

УДК 37.01



А.Е.Иванова

*Младший научный сотрудник, Институт образования, Национальный исследовательский университет «Высшая школа экономики», г. Москва
E-mail: aeivanova@hse.ru*

Alina E. Ivanova
Junior Researcher, Institute of Education, National Research University Higher School of Economics, Moscow, Russia

ПРОБЛЕМА СОПОСТАВИМОСТИ РЕЗУЛЬТАТОВ В МЕЖДУНАРОДНЫХ СРАВНИТЕЛЬНЫХ ИССЛЕДОВАНИЯХ ОБРАЗОВАТЕЛЬНЫХ ДОСТИЖЕНИЙ

Международные сравнительные исследования образовательных достижений становятся движущей силой происходящих перемен в системе образования во многих странах, включая Россию. Крупнейшие международные исследования во многом служат примером эффективного проектирования и реализации изучения образовательных достижений в глобальном масштабе. Но сегодня все чаще проводятся и небольшие межстрановые исследования, предполагающие сравнение образовательных результатов, полученных с помощью различных языковых версий инструмента. Тем не менее, независимо от масштаба, ресурсов и числа вовлеченных участников, любые исследования, предполагающие проведение сравнений, требуют убедительных доказательств высокого качества каждого этапа своего жизненного цикла — от разработки инструментов до интерпретации полученных данных. В современной практике известны примеры, когда исследователям не удавалось обеспечить сопоставимость измерений, а значит, сравнение результатов оценивания оказывалось невозможным. Целью данной статьи является знакомство с международным опытом применения стратегий и методов обеспечения сопоставимости результатов сравнительных исследований. В работе рассмотрены основные проблемы и вызовы, с которыми сталкиваются организаторы международных сравнительных

Как цитировать статью: Иванова А. Е. Проблема сопоставимости результатов в международных сравнительных исследованиях образовательных достижений // Отечественная и зарубежная педагогика. 2018. Т. 1, № 2(48). С.68–81.

исследований, приведены процедуры оценки возможных угроз, а также предложены выработанные международным сообществом механизмы обеспечения сопоставимости данных подобных исследований. Только в случае строгого следования процедурам обеспечения качества международного исследования сделанные на его основе выводы, сравнения и интерпретации могут быть признаны надежными и справедливыми.

Ключевые слова: международные сравнительные исследования, адаптация, оценивание, сопоставимость, эквивалентность результатов, сравнения.

Международные исследования как глобальный феномен

Международные широкомасштабные сравнительные исследования (МСИ) в сфере образования, такие как, например, Международное исследование качества чтения и понимания текста (PIRLS), Международная программа по оценке образовательных достижений учащихся (PISA) или Программа международной оценки компетенций взрослых (PIAAC), производят значительные объемы объективных данных для исследователей и политиков. Сегодня МСИ до некоторой степени формируют то, как образование понимается и какую ценность оно имеет в странах-участницах [25].

Россия принимает участие в международных сравнительных образовательных исследованиях, проводимых Организацией экономического сотрудничества и развития (ОЭСР) и Международной ассоциацией по оценке образовательных достижений (IAE), с 1988 года [1]. В нашей стране успешно используются данные об образовательных результатах школьников, а также данные, получаемые из контекстных опросников родителей, учителей и директоров школ [2]. Эти обширные данные позволяют изучать эффекты различных факторов на образовательные результаты, оценивать взаимосвязи этих факторов и глубже понимать механизмы, лежащие в основе образовательных систем. Мы можем получить представление о содержании образовательных программ, системе институтов образования, структуре систем оценки образовательных результатов, квалификации педагогов и многом другом. И что наиболее важно, мы можем получить представление об уровне компетентности учащихся в той или иной области, о том, что они знают, как они могут это применять, и что они об этом думают, — на международном уровне.

