Изучение возможности проведения межстранового сравнительного исследования навыка чтения у учащихся на входе в школу

А. Е. Иванова, Е. Ю. Карданова

Статья поступила в редакцию в мае 2020 г.

Иванова Алина Евгеньевна
научный сотрудник Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики».
E-mail: aeivanova@hse.ru
Карданова Елена Юрьевна
кандидат физико-математических наук, доцент, ординарный профессор, директор Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики».
E-mail: ekardanova@hse.ru

Адрес: 101000, Москва, Потаповский пер., 16, стр. 10.

Аннотация. Первые годы обучения, когда ребенок только учится читать, критически важны для дальнейшего образования и развития. Межстрановые сравнительные исследования развивающегося навыка чтения имеют значительный потенциал с исследовательской и практической точки зрения, поскольку дают исследователям, практикам и политикам информацию о возможностях и перспективах для развития детей младшего школьного возраста в разных странах, условиях и контекстах. Публикаций такого рода немного, а исследований, в которых принимали бы участие русскоязычные дети на этапе поступления в школу, нет.

На примере данных, полученных с помощью двух языковых версий инструмента iPIPS на репрезентативных выборках первоклассников из Республики Татарстан и Шотландии, проанализирована возможность сопоставления результатов оценивания раннего навыка чтения у детей, начинающих школьное обучение в странах, которые различаются языком, культурой, возрастом начала школьного обучения.

С целью изучения возможных способов сопоставления результатов в условиях отсутствия общей метрической шкалы для оценивания выборки детей из разных стран проведены два исследования: в первом метод экспертного ранжирования применен для установления соответствия уровней развития чтения на русском и английском языках, а во втором полученная уровневая модель используется для установления пороговых баллов (бенчмарок) результатов оценивания учащихся в двух странах.

Ключевые слова: международные сравнительные исследования, начальная школа, экспертная оценка, метод попарных сравнений, Раш-подход.

DOI: 10.17323/1814-9545-2020-4-8-36

В изучении развития ребенка возрастает роль крупномасштабных внутри- и межстрановых исследований (МСИ). Современные МСИ позволяют верифицировать, уточнять и совершенствовать существующие теории развития [Peña, 2007; Shuttleworth-Edwards et al., 2004], они являются важным источником данных о предикторах, определяющих успешность обучения детей в разных странах, условиях, социальных и культурных контекстах [Ainley, Ainley, 2019; Carnoy et al., 2016; Caro, Cortés, 2012].

Интерес к МСИ во всем мире подтверждается стремительным ростом их числа с начала 2000-х годов. Например, количество участников PISA (*Programme for International Student Assessment*) выросло с 43 в 2000-м до 80 в 2018 г. [Liu, Steiner-Khamsi, 2020]. Исследователи отмечают, что правительства все большего числа стран стараются следовать логике МСИ в своей внутренней образовательной политике, стремясь добиться фиксированных, прогнозируемых, измеримых результатов [Espeland, 2015; Liu, Steiner-Khamsi, 2020].

Результаты международных исследований в сфере дошкольного и раннего школьного образования, и в частности проекты, посвященные ранней читательской грамотности, имеют особую значимость для исследователей и политиков. Она обусловлена повышающейся ролью читательской грамотности в современном обществе в целом, критической важностью именно первых лет обучения чтению для дальнейшего развития, необходимостью разумного расходования образовательных ресурсов, а также желанием исследователей и политиков принимать решения на доказательной основе, изучив опыт и лучшие практики других стран [Suggate, 2009]. Несмотря на то что каждая страна разрабатывает и реализует собственные образовательные цели и программы, ей необходимы внешние, международные ориентиры и информация о новых возможностях и перспективах для развития детей младшего школьного возраста [Buzhardt et al., 2019].

Сама возможность проведения научных исследований и сравнительного анализа с целью совершенствования образовательной политики существует только при условии надежных и валидных измерений в МСИ. Важнейшую роль в обеспечении валидности интерпретации результатов, полученных с помощью версий инструментов, принадлежащих разным странам, языкам или культурам, играет процедура адаптации. Научно-исследовательские организации, работающие в сфере оценивания, предлагают рекомендации относительно процедур, призванных обеспечить качество адаптации при проведении МСИ [American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014; Leong et al., 2016]. Цель этих процедур состоит в том, что-

бы максимально приблизиться к главному условию дальнейшего использования данных — обеспечению сопоставимости результатов оценивания.

Международная сопоставимость результатов МСИ возможна только если измерения, проводимые с помощью инструментов на разных языках, будут эквивалентны. Концепция эквивалентности измерений предполагает обеспечение и последующее эмпирическое обоснование 1) эквивалентности конструкта; 2) эквивалентности инструмента; 3) эквивалентности процедуры исследования [Ercikan, Roth, Asil, 2015]. Таким образом, чтобы свести к минимуму возможную культурно-языковую обусловленность результатов МСИ, используемые в них процедуры и методы разработки инструментов, а также валидизации результатов оценивания должны гарантировать оценку измеряемого поведения (навыка, компетенции, любого другого конструкта), свободную от эффектов других переменных (национальной или этнической принадлежности, социально-экономического статуса и др.).

Инструменты, призванные оценивать навык чтения, адаптировать на языки других стран и культур особенно трудно. Даже самые авторитетные исследования, такие как PIRLS (*Progress in International Reading Literacy Study*) и PISA, сталкиваются с проблемой несопоставимости результатов для отдельных языковых версий [Goodrich, Ercikan, 2019; Kreiner, Christensen, 2014]. При этом инструментов, которые позволяли бы на международном уровне сопоставлять данные о развивающемся навыке чтения у детей, только обучающихся читать, практически нет.

Так, международное исследование навыков чтения PIRLS ориентировано на учащихся, завершающих обучение в начальной школе, когда дети уже умеют читать. В недавно запущенном первом исследовании ОЭСР для детей на входе в начальную школу IELS (*Early Learning and Child Well-being*), в котором приняли участие пока только три страны, ранняя предчитательская грамотность измеряется через оценку понимания услышанного, фонематического восприятия и словарного запаса [OECD, 2020]. В другом авторитетном международном проекте, оценивающем чтение в начальной школе, — EGRA (*Early Grade Reading Assessment*) — задача международного сопоставления в принципе не ставится, инструмент, созданный на английском языке и адаптированный на языки принимающих стран-участниц, используется только для внутренних целей [Dubeck, Gove, 2015]. По мнению исследователей, развивающийся навык чтения крайне трудно оценить в кросс-культурном разрезе именно потому, что, когда дети только учатся читать, влияние языка в оценивании очень сильно [Ercikan, Roth, Asil, 2015].

В данной работе будет сделана попытка с помощью инструмента iPIPS (*international Performance Indicators in Primary Schools*) [Tymms, 1999] сопоставить развитие навыка чтения

у учащихся на входе в школу в двух странах — России и Великобритании. Инструмент iPIPS изначально создавался на английском языке, но в настоящее время используется в разных странах, включая не только англоязычные Австралию и Новую Зеландию, но и, к примеру, Германию, Бразилию, а также Россию [Карданова и др., 2018; Bartholo et al., 2019; Tymms et al., 2014; Vidmar et al., 2017].

При создании русскоязычной версии инструмента iPIPS было очевидно, что часть инструмента по оцениванию чтения у детей не может быть адаптирована, а должна быть только локализована. Локализация — это процесс преобразования некоторого продукта таким образом, чтобы в нем учитывалась культурная и языковая специфика целевой аудитории (страны, региона и т. д.) [Esselink, 2000]. Основное отличие локализации от адаптации заключается в том, что прямое международное сопоставление результатов тестирования между странами при локализации не проводится.

Необходимость локализации связана с существенными структурными различиями английского и русского языков, среди которых особенно важны номиноцентризм русского языка и вербоцентризм английского языка, несовпадение состава и функционала частей речи, фиксированный порядок слов в английском языке и ряд других. Из-за этих различий этапы становления языковой способности у англоговорящих и русскоговорящих детей не совпадают [Иванова, Карданова-Бирюкова, 2019]. А это, несомненно, сказывается на процессе формирования навыка чтения и, соответственно, его оценивания.

А. Е. Ивановой и К. С. Кардановой-Бирюковой [2019] описаны процедуры, использованные при локализации российской версии инструмента iPIPS в части чтения: показаны действия, призванные гарантировать эквивалентность измеряемого конструкта — это развивающийся навык чтения детей на входе в школу — на стадии разработки русскоязычной версии инструмента; обоснована невозможность достичь полной эквивалентности измерений; а также описаны процедуры сбора свидетельств конструктной валидности.

Первым шагом на пути создания русскоязычной версии iPIPS стал перевод и экспертиза заданий, созданных для британских детей и предназначенных для оценки их базовых навыков чтения. Перевод (прямой и обратный) осуществлялся в соответствии с рекомендациями Международной тестовой комиссии [Leong et al., 2016]. В часть iPIPS по чтению вошли несколько блоков заданий, соответствующих разным этапам теоретической модели чтения, разработанной авторами iPIPS: задания на понимание структуры текста; задания на распознавание букв; задания на распознавание графической оболочки слов; задания на механическое чтение (декодирование).

При создании русскоязычной версии iPIPS для оценивания навыков чтения локализовать эту часть заданий было достаточно легко. Однако разработка блока заданий для оценки сформированности навыка чтения с пониманием представляла сложную задачу. Блок заданий на чтение с пониманием представлял собой достаточно большие нарративные тексты со встроенными «ловушками» — пропусками, которые можно заполнить, выбрав одно из трех предложенных слов. «Ловушки» подбирались по разным принципам: графические, грамматические, фонетические, смысловые. Поскольку тексты, предъявляемые детям — носителям русского и английского языков, должны быть схожи по уровню сложности, а содержащиеся в них задания-«ловушки» должны быть направлены на проверку одинаковых навыков, локализация это блока потребовала значительных усилий.

