Измерение базовой математической грамотности в начальной школе

Д. А. Федерякин, Г. С. Ларина, Е. Ю. Карданова

Статья поступила в редакцию в сентябре 2020 г.

Федерякин Денис Александрович — стажер-исследователь Центра психометрики и измерений в образовании, Институт образования, Национальный исследовательский университет «Высшая школа экономики». E-mail: dafederiakin@hse.ru (контактное лицо для переписки)

Ларина Галина Сергеевна — кандидат наук об образовании, научный сотрудник Центра психометрики и измерений в образовании, Институт образования, Национальный исследовательский университет «Высшая школа экономики». E-mail: glarina@hse.ru

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

Адрес: 101000, Россия, Москва, ул. Мясницкая, 20.

Аннотация

Измерение математической грамотности затруднено в силу многокомпонентности данного конструкта, а также частой нагруженности заданий текстовой информацией. Пользователи результатов измерения, как правило, предъявляют спрос на информацию как об общем уровне развития математической грамотности у респондентов, так и об уровнях развития отдельных ее компонентов. Согласно стандартам образовательного и психологического тестирования для одновременного сообщения пользователям как общего балла по тесту, так и баллов по субшкалам теста требуется проведение дополнительных психометрических исследований, доказывающих валидность всех сообщаемых баллов. Проведено исследование, в результате которого показано, что тест базовой математической грамотности PROGRESS-ML, предназначенный для учащихся начальной школы, может быть использован в качестве одномерного инструмента измерения, что дает возможность сообщать общий тестовый балл по тесту. При этом чтение не вносит значимого вклада в вероятность решения заданий, и баллы, полученные по отдельным компонентам теста, могут быть самостоятельно представлены пользователям результатов тестов в дополнение к общему баллу.

Ключевые слова

сложные конструкты, композитные инструменты измерения, базовая математическая грамотность, PROGRESS-ML.

Для цитирования

Федерякин Д. А., Ларина Г. С., Карданова Е. Ю. (2021) Измерение базовой математической грамотности в начальной школе // Вопросы образования / Educational Studies Moscow. № 2. С. 199–226. https://doi.org/10.17323/1814-9545-2021-2-199-226

Measuring Basic Mathematical Literacy in Elementary School

D. A. Federiakin, G. S. Larina, E. Yu. Kardanova

Denis Federiakin, Intern Researcher, Center for Psychometrics and Measurement in Education, Institute of Education, National Research University Higher School of Economics. E-mail: dafederiakin@hse.ru (corresponding author)
Galina Larina, Candidate of Sciences in Education, Research Fellow, Center for Psychometrics and Measurement in Education, Institute of Education, National Research University Higher School of Economics. E-mail: glarina@hse.ru
Elena Kardanova, Candidate of Sciences in Mathematical Physics, Associate Professor, Tenured Professor, Director of the Center for Psychometrics and Measurement in Education, Institute of Education, National Research University Higher School of Economics. E-mail: ekardanova@hse.ru

Address: 20 Myasnitskaya Str., 101000 Moscow, Russian Federation.

Abstract Measuring mathematical literacy is not easy as this construct is multicomponent and tasks often involve a lot of reading. As a rule, intended users of measurement results want information about the overall level of respondents' mathematical literacy as well as its specific components. According to educational and psychological testing standards, reporting overall scores together with subscores simultaneously requires additional psychometric evaluation to provide evidence for validity of all scores reported. A study performed shows that PROGRESS-ML, a test measuring basic mathematical literacy in elementary school pupils, can be used as a one-dimensional measure, allowing overall test scores to be reported. Meanwhile, reading skills do not contribute significantly to the likelihood of item response, and subscores can be reported as complementary to the total score.

Keywords basic mathematical literacy, complex construct, composite measure, PROGRESS-ML.