На сегодняшний день в мировой практике накоплено немало примеров, когда международные сравнительные исследования стимулировали

происходящие в стране трансформации системы образования [9; 20]. Наиболее показательным является, пожалуй, пример Германии, где результаты PISA 2000 стали шокирующей новостью для образовательного сообщества, — по усредненному рейтингу страна оказалась на 20-м месте среди 32 стран ОЭСР. После волны критики, разразившейся в Германии относительно выводов и методов PISA, последовали образовательные реформы, которые привели к новой концептуализации всей школьной системы: от иного качества контрольно-измерительных инструментов и введения внутренних мониторингов в школах до изменения системы профессиональной подготовки учителей. Традиционные идеалы образования были переосмыслены, в стране стали больше ориентироваться на конкретные образовательные результаты [20].

Другой известный пример — Дания, которая демонстрировала результаты ниже среднего по ОЭСР в циклах PISA 2000, 2003, 2006 годов. Долгая общественная дискуссия внутри страны привела в конечном итоге к инициированию реформы общего обязательного образования: изменениям в содержании преподавания, особенно предметов естественнонаучного цикла, введению национального тестирования в компьютерной форме. Как отмечают исследователи внутри страны, важным эффектом недовольства Дании своим местом в международных сравнениях стало, как и в Германии, признание необходимости перемен [9].

Приведенные выше примеры показывают, насколько серьезно страны относятся к результатам МСИ. Обобщенная форма таких результатов — это чаще всего рейтинг, позволяющий идентифицировать страну или группу стран с самым высоким уровнем образовательных достижений. На основе таких рейтингов делается предположение, что образовательные практики, существующие в данной стране, — самые эффективные и потому заслуживают того, чтобы на них ориентировались другие участники [28]. Эти предположения, безусловно, требуют теоретического осмысления и эмпирической проверки. Существует ряд работ, призывающих исследователей и политиков относиться к этим предположениям осторожно и вдумчиво [12].

Тем не менее крупнейшие МСИ, такие как PIRLS или PISA, являются сегодня наиболее показательным примером эффективного проектирования и проведения исследований образовательных достижений (компетенций) на языках разных стран и культур. Однако в мире ежегодно проводятся международные сравнительные исследования, вовлекающие

не столь большое количество стран, ресурсов и внимания СМИ. Это могут быть различные совместные образовательные проекты и инициативы нескольких стран или образовательных организаций, предполагающие оценивание образовательных достижений учащихся. С другой стороны, даже при проведении внутренних национальных мониторингов могут существовать версии инструментов оценивания на различных языках, отличных от основного государственного языка. Или же речь может идти об экзаменах в учебные заведения, которые дают учащимся право выбрать язык, на котором сдавать экзамен. Эти примеры предполагают необходимость сравнения образовательных результатов участников, полученных с помощью культурно или лингвистически различающихся версий инструментов. А значит, разработчики языковых версий подобных исследований и инструментов должны ориентироваться на те же стандарты качества, что и крупные МСИ.

Ввиду того, что результаты проведения любых международных сравнительных исследований, независимо от их масштаба, влекут за собой серьезные последствия, они требуют убедительных доказательств высокого качества каждого этапа, начиная от разработки инструментов и заканчивая интерпретацией полученных данных. Целью нашей статьи является знакомство читателей с международным опытом применения стратегий и методов обеспечения сопоставимости результатов сравнительных исследований. Мы хотим рассмотреть базовые проблемы МСИ: как добиться оценок, сопоставимых для всех стран, с их различными языками и культурами? Какие механизмы обеспечивают принципиальную сопоставимость получаемых данных? Каким образом эту сопоставимость можно доказать?

Международные сравнительные исследования: основные задачи и вызовы

Началом существования МСИ в сфере образования можно считать 1958 год, когда исследователи ряда стран встретились в Институте образования ЮНЕСКО в Гамбурге, чтобы обсудить потенциал имевшихся на тот момент инструментов международных сравнений в сфере образования (в основном простых индикаторов, таких как уровень образования) и попытаться включить в них измерение образовательных результатов.