Работа по локализации этой части теста iPIPS состояла из следующих этапов: сначала изучались лингвистические характеристики оригинального текста, потом подбирались эквивалентные «ловушки» в русском языке, наконец, моделировался русскоязычный текст, содержащий эти «ловушки» и содержательно близкий англоязычному оригиналу.

Несмотря на то что локализация международного инструмента на русский язык не предполагает прямого сопоставления индивидуальных результатов, существует запрос на сравнение базовых навыков детей на входе в школу. Эта задача может быть решена путем непрямого сопоставления. Под непрямым сопоставлением мы понимаем сопоставление результатов оценивания в кросс-культурном контексте не на индивидуальном уровне, но на уровне групп.

Цель данной статьи — исследовать возможность проведения международного сравнения развивающегося навыка чтения на входе в школу у детей из двух стран — России и Великобритании. Для ее реализации проведена серия из двух исследований: первое предполагает использование метода экспертного ранжирования для установления соответствия уровней развития чтения на русском и английском языках, а во втором полученная уровневая модель будет применена для установления пороговых баллов (бенчмарок) в данных оценивания учащихся в двух странах. В итоге этой серии исследований впервые станет возможным сопоставление результатов тестирования по чтению у первоклассников из России и Великобритании.

1. Первое исследование. Экспертное ранжирование как основа для сопоставления

В рамках изучения возможностей сопоставительного исследования проведена экспертиза конструкта — уровневой модели навыков чтения детей на русском и английском языках. Методологической основой экспертизы стал метод экспертного ранжирования.

Данная методология применялась в исследовании, в котором изучалась сопоставимость экзаменационных заданий в Англии для учащихся в возрасте около 14 лет [Bramley, 2005] для когорт разных лет и для экзаменов, оцененных разными группами экспертов. Метод ранжирования, предлагаемый автором данного исследования, предоставляет возможность сопоставить результаты выполнения тестовых заданий для версий инструментов, не удовлетворяющих в полной мере требованиям эквивалентности измерений. Он основан на комбинации экспертной оценки и математического моделирования полученных в ходе этой оценки данных.

Цель нашего первого исследования состоит в том, чтобы показать, как экспертные оценки могут быть использованы для построения шкалы чтения, для выявления иерархии заданий внутри измеряемого нами конструкта, для сравнения иерархии заданий в двух языковых версиях и, наконец, для формирования основания для построения бенчмарок.

1.1. Методология первого исследования

1.1.1. Отбор заданий

В исследовании использованы оригинальная версия теста по чтению iPIPS на английском языке, состоящая из 44 заданий [Merrell, Tymms, 2007], а также локализованная русскоязычная версия данного инструмента [Иванова, Карданова-Бирюкова, 2019], состоящая из 40 заданий. В двух версиях различалось количество заданий-«ловушек» в блоке заданий на чтение с пониманием. В русскоязычной версии их было 14, а в английской — 18.

1.1.2. Участники

Для проведения исследования были приглашены 12 экспертов, свободно владеющих английским языком: практикующие учителя или преподаватели английского языка, лингвисты, филологи. 11 экспертов — женщины, один эксперт — мужчина, все с образованием не ниже магистратуры, с опытом работы в данной сфере от двух лет до более десяти лет.

1.1.3. Процедура

От всех экспертов было получено предварительное согласие на участие в исследовании. Каждый из них получил две папки, содержащие два набора распечатанных заданий на русском и на английском языках, специальную инструкцию для проведения экспертизы, а также короткую анкету.

Каждое задание было представлено на отдельном листе бумаги в виде картинки, к нему прилагалась инструкция, с которой оно предъявлялось ребенку на старте обучения в школе в России и в Великобритании, и краткая информация о том, на оценивание чего данное задание направлено. Инструкция для эксперта также содержала сведения о процедуре тестирования. Помимо этого, экспертам была предоставлена ссылка на видеоролик, чтобы они могли посмотреть, как проходит тестирование.

Эксперты были уведомлены, что задания на русском языке в одной папке и на английском языке в другой папке разложены случайным образом. Их просили внутри каждой папки проранжировать задания от самого простого до самого сложного. В основание ранжирования эксперты должны были положить собственную холистическую оценку степени трудности каждого задания, произведенную исходя из их экспертных знаний и опыта. Экспертов просили распределить время таким образом, чтобы между ранжированием отдельно русской и английской версий прошло как минимум два дня.

1.1.4. Аналитический подход

Ранговые данные в полной мере обладают характеристиками, которые позволяют использовать их для моделирования с помощью семейства моделей Раша. Ранги — это наблюдения, предполагающие качественное упорядочивание объектов вдоль гипотетического континуума эксплицитно или имплицитно выраженной переменной [Linacre, 2006]. Единичный набор рангов — иначе говоря, результат единичного ранжирования — уже содержит достаточно информации для упорядочивания объектов. Если имеется несколько наборов рангов одних и тех же объектов, то информации может быть достаточно для построения интервальной шкалы для измерения расстояний между объектами. Использование интервальной шкалы позволяет не только измерить расстояния между объектами, но и, зафиксировав начало шкалы (скажем, в среднем значении трудности всех заданий, как это принято в Раш-моделировании), определить трудность каждого задания. А значит, мы сможем сформулировать надежные выводы о результатах измерения и проанализировать устойчивость полученных результатов ранжирования.

Дж. Линакр [Linacre, 1989; 2006] разработал два подхода к моделированию ранжирования, которые базировались на методе попарных сравнений, предложенном Л. Терстоуном [Thurstone, 1927]:

1) из ранжирования множества объектов вывести попарные сравнения объектов, например на базе 10 проранжированных объектов будет произведено 45 операций попарных сравнений: 1-й против 2-го, 1-й против 3-го и т. д.;
2) рассмотреть каждую отдельную операцию ранжирования как задание в формате частичного оценивания (*partial credit item*).

Набор данных, который будет участвовать в последующем анализе, включает 1008 наблюдений: 12 экспертов оценили одну папку из 44 заданий плюс 12 экспертов оценили одну папку из 40 заданий. Ранжирование заданий внутри каждой языковой версии инструмента для каждого эксперта представляет собой

один набор рангов. Для построения шкалы оцениваемого экспертами развивающегося навыка чтения использованы оба подхода, предложенных Линакром.

Рассмотрим первый подход. В простейшем случае ранжирования пары заданий осуществляется попарное сравнение заданий и их упорядочивание: одно задание получает ранг выше или ниже другого задания. Здесь важно количество раз, когда одно задание будет получать от экспертов более высокий ранг, чем другое [Linacre, 1989]. Иначе говоря, задание n с уровнем трудности B_n может быть ранжировано «выше», чем задание m с уровнем трудности B_m общее количество H раз в процессе операции ранжирования, осуществленной определенным числом экспертов. С другой стороны, это же задание n может быть ранжировано «ниже», чем задание m, общее количество L раз. Отношение H/L — это данные, необходимые для оценки «расстояния» между мерами двух заданий n и m, которое можно выразить как $B_n - B_m$.

Таким образом, модель измерения для такого случая ранжирования можно представить следующим образом:

$$Ln\ (P_{nm} / P_{mn}) = B_n - B_m,$$

где P_{nm} — это вероятность того, что задание n будет ранжировано выше, чем задание m; P_{mn} — это вероятность того, что задание m будет ранжировано выше, чем задание n, и $P_{mn} + P_{nm} = 1$.

Отношение P_{nm} / P_{mn} позволяет использовать эмпирические данные для оценки параметров. Однако для ранжирования количества заданий большего, чем два, необходимо добавить в эту модель еще один параметр, так как задания оцениваются не независимо, а в определенном порядке внутри одного набора рангов.

В предложенной Линакром модели для сформулированного Терстоуном метода попарных сравнений (далее ТМ, *Thurstone Method*) ранги декомпозируются в попарные сравнения. Тогда модель измерения будет выглядеть следующим образом:

$$Ln\ (P_{nrk} / P_{nrk+1}) = B_n - B_r - F_{rk},$$

где P_{nrk} — это вероятность того, что в упорядоченном рейтинге эксперта r задание n получит ранг k; P_{nrk+1} — это вероятность того, что в упорядоченном рейтинге эксперта r задание n получит ранг $k+1$; B_n — это трудность задания n; B_r — это средняя трудность всех заданий, включенных в рейтинг эксперта r; F_{rk} — это трудность перехода от ранга $k+1$ к рангу k в рейтинге эксперта r.

Преимущество этой модели состоит в том, что ни число экспертов, ни количество заданий, которые нужно проранжировать,

не имеют значения. Оценки всех параметров выводятся из данных о положении каждого задания в каждом рейтинге [Linacre, 1989].

Другой подход предполагает моделирование рейтинга каждого эксперта как отдельного политомического задания, где количеству ответных опций будет соответствовать количество рангов, назначенных экспертом в процессе ранжирования. Аналитически это реализуется так: 12 экспертных рейтингов (по количеству ранжирований, осуществленных 12 экспертами) будут представлять собой «задания», а реальные задания в англоязычной и русскоязычной версиях нашего инструмента будут рассматриваться как «испытуемые». Для этого может быть применена модель частичного оценивания (*Partial Credit Model*, PCM [Masters, 1982]).

Этот подход реализован в работе Т. Брамли [Bramley, 2005], где модель PCM сформулирована как раз для случая ранжирования заданий экспертами:

$$Ln\ (P_{nrk} / P_{nr(k+1)}) = B_n - D_{rk},$$

где P_{nrk} — вероятность того, что задание n будет расположено на позиции k в процессе ранжирования экспертом r; $P_{nr(k+1)}$ — вероятность того, что задание n будет расположено на позиции $k+1$ в процессе ранжирования экспертом r; B_n — уровень трудности задания n; D_{rk} — трудность достижения позиции k относительно позиции $k+1$ в ранжировании экспертом r.

Анализ наших данных осуществлен в логике обоих подходов с применением программного обеспечения *Facets* [Linacre, Wright, 1994] и *Winsteps* [Linacre, 2011] соответственно.