For citing Federiakin D. A., Larina G. S., Kardanova E. Yu. (2021) Izmerenie bazovoy matematicheskoy gramotnosti v nachal'noy shkole [Measuring Basic Mathematical Literacy in Elementary School]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 2, pp. 199–226. https://doi.org/10.17323/1814-9545-2021-2-199-226

В современном образовании и, шире, социальных науках наблюдается растущая потребность в композитных средствах измерения — инструментах, состоящих из субшкал, которые определенным образом вкладываются (суммируются прямо или с разными весами) в итоговый тестовый балл. Такого рода исследовательские инструменты необходимы для измерения сложных конструктов, например так называемых навыков XXI века или новых компетенций. Эти конструкты состоят из различных компонентов, и их сложно представить в виде классической одномерной (однокомпонентной) характеристики респондентов. В то же время для практиков и лиц, принимающих решения на основе результатов измерений, ценной является информация об уровне развития как целостной характеристики, так и ее составных частей. Такая информация позволяет учитывать развитые и развивающиеся чер-

ты или способности респондентов, улучшая, например, работу системы образования или психологическую практику.

На языке психометрики композитные инструменты являются многомерными, и их задача состоит в том, чтобы оценить как общую способность респондентов, так и составляющие ее отдельные способности.

В «Стандартах образовательного и психологического тестирования» [American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014] отмечается, во-первых, что тестовые баллы не следует сообщать пользователям до тех пор, пока не будет установлена их валидность, сопоставимость и надежность, и, во-вторых, что если инструмент выдает более чем один тестовый балл, то психометрическое качество *всех* сообщаемых пользователю тестовых баллов должно быть установлено. Эти требования важны, поскольку как неточная информация на уровне общего тестового балла может спровоцировать решения с нежелательными социальными последствиями, так и неточная информация на уровне балла по субшкалам может привести к неправильным решениям по исправлению или улучшению ситуации [Sinharay, Puhan, Haberman, 2011]. В академической среде использование низкокачественных баллов по субшкалам может породить неверные выводы о природе изучаемого явления.

Базовая математическая грамотность — один из примеров композитных конструктов. Базовым грамотностям дано огромное количество определений, учитывающих совершенно различные аспекты этого конструкта — от предметной области до их распространенности в популяции [Фрумин и др., 2018]. Общей чертой всех этих формулировок остается то, что грамотность определяется как способность решать повседневные задачи. В силу разнообразия таких задач инструменты, направленные на измерение базовых грамотностей, погружают задания в совершенно разнородные контексты в попытках измерить в них успешность решения тех или иных задач. Именно так и возникает композитность таких инструментов (измерения базовой математической грамотности): они являются некоторой общей интегральной мерой успешности респондентов в решении различных задач (с использованием математических операций). Многие исследователи отмечают, что такая структура композитных инструментов приводит к «вмешательству» посторонних компетенций в процесс решения заданий (см., например, [Grimm, 2008]). Как следствие, одним из главных психометрических вопросов, относящихся к таким инструментам, является вопрос о возможности извлечения как общего валидного и психометрически состоятельного балла по таким инструментам, так и баллов по их компонентам.

Целью настоящей статьи является проверка и обоснование различных подходов к моделированию данных теста PROGRESS-ML,

направленного на измерение базовой математической грамотности в начальной школе. Статья построена следующим образом. Сначала мы приводим описание теоретических трудностей в измерении базовой математической грамотности. Затем рассматриваем теоретическую рамку и структуру теста PROGRESS-ML. Поскольку фокусом статьи является психометрическое моделирование этого инструмента, мы не анализируем подробно варианты осмысления теоретической рамки этого инструмента, его разработку, а также не описываем детально связь содержания этого инструмента с содержанием образования в начальной школе, нисколько не умаляя важность подобной работы. Далее мы описываем существующие психометрические подходы к моделированию композитных инструментов измерения и обосновываем выбор в пользу одного из них. После этого мы проводим анализ эмпирических данных, собранных с помощью теста PROGRESS-ML. На первом этапе этого анализа мы проверяем гипотезу о возможности сообщать пользователю общий балл базовой математической грамотности по тесту. На втором этапе мы тестируем предположение о важности читательских навыков при выполнении заданий теста на базовую математическую грамотность. Наконец, на третьем этапе мы оцениваем возможность сообщать пользователю баллы по субшкалам — отдельным тематическим областям и отдельным группам когнитивных операций. Мы завершаем данную работу психометрическим осмыслением полученных результатов и описанием их вклада в методологию образовательного и психологического тестирования.