Основная идея состояла в том, чтобы провести такое исследование, которое включало бы объективные индикаторы образовательных резуль-

татов в виде потенциально сравнимых на межстрановом уровне тестов, что позволило бы изучить институциональные характеристики образовательных систем, обеспечивающих успешность учащихся в разных странах [27]. Проведенное в 12 странах в 1959–1962 годах пилотное исследование, целью которого было изучить саму возможность подобной амбициозной идеи, доказало ее потенциальную реализуемость. Тем не менее только в 90-е годы прошлого века международные исследования стали принимать ту форму, которую мы знаем сейчас, что было связано с развитием вычислительных технологий и теории педагогических измерений [8].

Международные широкомасштабные сравнительные исследования сталкиваются с большим количеством методологических вызовов, связанных с созданием качественного измерительного инструментария и его адаптацией для стран-участниц. В основном они связаны с переводом инструментов на языки культурных или языковых групп, а также с формированием выборки испытуемых.

Методологические проблемы сравнительных исследований чреваты повышенной вероятностью ошибочных выводов из их результатов. К примеру, выявленные в ходе исследования межстрановые различия могут быть обусловлены артефактом измерения, а не существенными различиями в измеряемом конструкте. Таким образом, фундаментальной методологической задачей международных сравнительных исследований является обеспечение сопоставимости результатов.

Концепция сопоставимости результатов в МСИ

В научной литературе можно встретить различные синонимы понятия сопоставимости, например инвариантность или сравнимость результатов. Говоря о том, что некоторая оценка сопоставима в разрезе различных культур или стран, мы имеем в виду, что она измеряет один и тот же конструкт одним и тем же образом в этих культурах или странах. В контексте международных сравнительных исследований сопоставимость результатов обеспечивается через достижение *эквивалентности выборок* и *эквивалентности измерений* в различных языковых и национальных версиях инструмента оценивания.

Обеспечение сопоставимости выборок

Вопрос формирования выборки участников исследования является исключительно важным элементом с точки зрения обеспечения его ка-

чества и получения интерпретируемых данных. В идеале выборка для МСИ должна репрезентировать всех представителей целевой аудитории исследования в рамках каждой страны. Ошибки выборки включают в себя ошибку охвата (т.е. неполный охват или излишнюю представленность какой-то группы в выборке) и ошибку отбора (т.е. ошибку, связанную с процедурами отбора участников) [29].

В частности, серьезной проблемой МСИ является включение или исключение из выборки некоторых специфических групп участников. Например, учеников с ограниченными возможностями здоровья. В некоторых странах для них предусмотрены отдельные школы, в других применяется инклюзивный подход к образованию. То же касается включения в выборку частных школ или наличия в некоторых образовательных системах отдельных треков в обучении. Сложности с отбором участвующих школ хорошо иллюстрирует пример исследования PISA, когда выбранные школы могут, но не обязаны участвовать в исследовании. Так, в PISA-2015 уровень участия выбранных школ варьировался от 65% до 85% в разных странах [22].

Возрастные характеристики выборки также являются предметом многочисленных дискуссий в исследовательском сообществе. К примеру, возраст учеников 8-х классов в исследовании TIMSS сильно варьируется по странам: в Англии дети идут в школу с 4 лет, а в России — с 7 лет. С другой стороны, в исследовании PISA возраст участников должен быть в районе 15 лет, но это значит, что дети к этому моменту могут быть в 9-м, 10-м или в 11-м классе в разных странах.

Обеспечение эквивалентности измерений

Эквивалентность измерений включает в себя три важных компонента: 1) эквивалентность конструкта, 2) эквивалентность инструмента, 3) эквивалентность процедуры исследования [12].