1.2. Результаты первого исследования

Мы рассмотрели согласованность оценок всех экспертов, чтобы убедиться в том, что собранные данные достаточно надежны. Коэффициент конкордации Кендалла является классическим показателем согласованности оценок экспертов [Field, 2014]. Для русскоязычной версии коэффициент конкордации Кендалла составил 0,84, а для англоязычной — 0,87. Таким образом, оценки трудности заданий, которые дали эксперты для каждой языковой версии инструмента, очень высоко согласуются.

Далее мы представим результаты анализа ранжирования экспертами заданий отдельно для каждой языковой версии инструмента в рамках двух моделей, описанных выше: модели попарных сравнений (ТМ) и модели частичного оценивания (PCM). Анализ данных в обоих подходах дает схожие результаты. В табл. 1 приведены обобщенные результаты анализа для заданий. Ошибка оценивания трудности заданий (*Model S.E*) достаточно мала (особенно для модели PCM). При этом разброс трудности заданий (*Range of Measures*), согласно оценкам экс-

Таблица 1. **Общие показатели шкал**

Модель	Показатели качества измерений			
	Статистики согласия		Ошибка оценивания (Model S.E)	Разброс трудностей заданий (Range of Measures)
	Невзвешенная (INFIT MNSQ)	Взвешенная (OUTFIT MNSQ)		
Русскоязычная версия				
TM	1,0	0,6	0,16	9,86
PCM	1,3	1,3	0,07	2,99
Англоязычная версия				
TM	1,0	0,9	0,16	11,24
PCM	1,04	1,04	0,07	3,41

пертов, для модели TM значительно больше. Индикаторами того, насколько данные соответствуют выбранной модели измерения, являются статистики согласия. Они представляют собой среднеквадратичные отклонения эмпирических значений от ожидаемых моделью значений для каждого задания по каждому эксперту и обозначаются INFIT и OUTFIT MNSQ. Как видно из табл. 1, средние значения статистик согласия лежат в пределах рекомендуемых специалистами значений [0,6; 1,4] [Linacre, 2011].

В табл. 2 приведены данные по отдельным заданиям: уровень трудности заданий для обеих языковых версий согласно оценкам экспертов, ошибка измерения трудности, а также статистики согласия данных с моделью измерения[1].

Из табл. 2 можно заключить, что в целом данные подходят обеим моделям для обеих версий. Трудности заданий внутри каждой языковой версии, проанализированные с применением обоих подходов, очень высоко коррелируют. Для русскоязычной версии корреляция (Пирсона) оценок в рамках двух подходов составляет 0,95 ($p < 0,05$), а для англоязычной версии — 0,96 ($p < 0,05$).

Отразить иерархию трудности заданий графически удобно с помощью карт переменных для двух версий инструментов, представленных на рис. 1 и 2. Карты построены в рамках подхода PCM (карты для подхода TM имеют аналогичный вид).

[1] В программе *Facets* оценки трудности заданий представлены как оценки «легкости» задания. Для представления их в табл. 2 эти показатели были конвертированы в оценки трудностей. Трудности заданий представлены в логитах — специальных единицах измерения, принятых в современной теории тестирования.

Таблица 2. **Экспертная оценка трудности заданий для двух языковых версий**

Код задания	Задание	Русскоязычная версия						Англоязычная версия					
		PCM			TM			PCM			TM		
		Трудность (MEASURE)	Ошибка оценивания (S.E)	Статистика согласия (INFIT)	Трудность (MEASURE)	Ошибка оценивания (S.E)	Статистика согласия (INFIT)	Трудность (MEASURE)	Ошибка оценивания (S.E)	Статистика согласия (INFIT)	Трудность (MEASURE)	Ошибка оценивания (S.E)	Статистика согласия (INFIT)
Task01	Структура текста 1	–1,5	0,2	2,0	–2,6	0,1	1,4	–1,8	0,2	1,7	–3,4	0,2	1,5
Task02	Структура текста 2	–1,0	0,1	0,7	–1,9	0,1	1,3	–1,5	0,1	0,6	–2,5	0,2	1,4
Task03	Структура текста 3	–1,2	0,1	1,5	–2,6	0,1	1,5	–1,1	0,1	3,1	–2,6	0,1	1,5
Task04	Структура текста 4	–1,0	0,1	0,3	–2,2	0,1	1,3	–1,4	0,1	0,4	–3,1	0,1	1,4
Task05	Структура текста 5	–1,0	0,1	0,7	–1,9	0,1	1,3	–1,4	0,1	0,8	–2,6	0,1	1,5
Task06	Буквы 1	–1,2	0,1	0,5	–4,4	0,2	0,9	–1,5	0,1	0,6	–6,2	0,2	1,2
Task07	Буквы 12	–1,0	0,1	0,3	–4,2	0,2	1,1	–1,3	0,1	0,2	–4,7	0,2	0,9
Task08	Буквы 13	–0,9	0,1	0,3	–5,2	0,2	1,1	–1,3	0,1	0,6	–5,3	0,2	1,0
Task09	Буквы 14	–0,9	0,1	0,7	–3,4	0,2	0,9	–1,2	0,1	0,6	–5,1	0,2	1,0
Task10	Буквы 15	–0,8	0,1	0,5	–3,5	0,2	0,8	–1,2	0,1	0,4	–4,3	0,2	0,8
Task11	Буквы 16	–0,6	0,1	1,1	–3,5	0,2	0,9	–1,1	0,1	0,2	–4,7	0,2	0,9
Task12	Буквы 17	–0,8	0,1	0,4	–3,2	0,2	0,7	–1,1	0,1	0,2	–4,3	0,2	0,8
Task13	Буквы 18	–0,7	0,1	0,5	–3,0	0,1	0,7	–1,0	0,1	0,4	–3,9	0,2	0,7
Task14	Буквы 19	–0,7	0,1	0,3	–3,0	0,1	0,8	–1,1	0,1	2,3	–4,0	0,2	0,7
Task15	Слова 1	–0,6	0,1	0,8	–1,8	0,1	0,8	–1,0	0,1	2,8	–3,0	0,1	0,6
Task16	Слова 2	–0,6	0,1	0,8	–1,2	0,2	0,9	–1,0	0,1	1,9	–2,0	0,2	0,8
Task17	Слова 3	–0,6	0,1	0,2	–1,0	0,2	0,9	–1,0	0,1	2,7	–1,8	0,2	0,8
Task18	Слова 4	–0,5	0,1	1,5	–2,0	0,1	0,7	–0,8	0,1	0,5	–3,0	0,1	0,6
Task19	Слова 5	–0,8	0,1	1,4	–1,5	0,1	0,9	–0,9	0,1	0,3	–2,1	0,2	0,8
Task20	Слова 6	–0,9	0,1	3,1	–1,0	0,2	0,9	–0,8	0,1	0,6	–2,5	0,2	0,7

Task21	Слова 7	–0,9	0,1	2,4	–0,5	0,2	0,9	–0,8	0,1	0,5	–1,1	0,2	0,9
Task22	Слова 8	–0,8	0,1	2,3	–2,0	0,1	0,8	–0,9	0,1	0,5	–1,5	0,2	0,9
Task23	Слова 9	–0,8	0,1	2,5	–0,9	0,2	0,9	–0,7	0,1	1,1	–1,7	0,2	0,9
Task24	История, часть 1	0,4	0,1	6,3	1,0	0,2	1,1	0,1	0,2	0,8	0,7	0,2	0,6
Task25	История, часть 2	0,7	0,1	5,1	1,5	0,2	1,2	0,2	0,2	0,5	0,7	0,2	0,5
Task26	История, часть 3	0,8	0,1	3,4	1,9	0,2	1,1	0,2	0,2	1,4	0,5	0,2	0,6
Task27	Текст-«ловушка» 1	1,1	0,1	1,1	2,7	0,2	1,0	1,3	0,1	1,5	4,2	0,2	1,0
Task28	Текст-«ловушка» 2	1,2	0,1	0,7	4,0	0,2	1,0	1,0	0,1	1,4	3,4	0,2	1,4
Task29	Текст-«ловушка» 3	1,3	0,1	1,1	3,9	0,2	0,9	1,3	0,1	3,0	4,6	0,2	0,9
Task30	Текст-«ловушка» 4	1,4	0,1	0,9	3,6	0,2	0,9	1,4	0,1	0,7	4,2	0,2	1,0
Task31	Текст-«ловушка» 5	1,2	0,1	1,1	3,3	0,2	1,0	1,4	0,1	1,3	4,3	0,2	1,0
Task32	Текст-«ловушка» 6	1,4	0,1	0,8	4,1	0,2	1,0	1,2	0,1	0,6	4,4	0,2	1,0
Task33	Текст-«ловушка» 7	1,3	0,1	1,0	3,4	0,2	0,9	1,5	0,1	0,5	4,1	0,2	0,9
Task34	Текст-«ловушка» 8	1,3	0,1	1,1	3,1	0,2	0,9	1,5	0,1	0,8	5,0	0,2	1,0
Task35	Текст-«ловушка» 9	1,5	0,1	0,6	3,8	0,2	1,0	1,3	0,1	1,4	3,0	0,2	0,8
Task36	Текст-«ловушка» 10	1,4	0,1	0,9	3,6	0,2	0,9	1,5	0,1	1,0	4,4	0,2	1,0
Task37	Текст-«ловушка» 11	1,3	0,1	0,8	4,2	0,2	1,0	1,2	0,1	2,2	4,5	0,2	1,0
Task38	Текст-«ловушка» 12	1,1	0,1	1,5	3,3	0,2	1,0	1,5	0,1	0,7	2,4	0,2	0,8
Task39	Текст-«ловушка» 13	1,2	0,1	0,8	4,6	0,2	1,0	1,4	0,1	0,9	4,7	0,2	1,0
Task40	Текст-«ловушка» 14	1,3	0,1	0,7	4,1	0,2	1,0	1,5	0,1	0,8	4,8	0,2	1,0
Task41	Текст-«ловушка» 15							1,4	0,1	0,9	3,6	0,2	1,0
Task42	Текст-«ловушка» 16							1,6	0,1	0,4	3,1	0,2	1,3
Task43	Текст-«ловушка» 17							1,6	0,1	0,9	5,0	0,2	1,0
Task44	Текст-«ловушка» 18							1,5	0,1	1,1	3,5	0,2	1,0

Таблица 3. **Разница в трудности заданий для группировки**

Разница в трудности пограничных заданий в логитах	Русскоязычная версия	Англоязычная версия
Task24 — Task23	1,9	2,5
Task26 — Task25	0,8	3,7

На картах самые легкие задания расположены внизу (это задания на распознавание букв), чуть выше представлены задания на оценивание структуры текста и задания на распознавание графической оболочки слов. В средней части шкалы (около 0 логитов) расположены задания, позволяющие проверить навык механического чтения. Наконец, в верхней части находятся самые трудные задания — на чтение с пониманием.