1. Проблемы в измерении базовой математической грамотности

Математическая грамотность — компетенция, необходимая каждому в повседневной жизни: при покупке продуктов, приготовлении еды, оплате счетов. В России математика традиционно рассматривается как школьный предмет, имеющий принципиальное значение для успешного овладения другими дисциплинами и являющийся уникальным средством «интеллектуального развития в массовой школе» [Козлов, Кондаков, 2009]. Однако измерить математическую грамотность учащихся — задача непростая в силу сложности определения этого конструкта.

В мировой практике отсутствует единая и общепринятая концепция математической грамотности, которую можно было бы определить через конечный набор знаний и умений [Jablonka, 2003]. К математической грамотности относят навыки вычисления, знание и понимание фундаментальных понятий, пространственное мышление, умение решать задачи в повседневном контексте, способность разработать сложные алгоритмы решения и способность проанализировать данные на графике. Даже при оценке математической грамотности у детей в дошкольном возрасте авторы диагностических систем опираются на разные наборы из внуши-

тельного списка базовых знаний и умений: устный счет, знание чисел, сравнение чисел, счет с опорой на предметы, распознавание геометрических фигур, задачи на измерения.

Сложность определения и, соответственно, измерения математической грамотности учащихся обусловлена, в частности, зависимостью математики как школьного предмета от конкретной школьной программы, разрабатываемой в соответствии с целями, стоящими перед математическим образованием. Например, в ряде программ акцентируется внимание на системообразующей роли качественного математического образования для научно-технического прогресса (например, Концепция математического образования в России 2013 г.[1]). В рамках другого подхода — его придерживаются некоторые современные международные мониторинги качества образования, например PISA[2] [OECD, 2019], — на первый план выдвигается необходимость овладения школьником к концу обязательного этапа обучения такими навыками, которые позволят ему успешно справляться с разнообразными повседневными задачами — от покупки продуктов до расстановки мебели в квартире.

Среди исследователей достигнут консенсус только о траектории развития отдельных математических навыков в дошкольном и школьном периодах [Purpura, Baroody, Lonigan, 2013]. До сих пор остаются открытыми вопросы, например, о том, как и в какой момент происходит освоение символьных знаков — цифр и чисел [Benoit et al., 2013; Kolkman, Kroesbergen, Leseman, 2013], в какой степени развитие математических навыков опирается на другие когнитивные способности [Peng et al., 2016; Toll et al., 2015] и какие ранние способности в наибольшей степени предсказывают последующие математические достижения [Chen, Li, 2014; Libertus et al., 2016; Schneider et al., 2017]. Что касается формального обучения, навыки, которые приобретают ученики в школе, с переходом из класса в класс становятся все более разнородными в силу разных программ обучения, что затрудняет выделение единого конструкта для описания математической грамотности.

Сложность определения математической грамотности заключается и в том, что этот конструкт указывает не только на то, какими знаниями владеет учащийся, но и на то, как и в каких ситуациях он может их использовать. Разное предметное содержание оценивается с помощью заданий, предполагающих разную по интенсивности когнитивную нагрузку. Например, задача на определе-

[1] Концепция развития российского математического образования. Министерство образования и науки Российской федерации: http://www.math.ru/conc/vers/conc-3003.pdf

[2] Programme for International Students Assessment, проводится каждые три года для оценки математической, читательской и естественнонаучной грамотности учащихся в возрасте 15 лет.