Эквивалентность конструкта. Чтобы в контексте различных стран и культур, принимающих участие в МСИ, можно было адекватно измерить искомый конструкт, он должен быть схожим образом определен и концептуально эквивалентен в этих культурах [12]. Данные, полученные из МСИ, могут быть несопоставимы, если в реальности в разных странах оценивались разные конструкты. Было бы странно полагать, что некоторый инструмент, созданный в одной культуре, на основе ее ценностей и представлений, может быть абсолютно точно перенесен

в контекст другой страны и культуры. В научной литературе подчеркивается, что в образовании и психологии конструкты почти всегда содержат в себе культурные компоненты. Следовательно, для обеспечения сопоставимости результатов международных исследований культурные различия конструктов должны быть тщательно оценены в процессе планирования, апробации и проведения измерения.

Эквивалентность инструмента. Для проведения МСИ, как правило, создаются различные языковые версии инструмента измерения. С их культурными, содержательными, лингвистическими компонентами связана эквивалентность инструмента. В результате трансформации инструмента из одной языковой версии в другую могут возникать искажения в общем дизайне инструмента, в формулировке отдельных вопросов, в порядке вопросов или заданий, в предлагаемых вариантах ответа и др. [11]. Исследования показывают, что язык инструмента значительно влияет на то, как участники исследования отвечают на один и тот же вопрос [14]. Следует отметить, что именно ошибки и искажения, возникающие при переводе языковых версий инструмента, являются наиболее частой проблемой МСИ [7; 11].

Эквивалентность процедуры. Достижение эквивалентности процедур в МСИ предполагает, что а) каждый шаг в процессе администрирования процедуры исследования одинаков, независимо от времени и места проведения, б) формат тестирования является единообразным [14]. Ряд недавних исследований показывает, что к оценке эквивалентности процедур следует также отнести проблему стиля ответов участников исследования. Последние могут стать серьезным источником вариации в результатах международных исследований [15]. Стили ответов могут быть связаны с культурно-обусловленными различиями в мотивации участия в исследовании, различиями в стратегиях решения заданий (например, склонности к угадыванию), социальной желательности и пр. [11; 14].

Важно отметить, что любой международный инструмент оценивания, предполагающий проведение сравнений в разрезе различных стран и культур, неизбежно содержит в себе посторонний по отношению к искомому конструкту культурно обусловленный компонент. Сопоставимость данных МСИ не следует принимать как данность, она должна быть заранее продумана и в дальнейшем доказана, если целью исследования является сравнительная интерпретация баллов и обобщение результатов на различные страны и культуры.

Стандарты в адаптации инструментов МСИ

Сопоставимость оценок международного сравнительного исследования обеспечивается, в первую очередь, с помощью специально разработанных процедур адаптации инструментов оценивания для языков стран и культур, принимающих участие в МСИ. Ранее под адаптацией понимался простой перевод инструмента с одного языка на другой [26]. Однако современный подход предполагает, что адаптация — это сложный многоступенчатый процесс, позволяющий осуществить трансфер инструмента измерения, созданного в рамках одной культуры, в контекст и на язык другой культуры [17]. Такое широкое понимание адаптации связывает ее буквально со всеми видами деятельности, возникающими в ходе проведения международного сравнительного исследования.

Научно-исследовательские организации в сфере оценивания предлагают различные руководства и рекомендации, призванные обеспечить качество адаптации при проведении МСИ [4; 17]. Ведущие исследователи также предлагают свои решения по узкоспециальным проблемам МСИ, например контролю качества перевода [26], мерам обеспечения единообразия и согласованности механизмов администрирования исследования на местах [18] или методам эмпирического анализа сопоставимости полученных результатов [23].

Сегодня мировым сообществом признаны *методические рекомендации по переводу и адаптации тестов*, предложенные Международной тестовой комиссией (International Test Commission). В актуальной версии рекомендаций, которая вышла в 2016 году, содержится шесть тематических разделов, охватывающих основные этапы процедуры адаптации [17].