Обе карты демонстрируют кластеризацию заданий в верхней, средней и нижней части шкалы. Более того, расстояния на измерительном континууме между пограничными заданиями верхнего и среднего, среднего и нижнего кластеров достаточно большие (табл. 3) и значительно превышают две ошибки измерения — это три крупные группы заданий, соответствующие разным уровням развития навыка чтения.

Таким образом, в ходе проведенного исследования мы получили экспертные оценки иерархии трудности заданий для двух языковых версий инструмента iPIPS. Мы провели психометрический анализ результатов экспертной оценки в русле двух измерительных подходов, благодаря чему смогли показать, что для обеих языковых версий инструмента оценки чтения iPIPS можно выделить три кластера различающихся в пределах двух ошибок измерения уровней трудности заданий. Эти кластеры представлены одними и теми же группами заданий и на русском, и на английском языке.

К первому кластеру (в частности, на рис. 1 и 2 он расположен в нижней части карты) относятся самые легкие задания инструмента, направленные на оценку понимания ребенком общей структуры печатного текста на своем языке, на распознавание букв и на схватывание графической оболочки слова. Этот кластер соответствует начальному периоду овладения ребенком навыком чтения. Ко второму кластеру (средняя часть карты) принадлежат задания средней трудности. В обеих языковых версиях они направлены на оценку механического чтения (чтения-декодирования). Наконец, третий кластер (верхняя часть карты) включает задания, позволяющие оценить чтение на понимание.

Рис. 1. **Карта переменных русскоязычной версии.** *Partial Credit Model. Winsteps*

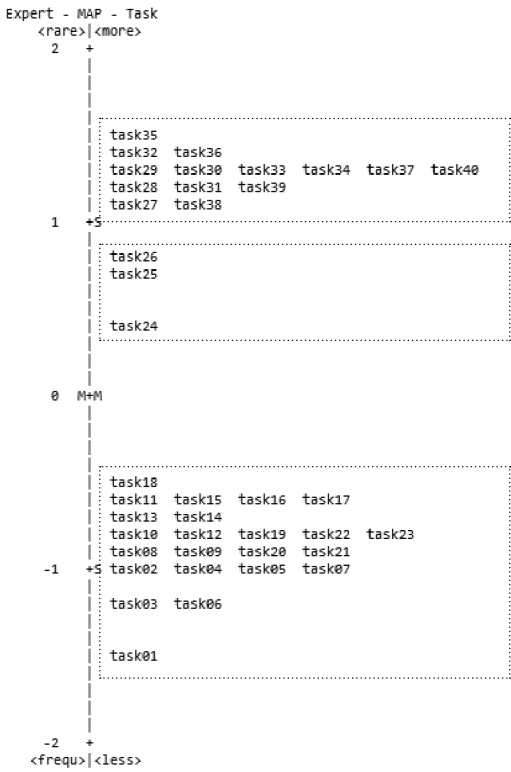
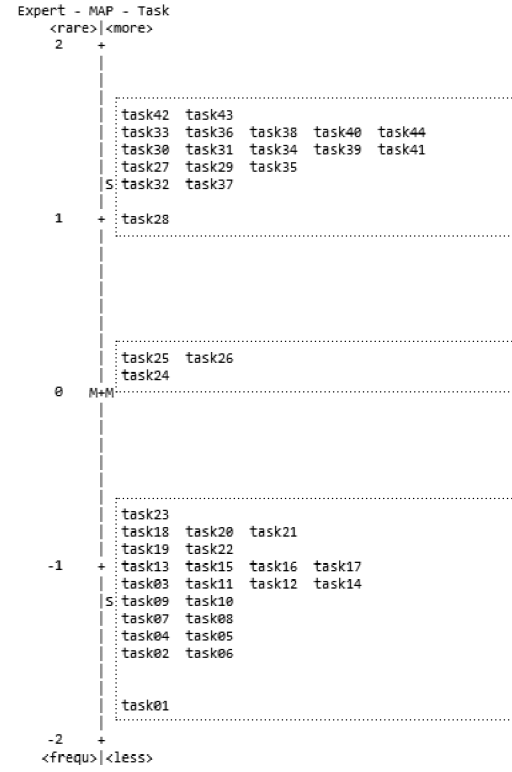


Рис. 2. **Карта переменных англоязычной версии.** *Partial Credit Model. Winsteps*



Выделенные кластеры заданий являются эмпирическим подтверждением теоретической модели чтения, представленной в оригинальной версии инструмента iPIPS, для русскоязычной версии. Эти кластеры заданий могут быть использованы в качестве единого основания для построения сопоставимых пороговых баллов (бенчмарок) на шкале чтения в двух языковых версиях инструмента.

2. Второе исследование. Установление пороговых баллов для сравнения групп учащихся по уровню развития навыка чтения на русском и английском языках

В ходе процедур адаптации/локализации iPIPS на русский язык достичь полной эквивалентности измерений навыка чтения было невозможно [Иванова, Карданова-Бирюкова, 2019], а значит, невозможно провести прямое сопоставление индивидуальных оценок детей в двух странах. Тем не менее, используя реальные результаты тестирования учащихся в двух странах, проведенного на основе эквивалентных процедур, и подтвердив возможность использования единой уровневой модели

развивающегося навыка чтения (первое исследование), можно попытаться провести непрямое сопоставление результатов учащихся двух стран. В частности, используя принципы Раш-измерений, мы можем сопоставить группы учащихся в разных странах по уровню развития чтения.

Мы используем реальные эмпирические данные, полученные на выборках учащихся в России и Великобритании. В качестве референтных данных для русскоязычной версии были взяты данные выборки учащихся 1-х классов на входе в школу в Республике Татарстан. Для англоязычной версии — данные выборки учащихся 1-х классов на входе в школу в Шотландии. Задания, использованные ниже, были описаны в исследовании 1.

2.1. Методология второго исследования

2.1.1 Выборка

В России все необходимые данные для формирования выборки собраны в сотрудничестве с Республиканским центром мониторинга качества образования Республики Татарстан в 2017 г. Выборка составлена как репрезентативная, общим объемом более 5000 детей (44% генеральной совокупности). Генеральной совокупностью данной выборки являлись все учащиеся 1-х классов выбранных районов Татарстана. Стратификация выборки осуществлялась по типу школы и ее местоположению. Единицей выборки являлся класс, случайным образом выбираемый из всей параллели 1-х классов конкретной школы-участницы. В исследовании принимали участие только те дети, которые получили согласие родителей на участие. Общий объем выборки первоклассников, результаты которых в дальнейшем будут использованы для анализа, составил 4940 человек. Средний возраст детей в начале школы составлял 7,4 года.

Выборку детей, представляющих в данном исследовании Великобританию, составили 6627 учащихся из Шотландии — репрезентативная выборка для данного государственного образования (сведения за 2014/2015 учебный год) [Tymms, Merrell, Buckley, 2015]. Средний возраст детей в начале школы составлял 5 лет, хотя отмечена некоторая доля детей старшего возраста.

2.1.2. Аналитический подход для проведения сопоставительного исследования

Основываясь на принципах измерений Г. Раша, мы предлагаем рассматривать результаты, полученные при тестировании детей, как континуум развития навыка чтения, условный «путь», который учащиеся проходят для овладения чтением с пониманием. Здесь мы ориентировались на теоретическую модель чтения, которую предлагает iPIPS, а также на уровневую модель, полученную в ходе анализа результатов ранжирования заданий по трудности с помощью экспертов в первом исследовании.

Используя шкалу экспертных оценок трудности заданий для измерения развивающегося навыка чтения, трансформированную в ходе моделирования в интервальную шкалу логитов, мы смогли выделить три кластера заданий, или три теоретически

интерпретируемых и с помощью экспертных оценок подтвержденных этапа развития навыков чтения у детей. Предлагаемая нами модель позволяет идентифицировать уровни развития навыка чтения: от тех учащихся, кто еще не достиг первого уровня и только начинает знакомиться с концепцией чтения, до продвинутого уровня — детей, способных читать и понимать прочитанное.

Далее мы можем проверить, насколько эмпирическая иерархия заданий соответствует экспертной иерархии и выделенной нами уровневой модели. Если при психометрическом анализе эмпирических данных мы будем наблюдать кластеризацию заданий, соответствующую нашей уровневой модели, мы сможем говорить о возможности построить пороговые оценки перехода детей с одного уровня читательского мастерства на другой.

Все задания, попадающие в каждый выделенный в первом исследовании кластер, можно рассматривать как отдельный субтест, репрезентирующий определенный уровень. Для установления пороговых баллов (бенчмарок) важно определить критерий перехода с одного уровня развития навыка на другой. Опираясь на теоретические работы отечественных специалистов [Беспалько, 1989], мы предположили, что некоторый гипотетический уровень развития навыка можно рассматривать как достигнутый, если по крайней мере 70% заданий данного уровня выполнены верно (речь идет о вероятностных оценках). Согласно данной концепции, усвоение материала на 70% свидетельствует о готовности к усвоению нового материала, а также о сформированности навыка.