ние высоты столбца на простой и знакомой ученику гистограмме требует меньших умственных усилий, чем задача на работу с информацией, представленной на графике незнакомого ему типа. Поэтому во многих исследованиях математическая грамотность рассматривается еще и с точки зрения используемых при решении задачи когнитивных операций. Например, в тесте PISA оценивается, насколько хорошо 15-летние учащиеся владеют теоретическим материалом по отдельным темам (числа, пространства и формы и др.) и как они применяют усвоенные знания на каждом этапе решения задачи [OECD, 2013].

Наконец, сложность измерения математической грамотности обусловлена еще и тем, что математические способности тесно связаны с читательской грамотностью. В лонгитюдном исследовании на выборке малообеспеченных детей из школ Чикаго установлено, что читательские навыки в 3-м классе являются значимым предиктором прироста математических способностей вплоть до 8-го класса, даже при учете предыдущих математических достижений. Наибольшие эффекты навыков чтения были обнаружены для предметных областей «решение проблем» и «анализ данных» [Grimm, 2008]. Схожие результаты получены и при анализе средних баллов учащихся из 22 стран, участвовавших в 2003 г. в двух международных исследованиях: балл по чтению в PISA в наибольшей степени связан с достижениями учеников в предметной области «анализ данных» в исследовании TIMSS[3], которая предполагает чтение и интерпретацию данных на диаграммах (r = 0,91, остальные коэффициенты варьируют от 0,57 до 0,79) [Wu, 2010]. Трудностям в освоении математических навыков (дискалькулия) зачастую сопутствует дислексия[4] [Joyner, Wagner, 2020]. Причем дети, у которых одновременно присутствуют оба эти нарушения, хуже справляются с заданиями по математике, чем дети только с дискалькулией [Jordan, Hanich, Kaplan, 2003].

Природа связи между математической и читательской грамотностью комплексна и до конца не выяснена. С одной стороны, оба этих конструкта могут опираться на общие когнитивные функции — например, показано, что ранние языковые навыки определяют развитие концепции числа в раннем детстве [Gelman, Butterworth, 2005]. Другой возможной причиной этой связи могут быть ошибки в понимании текста задания [Cummis et al., 1988]. Так, согласно классификации ошибок, совершаемых при решении текстовых задач [Newman, 1977; Casey, 1978], первые два этапа решения любой текстовой задачи прямо опираются на навыки чтения —

[3] Trends in Mathematics and Science Study, проводится каждые четыре года. Измеряет математическую и естественнонаучную грамотность учащихся 4-х и 8-х классов.

[4] Специфическое нарушение способности к овладению навыками чтения и письма.

это этапы декодирования (*decoding*) и понимания (*comprehension*) письменной информации, которые предполагают внимательное прочтение задания, понимание контекста задачи и поставленного вопроса, сбор всей необходимой информации. На эти этапы решения задачи приходится от 12 до 22% всех ошибок в решении (см., например, обзор исследований в [Clements, Ellerton, 1996]). То есть правильное выполнение любой текстовой задачи зависит от того, совершил ли учащийся ошибки на первых двух этапах решения — например, не прочитал ли случайно слово «минуты» вместо слова «минус» [Clements, 1980].

Таким образом, математическая грамотность является комплексным, не бинарным, многогранным конструктом, опирающимся на широкое разнообразие математических навыков. Представляется затруднительным поместить на одну шкалу весь процесс развития математических навыков — от неформального математического знания до усвоения сложных математических методов в старшей школе. Кроме того, математические способности ученика, измеряемые с помощью текстовых задач и графиков, в значительной степени зависят от его навыков чтения. Комплексный характер конструкта и его связь с другими интеллектуальными навыками усложняют задачу измерения математической грамотности и требуют применения сложного психометрического моделирования для валидизации инструментов.