В первом разделе — «Предварительные условия» — подчеркивается, что решению о проведении адаптации должны предшествовать исследования по оценке понимания измеряемого конструкта в разных странах, по оценке специфических культурных эффектов, языковых и национальных особенностей испытуемых в разных странах. Во втором разделе — «Руководство по разработке тестов» — основное внимание уделяется фактическому процессу адаптации теста — созданию инструментов на языках стран — участниц исследования, включая сбор доказательств о том, что язык и терминология являются приемлемыми для всех участников; что формат заданий, процедуры и правила тестирования им знакомы; что содержание заданий им известно и др. Третий раздел — «Подтверждение (эмпирический анализ)» — включает в себя

те рекомендации, которые связаны со сбором эмпирических доказательств эквивалентности, надежности и достоверности инструмента на разных языках. Раздел «Администрирование» посвящен организации процедур международного исследования. Пятый раздел — «Шкалы и интерпретация» — обсуждает проблему межгрупповых различий в баллах и собственно сравнение результатов. Наконец, шестой раздел, «Документация», призывает к четкой и технически выверенной фиксации всего процесса адаптации, а также освещает вопросы информирования пользователей измерительных инструментов.

Анализ сопоставимости результатов исследований в современной исследовательской практике

Как показывают многочисленные исследования, строгое соблюдение всех рекомендаций адаптации абсолютно необходимо, но даже оно не гарантирует, что в итоге проведенного международного исследования результаты всех стран окажутся сопоставимы [19]. Именно поэтому обычной практикой сегодня являются не только исследования, проводимые в процессе адаптации, но и дополнительные независимые исследования по конкретным узким темам, которые проводятся специалистами уже после открытия и публикации данных. В ходе всего комплекса таких исследований должны быть собраны убедительные эмпирические свидетельства эквивалентности измерений. Сегодня для этой цели создан большой арсенал количественных методов, включая, например, средства и методы современной теории тестирования (IRT), конфирматорного факторного анализа, а также качественные методы для решения отдельных задач.

Отметим, что методология современной теории тестирования позволяет в принципе прогнозировать и оценивать качество проводимых измерений. Выбор математической модели, описывающей взаимосвязи измеряемого конструкта, характеристик инструмента и первичных результатов выборки, психометрический анализ качества заданий и инструмента, анализ структуры измеренного конструкта, создание шкал и итоговых оценок — в современных МСИ в сфере образования все это производится на базе IRT. Обоснование эквивалентности результатов МСИ является важной частью общего исследовательского процесса.

Для подтверждения эквивалентности измерений в рамках IRT прежде всего оценивается адекватность функционирования заданий и инструмента в целом внутри каждой страны. Затем проводится оценка

структуры (размерности) измеренного конструкта и ее схожести во всех странах. Наконец, проводится анализ возможных искажений в заданиях в различных языковых версиях инструментов [23].

Сигналом наличия в результатах искажений является, в частности, различное функционирование заданий (Differential Item Functioning, DIF) или инструмента в целом (Differential Test Functioning). Иногда в инструментах МСИ присутствуют задания, которые оказываются несправедливо легче или сложнее для какой-то группы участников (или для целой страны), хотя по своим способностям эта группа не отличается от других [23]. К примеру, в отчете ОЭСР было показано, что несколько заданий из области анализа данных в PISA-2006 демонстрировали искажения в отношении участников из России и Словакии [21]. Это были задания, материал которых еще не изучался большинством учеников в школах этих двух стран. Причиной различного функционирования заданий могут быть также формат заданий, специфические языковые формулировки, различная длина предложений и текстов и др. Очевидно, что задания с DIF вносят существенные искажения в результаты. Соответственно, необходимы исследования, демонстрирующие, что для шкалирования и получения баллов по результатам МСИ используются задания, свободные от DIF.