Пороговые баллы перехода с одного уровня развития навыка чтения на другой были установлены с помощью аппарата современной теории тестирования. Для этого использовалась следующая процедура. Сформированы три гипотетических задания, которые репрезентируют каждый уровень. Значения трудности этих гипотетических заданий определялись как среднее значение всех заданий соответствующего уровня. Далее для каждого уровня шкалы по чтению определялось значение порога его достижения как уровень навыка, при котором вероятность верно выполнить гипотетическое «усредненное» задание на данном уровне составляет 0,7 (принятый нами 70%-ный порог достижения уровня). Все участники, результат тестирования которых находится ниже данной границы, считаются не достигшими этого уровня, равно как и всех последующих.

2.2. Результаты второго исследования

Следуя описанному выше аналитическому подходу, рассмотрим результаты предварительного психометрического анализа имеющихся эмпирических данных и графически сопоставим иерархии заданий, полученных для русскоязычной и англоязычной версий теста с помощью экспертов и в ходе тестирования детей. Для перевода первичных баллов детей в оценки их способности к чте-

Таблица 4. **Результаты анализа надежности эмпирических данных**

Показатели	Надежность Альфа	Person Reliability	Person Separation
Русскоязычная версия	0,97	0,87	2,56
Англоязычная версия	0,75	0,71	1,58

нию применялась однопараметрическая дихотомическая модель Раша [Wright, Stone, 1979]. Проведен психометрический анализ заданий, анализ размерности и надежности шкалы, анализ соответствия данных модели измерения. Для психометрического анализа теста и оценки параметров заданий и испытуемых использовалась программа *Winsteps* [Linacre, 2011].

В табл. 4 представлены общие психометрические показатели качества шкал чтения для эмпирических данных, полученных на выборках учеников в России и Великобритании.

Для обеих версий инструмента характерны высокие показатели надежности — классической и Раш-надежности (*Person Reliability*), а также достаточно высокий общий уровень чувствительности шкалы (*Person Separation*), позволяющий различать по крайней мере три группы[2] испытуемых, выделенных на основании уровня навыка чтения.

Убедившись в психометрическом качестве шкалы чтения на выборке учащихся из России и Великобритании, мы можем рассмотреть иерархию заданий на картах переменных и сравнить ее с иерархией в первом исследовании (рис. 3, 4). На рисунках задания представлены справа, а распределение испытуемых слева.

Хотя разграничение кластеров заданий вдоль оси на картах переменных менее явное, чем в случае экспертных оценок, их общие структуры полностью аналогичны. В нижней части карты расположены самые легкие задания инструмента, направленные на оценку понимания ребенком общей структуры печатного текста (I1–I5), на распознавание букв (L1–L8), на схватывание графической оболочки слова (W1–W9), что соответствует заданиям первого кластера в исследовании 1. В средней части карты находятся задания на оценку механического чтения (S1–S3). В верхней части карты лежат задания, позволяющие оценить

[2] Для определения количества групп (страт), на которые могут быть поделены испытуемые, используется формула перевода индекса *Separation* по процедуре, изложенной в руководстве *Winsteps* [Linacre, 2011].

Рис. 3. **Карта переменных. Результаты тестирования учащихся на английском языке**

