pp. 472–490.

Brislin R. W. (1970) Back-Translation for Cross-Cultural Research. *Journal of Cross-Cultural Psychology*, vol. 1, no 3, pp. 185–216.

Brodersen J., Meads D., Kreiner S., Thorsen H., Doward L., McKenna S. (2007) Methodological Aspects of Differential Item Functioning in the Rasch Model. *Journal of Medical Economics*, vol. 10, no 3, pp. 309–324.

Cantwell B., Marginson S., Smolentseva A. (eds) (2018) *High Participation Systems of Higher Education*. Oxford: Oxford University.

Carmines E. G., Zeller R. A. (1979) *Reliability and Validity Assessment*. Vol. 17. Thousand Oaks, CA: Sage.

Church A. T., Alvarez J. M., Mai N. T., French B. F., Katigbak M. S., Ortiz F. A. (2011) Are Cross-Cultural Comparisons of Personality Profiles Meaningful? Differential Item and Facet Functioning in the Revised NEO Personality Inventory. *Journal of Personality and Social Psychology*, vol. 101, no 5, pp. 1068–1089.

Davidov E., Meuleman B., Cieciuch J., Schmidt P., Billiet J. (2014) Measurement Equivalence in Cross-National Research. *Annual Review of Sociology*, vol. 40, pp. 55–75.

Dumas D., Alexander P. A. (2016) Calibration of the Test of Relational Reasoning. *Psychological Assessment*, vol. 28, no 10, pp. 1303–1319.

Ercikan K., Lyons-Thomas J. (2013) Adapting Tests for Use in Other Languages and Cultures. APA Handbook of Testing and Assessment in Psychology (ed. K. F. Geisinger), Washington: American Psychological Association, vol. 3, pp. 545–569.

Goldhammer F., Martens T., Christoph G., Lüdtke O. (2016) *Test-Taking Engagement in PIAAC.* Paris: OECD.

Hambleton R. K. (1993) Translating Achievement Tests for Use in Cross-National Studies. *European Journal of Psychological Assessment*, vol. 9, no 1, pp. 57–68.

Hambleton R. K., Kanjee A. (1995) Increasing the Validity of Cross-Cultural Assessments: Use of Improved Methods for Test Adaptations. *European Journal of Psychological Assessment*, vol. 11, no 3, pp. 147–157.

Hanushek E. A., Woessmann L. (2008) The Role of Cognitive Skills in Economic Development. *Journal of Economic Literature*, vol. 46, no 3, pp. 607–668.

Holland P. W., Wainer H. (1993) *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Jamelske E. (2009) Measuring the Impact of a University First-Year Experience Program on Student GPA and Retention. *Higher Education*, vol. 57, no 3, pp. 373–391.

Jeon M., De Boeck P. (2016) A Generalized Item Response Tree Model for Psychological Assessments. *Behavior Research Methods*, vol. 48, no 3, pp. 1070–1085.

Kankaraš M., Moors G. (2014) Analysis of Cross-Cultural Comparability of PISA 2009 Scores. *Journal of Cross-Cultural Psychology*, vol. 45, no 3, pp. 381–399.

Kardanova E., Loyalka P., Chirikov I. et al. (2016) Developing Instruments to Assess and Compare the Quality of Engineering Education: The Case of China and Russia. *Assessment & Evaluation in Higher Education,* vol. 41, no 5, pp. 770–786.

Kuzminov Ya., Sorokin P., Froumin I. (2019) Obshchie i spetsialnye navyki kak komponenty chelovecheskogo kapitala: novye vyzovy dlya teorii i praktiki obrazovaniya [Generic and Specific Skills as Components of Human Capital: New Challenges for Education Theory and Practice]. *Foresight and STI Governance*, no 13 (S2), pp. 19–41.

Lenz S. A., Soler I. G., Dell'Aquilla J., Uribe P. M. (2017) Translation and Cross-Cultural Adaptation of Assessments for Use in Counseling Research. *Measurement and Evaluation in Counseling and Development*, vol. 50, no 4, pp. 224–231.

Liu O. L., Rios J. A., Borden V. (2015) The Effects of Motivational Instruction on College Students' Performance on Low-Stakes Assessment. *Educational Assessment*, vol. 20, no 2, pp. 79–94.

Loyalka P., Liu O. L., Li G. et al. (2019) Computer Science Skills Across China, India, Russia, and the United States. *Proceedings of the National Academy of Sciences*, vol. 116, no 14, pp. 6732–6736.

Magis D., Facon B. (2013) Item Purification Does Not Always Improve DIF Detection: A Counterexample with Angoff's Delta Plot. *Educational and Psychological Measurement*, vol. 73, no 2, pp. 293–311.

Marginson S. (2019) Limitations of Human Capital Theory. *Studies in Higher Education*, vol. 44, no 2, pp. 287–301.

Martin M. O., Mullis I. V.S., Hooper M. (eds) (2016) *Methods and Procedures in TIMSS2015*. Available at: http://timssandpirls.bc.edu/publications/timss/2015-methods.html (accessed 10 April 2020).