Другой подход к оценке эквивалентности измерений — это мультигрупповой конфирматорный факторный анализ, который используется исследователями для одновременной оценки сопоставимости конструктов и возможных различий между языковыми и культурными группами [24]. Применяя данный метод, исследователь на базе теоретической модели инструмента строит статистическую (факторную) модель и проверяет, действительно ли она демонстрирует одинаковые параметры во всех релевантных группах.

Проведение эмпирических исследований, которые доказывают сопоставимость данных международных исследований, является безусловным требованием качественного МСИ. «Побочным эффектом» таких исследований является периодическое обнаружение погрешностей в переводе языковых версий инструментов [6], в функционировании заданий для отдельных стран или групп стран [16], различий в измеренном конструкте [10], что отражается на репутации МСИ. В то же время такие исследования позволяют увидеть новые перспективы в развитии и совершенствовании методологии МСИ и более ответственно и рационально относиться к их выводам.

МСИ: ответственность за результаты независимо от масштаба

Международные широкомасштабные исследования стали довольно привычной частью современной реальности в образовании. Число стран, вовлеченных в такие сравнения, постоянно растет. Исследователи выделяют различные причины, по которым страны принимают решение об участии в МСИ. Например, сбор данных в целях образовательной политики, создание технологической основы для развития национальных систем образования, получение финансовой помощи (как предпосылка или подтверждение эффективности использования внешних займов), поддержка и развитие отношений с другими странами (сигнал о принадлежности к определенному сообществу, «клубу»), ответы на актуальную политическую повестку дня в стране («скандализация» результатов в рейтингах как двигатель реформ), инвестиции в экономический рост (оценка человеческого капитала страны как индикатор экономической конкурентоспособности), и, наконец, информирование образовательного и исследовательского сообщества о тенденциях и содержании образования в разных странах [3; 27]. Страны относятся к результатам МСИ крайне серьезно, и влияние, которое они оказывают на образовательную политику, свидетельствует о политическом доверии стран к практике сравнения образовательных достижений, к лежащей в основе МСИ методологии измерения и сопоставления результатов. Организации, проводящие МСИ, должны это доверие оправдывать, в том числе через проведение исследований, развитие измерительных технологий, совершенствование стандартов разработки, адаптации и организации сравнительных исследований.

В данной статье были представлены наработанные за годы проведения крупных международных сравнительных образовательных исследований практики, позволяющие обеспечивать и доказывать сопоставимость их результатов в условиях многочисленных методологических вызовов, с которыми сталкиваются МСИ. В современном мире, когда глобальное или региональное сотрудничество в сфере образовательных исследований, заимствование инструментов оценки и сопоставления образовательных результатов является обычной практикой, крайне важно иметь представление о возможностях и трудностях проведения сравнительных исследований, независимо от их масштаба.

Существует ряд примеров, когда сравнение результатов оценивания оказывалось невозможным, когда не удавалось обеспечить сопостави-

мость измерений, в частности из-за того, что адаптация инструмента для использования в другой языковой или культурной группе была реализована недостаточно качественно. К примеру, в исследовании сопоставимости версий национального экзамена для приема в вузы Израиля на иврите и русском языке было выявлено, что треть заданий демонстрировала несправедливое функционирование в пользу участников, сдававших экзамен на русском языке. После публикации исследования инструмент был серьезно переработан [5]. Другой пример — адаптация американского выпускного экзамена для аттестации медицинских сестер в Канаде для английских и французских языков. Проведенное исследование выявило серьезные нарушения процедуры адаптации и фактическую несопоставимость языковых версий экзамена, что поставило вопрос об отказе от инструмента [13].