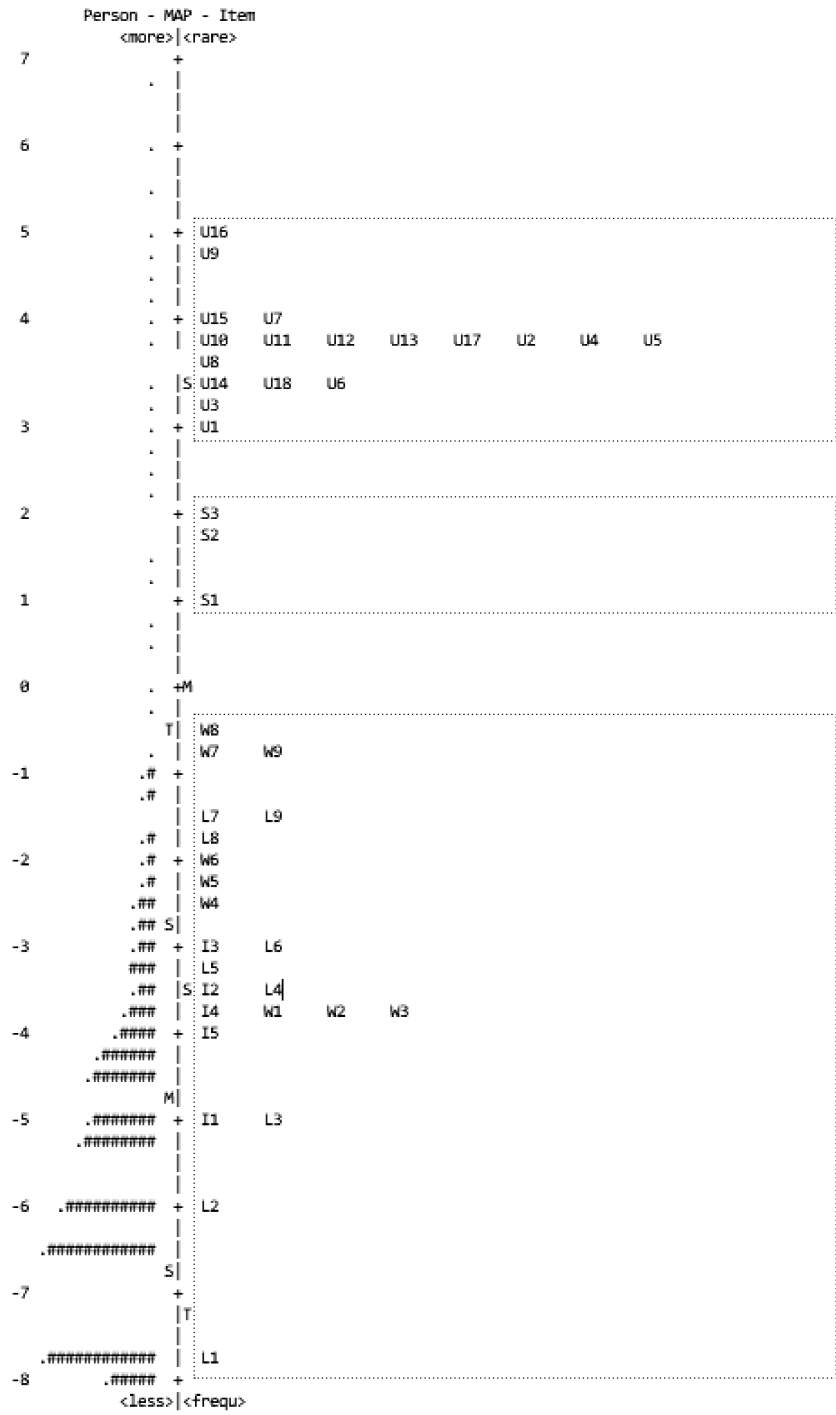


Рис. 4. **Карта переменных. Результаты тестирования учащихся на русском языке**

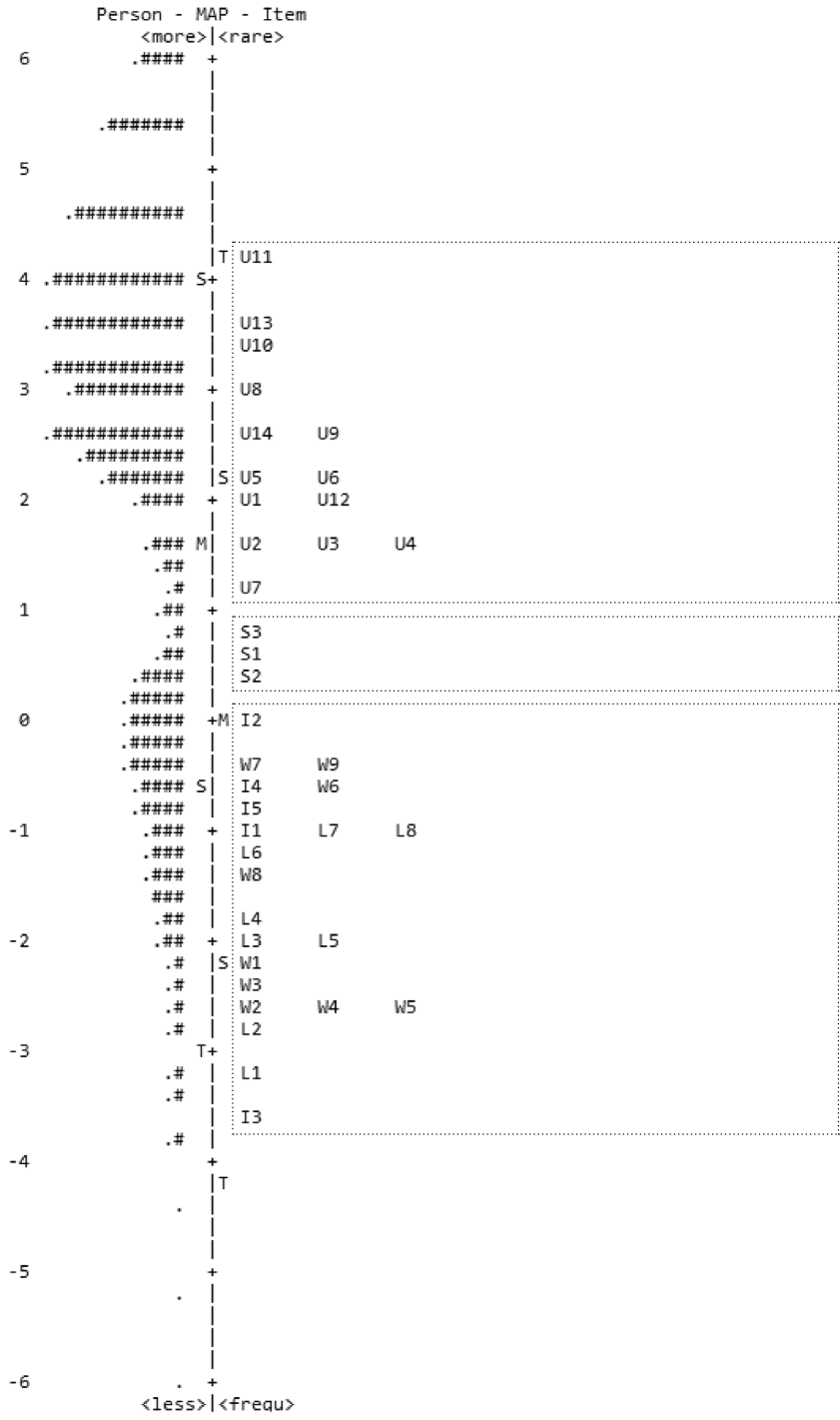


Таблица 5. **Установление пороговых значений**

Пороговые баллы для уровней	Выборка			
	Россия		Великобритания	
	Средняя трудность заданий уровня	Пороговый балл достижения уровня	Средняя трудность заданий уровня	Пороговый балл достижения уровня
Порог уровня 3. Чтение с пониманием	2,11	2,96	3,78	4,63
Порог уровня 2. Механическое чтение (декодирование)	0,61	1,46	1,53	2,38
Порог уровня 1. Понимание структуры печатного теста, распознавание букв и частотных слов	–1,63	–0,78	–3,16	–2,31

чтение на понимание (U1–U14), соответствующие третьему кластеру.

Чтобы оценить, насколько реалистично полученные с помощью экспертов оценки трудности заданий в каждой из языковых версий отражают иерархию заданий внутри шкалы чтения, мы провели корреляционный анализ результатов оценок трудности заданий от экспертов и полученных в ходе тестирования. Коэффициенты корреляции (Пирсона) достаточно высоки: 0,81 для русскоязычной версии и 0,88 для англоязычной версии.

Таким образом, анализ имеющихся эмпирических и экспертных данных подтверждает возможность использования выделенной уровневой модели чтения как основы для построения пороговых баллов (бенчмарок) и для проведения непрямого международного сопоставления.

В ходе выполнения описанной в разделе об аналитическом подходе процедуры нами определены пороговые баллы на шкале развития навыка чтения, которые задают границы освоения каждого из уровней овладения навыком чтения (табл. 5[3]). Эти уровни могут быть единым образом применены для результатов тестирования детей в двух языковых версиях инструмента.

Определение пороговых баллов для русскоязычной версии инструмента проиллюстрировано на рис. 5. По оси абсцисс на графике представлен уровень подготовленности

[3] Пороговые баллы в таблице представлены в логитах, однако они легко могут быть преобразованы в любую шкалу, на которой представляются результаты тестирования, например в 100-балльную.

Рис. 5. **Определение пороговых оценок при делении учеников по уровням овладения чтением** (русскоязычная версия инструмента оценивания)

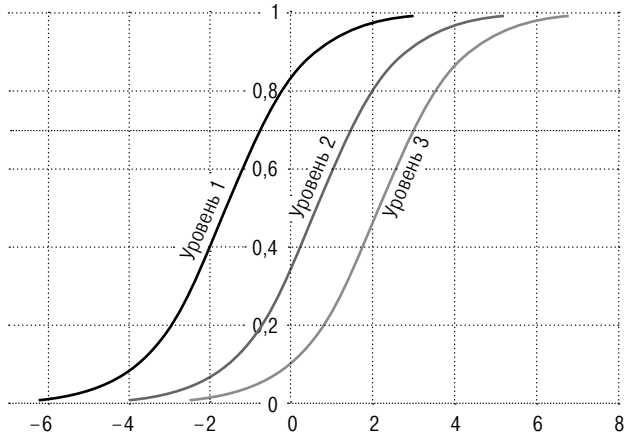
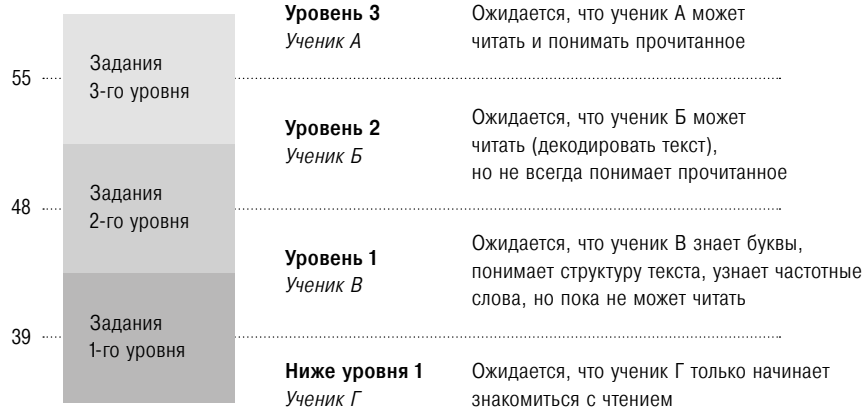


Рис. 6. **Категоризация пороговых значений** (пример для русскоязычной версии)



испытуемых в логитах, по оси ординат — вероятность выполнения задания. Три кривые на графике соответствуют характеристическим кривым[4] трех гипотетических заданий, представляющих собой усредненные задания трех уровней. Горизонтальная

[4] Характеристическая кривая задания представляет собой график вероятности правильного выполнения задания в зависимости от уровня подготовленности испытуемого.

Таблица 6. **Сравнение распределения учеников в двух странах по уровню овладения чтением на входе в школу**

Уровень навыка чтения	Описание уровня	Доля выборки, %	
		Россия	Великобритания
Уровень 3	Чтение с пониманием	32,7	0,3
Уровень 2	Механическое чтение	27,7	0,6
Уровень 1	Распознавание букв и частотных слов, понимание структуры текста	23,8	9,9
Ниже уровня 1	Первое знакомство с чтением	15,9	89,2

прямая задает принятый нами 70%-ный порог достижения уровня. Абсцисса точки пересечения горизонтальной линии и характеристической кривой для каждого уровня развития навыка чтения задает пороговый балл.

На рис. 6 показана шкала чтения с установленными пороговыми баллами. Пороговые баллы переведены на 100-балльную шкалу, которая используется для представления результатов участников тестирования (в стандартизированных баллах со средним значением 50 и стандартным отклонением 10). В итоге получаются следующие пороговые оценки: переход на 1-й уровень — 39 баллов; на 2-й уровень — 48 баллов; на 3-й уровень — 55 баллов.

2.2.1. Использование пороговых баллов в целях сопоставительного исследования

Таким образом, в ходе второго исследования были построены пороговые баллы (бенчмарки), которые задают границы каждого из уровней овладения навыком чтения для каждой из страновых версий теста. В совокупности результаты первого и второго исследований свидетельствуют о возможности провести международное сопоставление навыков чтения в группах детей на входе в школу. В табл. 6 представлено распределение по уровням освоения навыка чтения учащихся в России и Великобритании.

Распределения учащихся из двух стран по уровням освоения навыка чтения сильно различаются. Интерпретация полученных результатов не является фокусом данной статьи, тем не менее отметим, что используемые в данном исследовании выборки значительно различаются по возрасту — возможно, именно этим объясняются такие существенные различия в достижениях детей. Основным же результатом второго исследования можно считать то, что описанная методика пригодна для осуществления непрямого сопоставления результатов международного исследования в условиях отсутствия единой метрической шкалы.

3. Обсуждение результатов

Чтобы исследователи, политики и практики в сфере образования могли полагаться на данные межстрановых исследований, необходимо проверить и доказать, что эти данные являются надежными, справедливыми и осмысленными, что они адекватно репрезентируют измеряемый конструкт, что на их основе можно принимать обоснованные решения. При создании русскоязычной версии инструмента iPIPS, первоначально разработанного на английском языке, были использованы международные стандарты разработки и адаптации тестов [Leong, 2016].

Инструмент iPIPS широко применяется школами в Великобритании, а также в ряде других стран, включая Австралию, Бразилию, Германию, Южную Африку [Archer et al., 2010; Bartholo et al., 2019; Tymms et al., 2014; Vidmar et al., 2017]. Он использовался в некоторых ранее опубликованных исследованиях для сравнения результатов детей на международном уровне, например в англоязычных странах — Великобритании, Австралии и Новой Зеландии [Tymms et al., 2014], где авторы пытались оценить возможные различия в академических достижениях и прогрессе детей и эффективность образовательных систем.

Еще в одном исследовании проверялся потенциал iPIPS для сравнения результатов тестирования по математике детей из Великобритании (Англии, Шотландии) и России — стран с разным возрастом старта начальной школы, разными учебными программами и разными языками и культурами [Ivanova et al., 2018]. Показано, что, несмотря на явные трудности, прямое международное сопоставление результатов тестирования iPIPS по математике возможно.

В исследовании [Vidmar et al., 2017] сравниваются результаты первоклассников в чтении на базе инструмента iPIPS для Сербии и Германии. Однако в данной статье даются только средние оценки по выборке, при этом никаких доказательств сопоставимости языковых версий не приводится.

Таким образом, исследований, в которых с помощью инструмента iPIPS сопоставлялись бы результаты оценивания чтения на входе в школу в группах учащихся из разных стран, до настоящего времени не существовало.

В данной работе сделана попытка решить интересную научную проблему проведения международного сопоставления уровней развития навыка чтения у учащихся на входе в школу на примере результатов тестирования первоклассников в России и в Великобритании. В первом исследовании проведена экспертиза конструкта — уровневой модели навыков чтения у детей на входе в школу на русском и английском языках. Методологической основой экспертизы стал метод экспертного ранжирования.