2. Описание теста базовой математической грамотности PROGRESS-ML

Тест базовой математической грамотности PROGRESS-ML является частью инструмента PROGRESS[5], предназначенного для мониторинга базовых грамотностей у детей в начальной школе с целью сопровождения и совершенствования образовательного процесса. PROGRESS рассчитан на оценивание базовых грамотностей в нескольких областях: математическая грамотность, языковая грамотность, смысловое чтение и словарный запас. Оценивание проводится в формате компьютерного адаптивного тестирования.

PROGRESS-ML разработан с целью оценить, насколько хорошо учащийся ориентируется в математике после двух лет обучения в школе. В современном постиндустриальном мире математическое образование должно подготовить учащихся к тому, что им придется решать задачи в постоянно изменяющихся условиях — например, быстро принимать решения и адаптироваться к новым условиям задач, уметь решать незнакомые задачи и быстро ориентироваться в больших объемах информации. В таких условиях концепция базовой математической грамотности, необходимой для успешной жизни, изменяется. Например, она расширяется за счет увеличения количества контекстов, в которых необходимо обработать информацию или решить задачу. При разработке те-

[5] Инструмент разработан в Институте образования НИУ ВШЭ.

Таблица 1. **Тематические области, оцениваемые в тесте PROGRESS-ML**

Предметная область	Количество заданий	Описание
Пространственные представления	7	Задания предназначены для измерения способности школьников понимать пространственные отношения между фигурами, мысленно представлять плоские и объемные фигуры в пространстве. Для их выполнения требуется не только распознавать отдельные геометрические фигуры, но и уметь видеть новые геометрические объекты, образованные путем объединения плоских или объемных фигур в единую композицию
Измерения величин	6	Выполняя задания этого блока, учащийся демонстрирует понимание того, что число может не только показывать место объекта в последовательности, но и являться характеристикой данного объекта (длина, площадь). Задачи направлены на проверку способности оперировать числами как мерами объектов
Закономерности	6	Задания оценивают способность школьников распознавать и продолжать числовые и геометрические последовательности, проверяют степень сформированности у учащихся алгоритмических универсальных учебных действий. Для решения задач учащийся должен видеть принципы (одно или несколько правил) построения последовательностей
Моделирование	6	В этом блоке заданий измеряется способность учащихся формально выражать (с помощью чисел) модели, репрезентированные с помощью текста или геометрических последовательностей. Задания данного блока, как и блока «Закономерности», проверяют степень сформированности у учащихся алгоритмических универсальных действий. Только, в отличие от закономерностей, здесь учащийся должен не просто понять модель, но и суметь ее записать на языке математики
Работа с информацией	5	Задания оценивают способность учащихся понимать и интерпретировать информацию, представленную в таблице и на графике. Кроме того, для успешного решения заданий в этом блоке ученик должен понимать, как проводить вычисления на основе информации на графике, а также выносить суждения, привлекая дополнительную информацию

ста мы опирались на следующее определение: «Базовая математическая грамотность (включая работу с данными) — способность применять математические инструменты, аргументацию, моделирование в повседневной жизни, в том числе в цифровой среде» [Фрумин и др., 2018].

Тест состоит из 30 заданий с выбором одного варианта ответа из предложенных. Выполнение всех заданий оценивалось дихо-

Таблица 2. **Группы когнитивных операций, оцениваемых в тесте PROGRESS по математике**

Группа когнитивных операций	Количество заданий	Описание
Знание	12	Оценивается знание фактической информации по математике — фундамента для решения любых задач. Например, в задании необходимо выполнить сложение или вычитание двух целых чисел, определить величину столбца на простой гистограмме, посчитать, сколько раз геометрическая фигура помещается на картинке
Применение	14	Оценивается использование учащимися усвоенных знаний и навыков для решения задач и проблемных ситуаций, контекст и алгоритм решения которых им хорошо знакомы. Например, в задаче необходимо определить правило построения последовательности (числовой или с геометрическими фигурами)
Рассуждение	4	Задания требуют тщательного анализа предоставленной информации, чтобы связать факты из нескольких областей знаний и рассмотреть несколько вариантов решения. Эти задания незнакомы учащимся и поэтому требуют большего их внимания

томически. Содержание теста отбиралось таким образом, чтобы оно, с одной стороны, отвечало определению базовой математической грамотности, а с другой — учитывало содержание программы начального общего образования. В результате были выделены пять тематических областей (табл. 1). Задания в тесте сгруппированы в блоки в соответствии с тематической областью.