Проведение сравнений образовательных достижений, будь то крупномасштабные международные исследования или локальные проекты, предполагающие сопоставление результатов различных языковых и культурных групп, требует серьезных ресурсов и усилий. Базовое предположение о том, что результаты оценивания в принципе сравнимы, должно быть тщательно проверено и доказано исследователями. Только в этом случае сделанные на основе исследований выводы, сравнения, интерпретации могут быть признаны надежными, валидными и справедливыми, а значит, по-настоящему полезными для всех участников.

Исследование выполнено за счет гранта Российского научного фонда (проект № 16–18–10401).

PROBLEM OF COMPARABILITY OF RESULTS IN INTERNATIONAL COMPARATIVE STUDIES OF EDUCATIONAL ACHIEVEMENTS

International comparative studies of educational achievements became a driver of the ongoing changes in the educational system in many countries, including Russia. The largest international research is in many ways an example of effective design and implementation of the study of educational achievements on a global scale. However, today many studies of a lesser scale are being conducted. These studies also involve comparisons of the educational results obtained with various language versions of the assessment instrument. Nevertheless, regardless of the scale, resources and the number of participants, the studies that suppose comparisons require convincing evidence of the high quality on each stage of their life cycle from developing tools to interpreting their data. In modern practice, the examples are known, when researchers were not able to ensure comparability of measurements, and therefore, comparisons of the evaluation results were impossible. The purpose of the current paper is to show international experience in applying strategies and methods to ensure comparability of the results for comparative studies. The article examines the main problems and challenges

that the organizers of international comparative studies face; it proposes procedures for assessing possible threats; and, finally, it considers the mechanisms developed by the international community to ensure comparability of these studies. Only in the case of strict adherence to the procedures for ensuring the quality of an international study, the conclusions, comparisons and interpretations made on its basis can be recognized reliable and fair.

Keywords: international comparative studies, adaptation, evaluation, comparability, equivalence of results, comparisons.

Литература/References

1. Болотов, В. А., Вальдман, И. А., Ковалёва, Г. С. и др. Российская система оценки качества образования: главные уроки // Качество образования в Евразии. — 2013. — № 1. С 85–121.
2. Ковалева Г. С. Финансовая грамотность как составляющая функциональной грамотности: международный контекст // Отечественная и зарубежная педагогика. 2017. Т. 1, № 2 (37). С. 31–43.
3. Addey, C., Sellar, S., Steiner-Khamsi et al. The rise of international large-scale assessments and rationales for participation // Compare: A Journal of Comparative and International Education. 2017. Т. 47, № 3. С. 434–452.
4. AERA, APA, NCME. Standards for educational and psychological testing. Amer Educational Research Assn, 2014, С. 57.
5. Allalouf A., Hambleton R. K., Sireci S. G. Identifying the causes of DIF in translated verbal items // Journal of educational measurement. 1999. Т. 36, № 3. С. 185–198.
6. Asil M., Brown G. T. L. Comparing OECD PISA reading in English to other languages: Identifying potential sources of non-invariance // International Journal of Testing. 2016. Т. 16, № 1. С. 71–93.
7. Benitez I., Padilla J. L. Analysis of nonequivalent assessments across different linguistic groups using a mixed methods approach: Understanding the causes of differential item functioning by cognitive interviewing // Journal of Mixed Methods Research. 2014. Т. 8, № 1. С. 52–68.
8. Braun H. Prospects for the future: A framework and discussion of directions for the next generation of international large-scale assessments // The role of international large-scale assessments: Perspectives from technology, economy, and educational research. Springer Netherlands, 2013. С. 149–160.
9. Dolin J., Krogh L. B. The relevance and consequences of PISA science in a Danish context // International Journal of Science and Mathematics Education. 2010. Т. 8, № 3. С. 565–592.
10. Ercikan K., Koh K. Examining the construct comparability of the English and French versions of TIMSS // International Journal of Testing. 2005. Т. 5, № 1. С. 23–35.
11. Ercikan, K., & Lyons-Thomas, J. Adapting tests for use in other languages and cultures. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, et al. (Eds.). **APA handbook of testing and assessment in psychology. Testing and assessment in school psychology and education.** Washington, DC, US: American Psychological Association. 2013. Т. 3, С. 545–569.
12. Ercikan K., Roth W. M., Asil M. Cautions about inferences from international assessments: The case of PISA 2009 // Teachers College Record. 2015. Т. 117, № 1. С. 1–28.
13. Hall L. M. G., Lalonde M., Kashin J. People are failing! Something needs to be done: Canadian students' experience with the NCLEX-RN // Nurse education today. 2016. Т. 46. С. 43–49.
14. Hambleton R. K. Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures // Adapting educational and psychological tests for cross-cultural assessment. 2005. Т. 1. С. 3–38.
15. He J., Van de Vijver F. J. R. Response styles in factual items: Personal, contextual and cultural correlates // International Journal of Psychology. 2016. Т. 51, № 6. С. 445–452.
16. Huang X., Wilson M., Wang L. Exploring plausible causes of differential item functioning in the PISA science assessment: language, curriculum or culture // Educational Psychology. 2016. Т. 36, № 2. С. 378–390.
17. International Guidelines for Test Use // ITC [International Test Commission], 2016 [Электронный ресурс]. URL: www.InTestCom.org. Дата обращения: 11.08.2017. // International Journal of Testing.
18. Jowell, R., Roberts, C., Fitzgerald, R., & Eva, G. Measuring attitudes cross-nationally: Lessons