Сравнение иерархии трудности заданий, полученной в результате калибровки данных с использованием двух подходов

Раш-моделирования, показало, что в двух языковых версиях можно выделить три кластера заданий. Эти кластеры представлены одними и теми же заданиями и на русском, и на английском языках.

В первом исследовании показано, что экспертные оценки трудности заданий, репрезентирующих развивающийся навык чтения у учащегося на входе в школу, могут быть использованы для выявления иерархии заданий вдоль континуума данного конструкта, для сравнения иерархии заданий в двух языковых версиях и, наконец, для формирования основания для построения бенчмарок — пороговых баллов между уровнями развития навыка чтения на двух языках, русском и английском.

Во втором исследовании с использованием данных тестирования выборок русско- и англоговорящих учащихся в двух странах установлены пороговые значения и определены уровни развития навыка чтения. Эти уровни единым образом применены для результатов тестирования детей в двух языковых версиях инструмента и для выделения групп детей, находящихся на том или ином уровне развития чтения в двух странах.

Мы предполагаем, что если структура предложенной теоретической модели чтения iPIPS подтверждается в двух любых рассматриваемых для целей сопоставления странах (т. е. выделенные экспертами, а также подтвержденные в ходе психометрического анализа кластеры заданий инструмента отражают один и тот же конструкт), то этот факт может стать основой для построения международных эталонных показателей — пороговых баллов, на основании которых будет возможно сравнивать совокупный процент детей, достигших в той или иной стране того или иного уровня чтения. Данное предположение в будущем следует проверить для других языковых версий инструмента iPIPS.

Практическая значимость выполненной работы состоит в том, что она знакомит заинтересованное сообщество с проблемами, с которыми можно столкнуться при оценивании конкретного конструкта (навыка чтения) на специфической выборке участников (дети на входе в школу) из разных стран. Методология данной работы применима для решения проблем исследований, рассматривающих схожие конструкты и ориентированных на младших школьников.

В перспективе данную методику можно использовать как для международного сопоставления данных, так и для других целей. Например, для сопоставления по годам результатов тестирования, полученных на разных выборках учащихся и с помощью различающихся версий одного инструмента оценивания. В Великобритании [Bramley, 2005] такая практика реализуется уже несколько лет для сопоставления результатов некоторых письменных экзаменов.

Литература

1. Беспалько В. П. (1989) Слагаемые педагогической технологии. М.: Педагогика.
2. Иванова А. Е., Карданова-Бирюкова К.С. (2019) Создание русскоязычной версии международного инструмента оценивания ранних навыков чтения // Вопросы образования/Educational Studies Moscow. № 4. С. 93–115. DOI:10.17323/1814-9545-2019-4-93-115.
3. Карданова Е. Ю., Иванова А. Е., Сергоманов П. А., Канонир Т. Н., Антипкина И. В., Кайкы Д. Н. (2018) Обобщенные типы развития первоклассников на входе в школу. По материалам исследования iPIPS // Вопросы образования/Educational Studies Moscow. № 1. С. 8–37. DOI:10.17323/1814-9545-2018-1-8-37.
4. Ainley M., Ainley J. (2019) Non-Cognitive Attributes: Measurement and Meaning // L. E. Suter et al. (eds) The SAGE Handbook of Comparative Studies in Education. London: SAGE. P. 103–125.
5. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME) (2014) Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.
6. Archer E., Howie S. J., Scherman V., Coe R. (2010) Finding the Best Fit: The Adaptation and Translation of the Performance Indicators for Primary Schools for the South African context // Perspectives in Education. Vol. 28. No 1. P. 77–88.
7. Bartholo T. L., Koslinski M. C., Costa M. D., Barcellos T. (2019) What Do Children Know upon Entry to Pre-School in Rio de Janeiro? Ensaio: Avaliação e Políticas Públicas em Educação.
8. Bramley T. (2005) A Rank-Ordering Method for Equating Tests by Expert Judgment // Journal of Applied Measurement. Vol. 6. No 2. P. 202–223.
9. Buzhardt J., Greenwood C. R., Hackworth N. J., Jia F., Bennetts S. K., Walker D., Matthews J. M. (2019) Cross-Cultural Exploration of Growth in Expressive Communication of English-Speaking Infants and Toddlers // Early Childhood Research Quarterly. Vol. 48. 3rd Quarter 2019. P. 284–294.
10. Carnoy M., Khavenson T., Loyalka P., Schmidt W. H., Zakharov A. (2016) Revisiting the Relationship between International Assessment Outcomes and Educational Production: Evidence from a Longitudinal PISA-TIMSS Sample // American Educational Research Journal. Vol. 53. No 4. P. 1054–1085.
11. Caro D. H., Cortés D. (2012) Measuring Family Socioeconomic Status: An Illustration Using Data from PIRLS2006 // IERI Monograph Series Issues and Methodologies in Large-Scale Assessments. No 5. P. 9–33.
12. Dubeck M. M., Gove A. (2015) The Early Grade Reading Assessment (EGRA): Its Theoretical Foundation, Purpose, and Limitations // International Journal of Educational Development. Vol. 40. January. P. 315–322.
13. Ercikan K., Roth W. M., Asil M. (2015) Cautions about Inferences from International Assessments: The Case of PISA 2009 // Teachers College Record. Vol. 117. No 1. P. 1–28.
14. Espeland W. (2015) Narrating Numbers // R. Rottenburg, S. Merry, S. Park, J. Mugler (eds). The World of Indicators: The Making of Governmental Knowledge through Quantification, Cambridge Studies in Law and Society. Cambridge: Cambridge University. P. 56–75.
15. Esselink B. (2000) A Practical Guide to Localization. Vol. 4. Amsterdam, Philadelphia: John Benjamins.
16. Field A. P. (2014) Kendall's Coefficient of Concordance // Wiley StatsRef: Statistics Reference Online.
17. Goodrich S., Ercikan K. (2019) Measurement Comparability of Reading in English and French Canadian Population: Special Case of the 2011 Pro-

gress in International Reading Literacy Study // Frontiers in Education. Vol. 4. https://www.frontiersin.org/articles/10.3389/feduc.2019.00120/full
18. Ivanova A., Kardanova E., Merrell C., Tymms P., Hawker D. (2018) Checking the Possibility of Equating a Mathematics Assessment between Russia, Scotland and England for Children Starting School // Assessment in Education: Principles, Policy & Practice. Vol. 25. No 2. P. 141–159.
19. Kreiner S., Christensen K. B. (2014) Analyses of Model Fit and Robustness. A New Look at the PISA Scaling Model Underlying Ranking of Countries According to Reading Literacy // Psychometrika. Vol. 79. No 2. P. 210–231.
20. Leong F. T., Bartram D., Cheung F., Geisinger K. F., Iliescu D. (2016) The ITC International Handbook of Testing and Assessment. New York, NY: Oxford University.
21. Linacre J. M. (2011) Winsteps Rasch Measurement. Version 3.71. Winsteps. com.
22. Linacre J. M. (2006) Rasch Analisis of Rank-Ordered Data // Journal of Applied Measurement. Vol. 7. No 1. P. 129–139.
23. Linacre J. M. (1989) Rank Ordering and Rasch Measurement // Rasch Measurement Transactions. Vol. 2. No 4. P. 41–42.
24. Linacre J. M., Wright B. D. (1994) A User's Guide to FACETS: Rasch Measurement Computer Program. Chicago: MESA.
25. Liu J., Steiner-Khamsi G. (2020) Human Capital Index and the Hidden Penalty for Non-Participation in ILSAs // International Journal of Educational Development. Vol. 73. Iss. C. P. 1–9.
26. Masters G. N. (1982) A Rasch Model for Partial Credit Scoring // Psychometrika. Vol. 47. No 2. P. 149–174.
27. Merrell C., Tymms P. (2007) Identifying Reading Problems with Computer-Adaptive Assessments // Journal of Computer Assisted Learning. Vol. 23. No 1. P. 27–35.
28. OECD (2020) Early Learning and Child Well-Being: A Study of Five-Year-Olds in England, Estonia, and the United States. Paris: OECD.
29. Peña E. D. (2007) Lost in Translation: Methodological Considerations in Cross-Cultural Research // Child Development. Vol. 78. No 4. P. 1255–1264.
30. Shuttleworth-Edwards A.B., Kemp R. D., Rust A. L., Muirhead J. G., Hartman N. P., Radloff S. E. (2004) Cross-Cultural Effects on IQ Test Performance: A Review and Preliminary Normative Indications on WAIS-III Test Performance // Journal of Clinical and Experimental Neuropsychology. Vol. 26. No 7. P. 903–920.
31. Suggate S. P. (2009) School Entry Age and Reading Achievement in the 2006 Programme for International Student Assessment (PISA) // International Journal of Educational Research. Vol. 48. No 3. P. 151–161.
32. Thurstone L. L. (1927) A Law of Comparative Judgment // Psychological Review. Vol. 34. No 4. P. 273–286.
33. Tymms P. (1999) Baseline Assessment, Value-Added and the Prediction of Reading // Journal of Research in Reading. Vol. 22. No 1. P. 27–36.
34. Tymms P., Merrell C., Buckley H. (2015) Children's Development at the Start of School in Scotland and the Progress Made during their First School Year: An Analysis of PIPS Baseline and Follow-Up Assessment Data. Edinburgh, UK: The Scottish Government. http://dro.dur.ac.uk/17417/
35. Tymms P., Merrell C., Hawker D., Nicholson F. (2014) Performance Indicators in Primary Schools: A Comparison of Performance on Entry to School and the Progress Made in the First Year in England and Four Other Jurisdictions. http://dro.dur.ac.uk/23562/1/23562.pdf
36. Vidmar M., Niklas F., Schneider W., Hasselhorn M. (2017) On-Entry Assessment of School Competencies and Academic Achievement: A Comparison between Slovenia and Germany // European Journal of Psychology of Education. Vol. 32. No 2. P. 311–331.
37. Wright B. D., Stone M. H. (1979) Best Test Design. Chicago, IL: MESA.