Дополнительно тест PROGRESS-ML оценивает когнитивные процессы учащихся, необходимые для выполнения заданий. При разработке теста мы опирались на группы когнитивных операций, выделенные в теоретической рамке международного исследования TIMSS для 4-го класса [Harris, Martin, 2013]: знание (*knowing*), применение (*applying*), рассуждения (*reasoning*). При этом тест разрабатывался таким образом, чтобы его задания согласовывались с моделями современной теории тестирования из семейства моделей Раша.

Предложенная в TIMSS таксономия когнитивных операций схожа с таксономией учебных действий Блума [Bloom, 1956], однако не идентична ей и выделяет только три группы когнитивных операций, а не шесть. Кроме того, таксономия TIMSS имеет и принципиальное отличие: три группы когнитивных операций не упорядочены в плане возрастания трудности или абстрактности операций. Таким образом, внутри каждой группы когнитивных операций присутствуют задания разной трудности.

Решение заданий в тесте PROGRESS-ML включает весь спектр когнитивных процессов таксономии TIMSS. Выполнение большей части заданий в тесте опирается в равных пропорциях на группы когнитивных процессов «знание» и «применение», а решение оставшейся небольшой части заданий — на рассуждения[6] (табл. 2).

Часть заданий теста PROGRESS-ML (около 50%) могут рассматриваться как нагруженные чтением, так как для их решения необходимо прочитать и понять условие.

Таким образом, тест PROGRESS-ML является сложным по структуре инструментом: он включает пять тематических областей и отражает три группы когнитивных операций. Такие инструменты называются композитными. Предполагается, что по итогам тестирования пользователю сообщаются общий тестовый балл учащегося (в данном случае уровень его базовой математической грамотности) и баллы по субшкалам (в данном случае это могут быть тематические области и/или когнитивные операции).

3. Психометрика композитных инструментов измерения

Психометрическое моделирование композитных инструментов состоит из нескольких этапов. В первую очередь необходимо проверить, является ли тест существенно одномерным. Если да, то мы можем сообщать пользователю общий балл по тесту — разумеется, при условии, что будет доказана его валидность и психометрическая состоятельность. Если тест не является одномерным, то необходимо использовать многомерные модели, и в этом случае сообщать общий балл можно только после дополнительного исследования с использованием иерархических моделей [Schmid, Leiman, 1957]. Особой популярностью пользуются два класса иерархических моделей: бифакторные модели [Reise, 2012] и модели с факторами высокого порядка [Gignac, 2008]. Несмотря на алгебраические сходства и то, что обе группы моделей предполагают использование тестового балла по всему тесту, интерпретации этих моделей различаются [Rijmen, 2010; Mansolf, Reise, 2017]. Модели с факторами высокого порядка оценивают общий фактор, который проявляется в заданиях через баллы по субшкалам, бифакторные модели предполагают полное разделение эффектов общего фактора и баллов по субшкалам.

Если авторы инструмента намерены помимо общего балла сообщать дополнительно баллы по отдельным субшкалам (например, когнитивным операциям или содержательным областям), возможны несколько подходов к моделированию. Первый состоит в том, чтобы одномерную модель применять к каждой субшкале отдельно [Davey, Hirsch, 1991]. Однако число заданий в каждой

[6] Правильность отнесения заданий к каждой группе когнитивных операций была подтверждена экспертами.