from the European Social Survey. London: Sage, 2007. С. 18, 117.

19. *Laschke C., Blömeke S.* Measurement of job motivation in TEDS-M: testing for invariance across countries and cultures // *Large-scale Assessments in Education*. 2016. Т. 4, № 1. С. 16.

20. *Niemann D., Martens K., Teltemann J.* PISA and its consequences: Shaping education policies through international comparisons // *European Journal of Education*. 2017. Т. 52, № 2. С. 175–183.

21. OECD. Learning Mathematics for Life. A perspective from PISA. OECD Publishing, 2010 [Электронный ресурс]. URL: <http://www.oecdbookshop.org/browse.asp?pid=title-detail&lang=en&ds=&ISBN=9789264075009>. Дата обращения: 09.09.2017

22. OECD. PISA 2015. Technical report. OECD Publishing, 2015 [Электронный ресурс]. URL: <http://www.oecd.org/pisa/data/2015-technical-report>. Дата обращения: 12.09.2017

23. *Oliveri, M. E., Olson, B. F., Ercikan, K., & Zumbo, B. D.* Methodologies for investigating item- and test-level measurement equivalence in international large-scale assessments // *International Journal of Testing*. 2012. Т. 12, № 3. С. 203–223.

24. *Putnick D. L., Bornstein M. H.* Measurement invariance conventions and reporting: The state of the art and future directions for psychological research // *Developmental Review*. 2016. Т. 41. С. 71–90.

25. *Sellar S., Lingard B.* The OECD and the expansion of PISA: New global modes of governance in education // *British Educational Research Journal*. 2014. Т. 40, № 6. С. 917–936.

26. *Sperber A. D., Devellis R. F., Boehlecke B.* Cross-cultural translation: methodology and validation // *Journal of cross-cultural psychology*. 1994. Т. 25, № 4. С. 501–524.

27. *Strietholt R., Scherer R.* The Contribution of International Large-Scale Assessments to Educational Research: Combining Individual and Institutional Data Sources // *Scandinavian Journal of Educational Research*. 2018. № 62, Т. 3 С. 368–385.

28. *Steiner-Khamsi, G., Waldow, F.* (2012). *World Yearbook in Education 2012: Policy Borrowing and Lending in Education*. New York, NY: Routledge. P. 46.

29. *Wu M.* Measurement, sampling, and equating errors in large-scale assessments // *Educational Measurement: Issues and Practice*. 2010. Т. 29, № 4. С. 15–27.