Meredith W. (1993) Measurement Invariance, Factor Analysis and Factorial Invariance. *Psychometrika*, vol. 58, no 4, pp. 525–543.

Meyer J. W., Ramirez F. O., Soysal Y. N. (1992) World Expansion of Mass Education, 1870–1980. *Sociology of Education*, vol. 65, no 2, pp. 128–149.

Mislevy R. J. (1988) Exploiting Collateral Information in the Estimation of Item Parameters. *ETS Research Report Series*, vol. 2, pp. 1–31.

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.

OECD (2017) PISA 2015 Technical Report. Available at: https://www.oecd.org/pisa/data/2015-technical-report/ (accessed 10 April 2020).

OECD (2016) *PISA 2018 Translation and Adaptation Guidelines*. Available at: https://www.oecd.org/pisa/pisaproducts/PISA-2018-translation-and-adaptation-guidelines.pdf (accessed 10 April 2020).

OECD (2015) *PISA 2018 Technical Standards*. Available at: https://www.oecd.org/pisa/pisaproducts/PISA-2018-Technical-Standards.pdf (accessed 10 April 2020).

OECD (2010) Education at a Glance 2010: OECD Indicators. www.oecd.org/edu/eag2010

Rauhvargers A. (2011) *Global University Rankings and Their Impact: EUA Report on Rankings 2011.* Brussels: European University Association.

Rogers H. J., Swaminathan H. (1993) A Comparison of Logistic Regression and Mantel-Haenszel Procedures for Detecting Differential Item Functioning. *Applied Psychological Measurement*, vol. 17, no 2, pp. 105–116.

Roussos L., Stout W. (1996) A Multidimensionality-Based DIF Analysis Paradigm. *Applied Psychological Measurement*, vol. 20, no 4, pp. 355–371.

Schmitt N., Kuljanin G. (2008) Measurement Invariance: Review of Practice and Implications. *Human Resource Management Review*, vol. 18, no 4, pp. 210–222.

Schneider S. L. (2009) *Confusing Credentials: The Cross-Nationally Comparable Measurement of Educational Attainment* (PhD Thesis). Oxford: Oxford University.

Schultz T. W. (1961) Investment in Human Capital. *The American Economic Review*, vol. 51, no 1, pp. 1–17.

Shavelson R. J., Zlatkin-Troitschanskaia O., Mariño J. P. (2018) International Performance Assessment of Learning in Higher Education (iPAL): Research and Development. *Assessment of Learning Outcomes in Higher Education* (eds O. Zlatkin-Troitschanskaia, M. Toepper, H. A. Pant, C. Lautenbach, C. Kuhn), Springer, pp. 193–214.

Tijmstra J., Bolsinova M. A., Jeon M. (2018) Generalized Mixture IRT Models with Different Item-Response Structures: A Case Study Using Likert-Scale Data. *Behavior Research Methods*, vol. 50, no 6, pp. 2325–2344.

Tremblay K. (2013) OECD Assessment of Higher Education Learning Outcomes (AHELO): Rationale, Challenges and Initial Insights from the Feasibility Study. *Modeling and Measuring Competencies in Higher Education* (eds S. Blömeke, O. Zlatkin-Troitschanskaia, Ch. Kuhn, Ju. Fege), Rotterdam: Brill Sense, pp. 113–126.

Van de Vijver F., Tanzer N. K. (2004) Bias and Equivalence in Cross-Cultural Assessment: An Overview. *European Review of Applied Psychology*, vol. 54, no 2, pp. 119–135.

Van der Linden W. J. (2007) A Hierarchical Framework for Modeling Speed and Accuracy on Test Items. *Psychometrika*, vol. 72, no 3, pp. 287–308.

Wang W. C., Wilson M. (2005) The Rasch Testlet Model. *Applied Psychological Measurement*, vol. 29, no 2, pp. 126–149.

Wang W. C., Shih C. L., Sun G. W. (2012) The DIF-free-then-DIF Strategy for the Assessment of Differential Item Functioning. *Educational and Psychological Measurement*, vol. 72, no 4, pp. 687–708.

Wynd C. A., Schmidt B., Schaefer M. A. (2003) Two Quantitative Approaches for Estimating Content Validity. *Western Journal of Nursing Research*, vol. 25, no 5, pp. 508–518.

Zhu W., Ennis C. D., Chen A. (1998) Many-Faceted Rasch Modeling Expert Judgment in Test Development. *Measurement in Physical Education and Exercise Science*, vol. 2, no 1, pp. 21–39.

Zlatkin-Troitschanskaia O., Pant H. A., Greiff S. (2019) Assessing Generic and Domain-Specific Academic Competencies in Higher Education. *Zeitschrift für Pädagogische Psychologie*, vol. 33, no 2, pp. 91–93. ·

Zlatkin-Troitschanskaia O., Pant H. A., Lautenbach C., Molerov D., Toepper M., Brückner S. (2017) *Modeling and Measuring Competencies in Higher Education. Approaches to Challenges in Higher Education Policy and Practice*. Wiesbaden: Springer.