Checking the Possibility of an International Comparative Study of Reading Literacy Assessment for Children Starting School

Authors **Alina Ivanova**
Research Fellow, Center for Psychometrics and Measurements in Education, Institute of Education, National Research University Higher School of Economics. E-mail: aeivanova@hse.ru

Elena Kardanova
Candidate of Sciences in Mathematical Physics, Associate Professor, Tenured Professor, Director of the Center for Psychometrics and Measurements in Education, Institute of Education, National Research University Higher School of Economics. E-mail: ekardanova@hse.ru

Address: Bld. 10, 16 Potapovsky Lane, 101000 Moscow, Russian Federation.

Abstract The early years of school, when a child is only learning to read, are critically important for later development and learning. Cross-cultural comparative assessments of reading literacy provide a rich source of data for researchers, practitioners and politicians on the opportunities and prospects of early childhood development in different countries, circumstances and contexts. There are few publications of this sort available, and none of them has involved Russian-speaking children on entry to school so far.

Data obtained using two language versions of the International Performance Indicators in Primary Schools (iPIPS) on representative samples of first-graders from the Republic of Tatarstan and Scotland is used to compare the early reading assessment results between children starting school in countries with linguistic, cultural, and school entry age differences.

Two studies are conducted to analyze the possible methods of comparing assessment results of children from different countries in the absence of a common measurement scale. The first study uses the rank-ordering method to establish a correspondence between the levels of reading literacy among Russian- and English-speaking children by expert judgment. In the second study, the obtained model of literacy levels is used to establish the cut-off scores (benchmarks) of student assessment outcomes.

Keywords cross-cultural comparative assessments, elementary school, expert judgment, pairwise comparison, Rasch modelling.

References Ainley M., Ainley J. (2019) Non-Cognitive Attributes: Measurement and Meaning. *The SAGE Handbook of Comparative Studies in Education* (eds L. E. Suter et al.), London: Sage, pp. 103–125.

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

Archer E., Howie S. J., Scherman V., Coe R. (2010) Finding the Best Fit: The Adaptation and Translation of the Performance Indicators for Primary Schools for the South African context. *Perspectives in Education*, vol. 28, no 1, pp. 77–88.

Bartholo T. L., Koslinski M. C., Costa M. D., Barcellos T. (2019) *What Do Children Know upon Entry to Pre-School in Rio de Janeiro?* Ensaio: Avaliação e Políticas Públicas em Educação.

Bespalko V. (1989) *Slagaemye pedagogicheskoy tekhnologii* [Components of Pedagogical Technology]. Moscow: Pedagogika.

Bramley T. (2005) A Rank-Ordering Method for Equating Tests by Expert Judgment. *Journal of Applied Measurement*, vol. 6, no 2, pp. 202–223.

Buzhardt J., Greenwood C. R., Hackworth N. J., Jia F., Bennetts S. K., Walker D., Matthews J. M. (2019) Cross-Cultural Exploration of Growth in Expressive Communication of English-Speaking Infants and Toddlers. *Early Childhood Research Quarterly*, vol. 48, 3rd Quarter 2019, pp. 284–294.

Carnoy M., Khavenson T., Loyalka P., Schmidt W. H., Zakharov A. (2016) Revisiting the Relationship between International Assessment Outcomes and Educational Production: Evidence from a Longitudinal PISA-TIMSS Sample. *American Educational Research Journal*, vol. 53, no 4, pp. 1054–1085.

Caro D. H., Cortés D. (2012) Measuring Family Socioeconomic Status: An Illustration Using Data from PIRLS2006. *IERI Monograph Series Issues and Methodologies in Large-Scale Assessments*, no 5, pp. 9–33.

Dubeck M. M., Gove A. (2015) The Early Grade Reading Assessment (EGRA): Its Theoretical Foundation, Purpose, and Limitations. *International Journal of Educational Development*, vol. 40, January, pp. 315–322.

Ercikan K., Roth W. M., Asil M. (2015) Cautions about Inferences from International Assessments: The Case of PISA 2009. *Teachers College Record*, vol. 117, no 1, pp. 1–28.

Espeland W. (2015) Narrating Numbers. *The World of Indicators: The Making of Governmental Knowledge through Quantification*. *Cambridge Studies in Law and Society* (eds R. Rottenburg, S. Merry, S. Park, J. Mugler), Cambridge: Cambridge University, pp. 56–75.

Esselink B. (2000) *A Practical Guide to Localization. Vol. 4*. Amsterdam, Philadelphia: John Benjamins.

Field A. P. (2014) *Kendall's Coefficient of Concordance*. Wiley StatsRef: Statistics Reference Online.

Goodrich S., Ercikan K. (2019) Measurement Comparability of Reading in English and French Canadian Population: Special Case of the 2011 Progress in International Reading Literacy Study. *Frontiers in Education*, vol. 4. Available at: https://www.frontiersin.org/articles/10.3389/feduc.2019.00120/full (accessed 27 September 2020).

Ivanova A., Kardanova-Biryukova K. (2019) Sozdanie russkoyazychnoy versii mezhdunarodnogo instrumenta otsenivaniya rannikh navykov chteniya [Constructing a Russian-Language Version of the International Early Reading Assessment Tool]. *Voprosy obrazovaniya/Educational Studies Moscow*, no 4, pp. 93–115. DOI:10.17323/1814-9545-2019-4-93-115.

Ivanova A., Kardanova E., Merrell C., Tymms P., Hawker D. (2018) Checking the Possibility of Equating a Mathematics Assessment between Russia, Scotland and England for Children Starting School. *Assessment in Education: Principles, Policy & Practice*, vol. 25, no 2, pp. 141–159.

Kardanova E., Ivanova A., Sergomanov P., Kanonire T., Antipkina I., Kayky D. (2018) Obobshchennye tipy razvitiya pervoklassnikov na vkhode v shkolu. Po materialam issledovaniya iPIPS [Patterns of First-Graders' Development at the Start of Schooling: Cluster Approach Based on the Results of iPIPS Project]. *Voprosy obrazovaniya/Educational Studies Moscow*, no 1, pp. 8–37. DOI:10.17323/1814-9545-2018-1-8-37.

Kreiner S., Christensen K. B. (2014) Analyses of Model Fit and Robustness. A New Look at the PISA Scaling Model Underlying Ranking of Countries According to Reading Literacy. *Psychometrika*, vol. 79, no 2, pp. 210–231.

Leong F. T., Bartram D., Cheung F., Geisinger K. F., Iliescu D. (2016) *The ITC International Handbook of Testing and Assessment*. New York, NY: Oxford University.

Linacre J. M. (2011) *Winsteps Rasch Measurement. Version 3.71. Winsteps. com*.
Linacre J. M. (2006) Rasch Analisis of Rank-Ordered Data. *Journal of Applied Measurement*, vol. 7, no 1, pp. 129–139.
Linacre J. M. (1989) Rank Ordering and Rasch Measurement. *Rasch Measurement Transactions*, vol. 2, no 4, pp. 41–42.
Linacre J. M., Wright B. D. (1994) *A User's Guide to FACETS: Rasch Measurement Computer Program*. Chicago: MESA.
Liu J., Steiner-Khamsi G. (2020) Human Capital Index and the Hidden Penalty for Non-Participation in ILSAs. *International Journal of Educational Development*, vol. 73, iss. C, pp. 1–9.
Masters G. N. (1982) A Rasch Model for Partial Credit Scoring. *Psychometrika*, vol. 47, no 2, pp. 149–174.
Merrell C., Tymms P. (2007) Identifying Reading Problems with Computer-Adaptive Assessments. *Journal of Computer Assisted Learning*, vol. 23, no 1, pp. 27–35.
OECD (2020) *Early Learning and Child Well-Being: A Study of Five-Year-Olds in England, Estonia, and the United States*. Paris: OECD.
Peña E. D. (2007) Lost in Translation: Methodological Considerations in Cross-Cultural Research. *Child Development*, vol. 78, no 4, pp. 1255–1264.
Shuttleworth-Edwards A.B., Kemp R. D., Rust A. L., Muirhead J. G., Hartman N. P., Radloff S. E. (2004) Cross-Cultural Effects on IQ Test Performance: A Review and Preliminary Normative Indications on WAIS-III Test Performance. *Journal of Clinical and Experimental Neuropsychology*, vol. 26, no 7, pp. 903–920.
Suggate S. P. (2009) School Entry Age and Reading Achievement in the 2006 Programme for International Student Assessment (PISA). *International Journal of Educational Research*, vol. 48, no 3, pp. 151–161.
Thurstone L. L. (1927) A Law of Comparative Judgment. *Psychological Review*, vol. 34, no 4, pp. 273–286.
Tymms P. (1999) Baseline Assessment, Value-Added and the Prediction of Reading. *Journal of Research in Reading*, vol. 22, no 1, pp. 27–36.
Tymms P., Merrell C., Buckley H. (2015) *Children's Development at the Start of School in Scotland and the Progress Made during their First School Year: An Analysis of PIPS Baseline and Follow-Up Assessment Data*. Edinburgh, UK: The Scottish Government. Available at: http://dro.dur.ac.uk/17417/ (accessed 27 September 2020).
Tymms P., Merrell C., Hawker D., Nicholson F. (2014) *Performance Indicators in Primary Schools: A Comparison of Performance on Entry to School and the Progress Made in the First Year in England and Four Other Jurisdictions*. Available at: http://dro.dur.ac.uk/23562/1/23562.pdf (accessed 27 September 2020).
Vidmar M., Niklas F., Schneider W., Hasselhorn M. (2017) On-Entry Assessment of School Competencies and Academic Achievement: A Comparison between Slovenia and Germany. *European Journal of Psychology of Education*, vol. 32, no 2, pp. 311–331.
Wright B. D., Stone M. H. (1979) *Best Test Design*. Chicago, IL: MESA.