субшкале, как правило, невелико, и поэтому надежность измерения будет недостаточно высокой, а ошибка измерения слишком велика. При таких условиях баллы по отдельным субшкалам сообщать будет неправомерно [American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014].

Второй подход предполагает применение некомпенсаторных многомерных моделей [Reckase, 2009]. Такие модели фактически представляют собой несколько одномерных моделей, записанных в одном уравнении правдоподобия. Каждая латентная характеристика переменных вычисляется на основе ответов респондентов только на соответствующие задания и с учетом оцененных корреляций между самими латентными переменными. Таким образом, многомерные модели используют информацию о каждой размерности и моделируют вероятность выполнения заданий как функцию не одной латентной переменной, а нескольких, с учетом связей между ними. В итоге надежность измерений будет выше, чем при первом подходе, и более вероятно, что можно будет сообщать баллы по отдельным субшкалам. Применение некомпенсаторных многомерных моделей может рассматриваться в контексте анализа коллатеральной информации — любой информации об инструменте, респондентах или их взаимодействии, включение которой в измерительную модель не меняет интерпретацию параметров, но которая позволяет уменьшить неопределенность в оценках этих параметров [Wang, Chen, Cheng, 2004].

Третий подход предполагает использование бифакторных моделей. Гипотетически они допускают одновременное сообщение общего балла и баллов по субшкалам в качестве дополнительной независимой информации. Однако, как показывают исследования, баллы по субшкалам, оцененные в рамках бифакторных моделей, редко обладают удовлетворительной надежностью, потому что они описывают информацию, которая не была извлечена с помощью балла по всему тесту, и она часто бывает подавлена случайными шумами [Haberman, Sinharay, 2010].

Таким образом, необходимым условием использования результатов композитных инструментов измерения является проведение психометрических исследований, на основании которых принимается решение, являются ли результаты по всему тесту и его субшкалам надежными и можно ли их сообщать пользователям.

3.1. Методология исследования

3.1.1. Выборка и процедура

Выборку составили 6078 учеников 3-х классов школы в двух регионах РФ. Относительно региональной совокупности третьеклассников выборки были репрезентативными. Средний возраст учащихся — 9,06 года (SD = 0,46), доля девочек в выборке — 52,36%. Проведено компьютеризированное адаптивное тестирование с правилом остановки.

3.1.2. Применяемые методы анализа данных

Весь психометрический анализ проводился в рамках современной теории тестирования (*Item Response Theory*) [van der Linden, 2018]. Она постулирует, что как задания, так и респонденты характеризуются ненаблюдаемыми параметрами, взаимодействие которых задает вероятность наблюдения правильного ответа каждого респондента на каждое задание теста. Конкретно анализ проводился в парадигме Раш-моделирования [Rasch, 1993], которое подразумевает, что задания различаются только одним параметром — трудностью, в отличие от иных моделей современной теории тестирования, в которых задания описываются большим количеством параметров [Wright, Stone, 1979]. При выборе психометрического подхода аргументами в пользу Раш-моделирования стали свойство специфической объективности, гарантирующее разделение параметров респондентов и заданий, однозначная иерархия заданий по трудности на всем континууме способности и численная стабильность моделей.

Важное преимущество Раш-моделирования — возможность оценить дисперсию ненаблюдаемых способностей респондентов [Kamata, 2001]. В случае если оценка дисперсии по какой-либо способности близка к нулю, эта способность не нагружает задания в достаточной мере, чтобы сказываться на вероятности наблюдения правильных ответов. Кроме того, как и все многомерные модели современной теории тестирования, примененные модели позволяют напрямую оценить корреляции и дисперсии ненаблюдаемых способностей, которые «очищены» от дисперсии случайной ошибки в распределении баллов по заданиям [De Boeck, Wilson, 2004].

Мы использовали одномерные и многомерные модели для проверки серии гипотез о возможных ненаблюдаемых способностях, которые требуются для решения заданий.

Для проверки гипотезы о возможности сообщать пользователю общий балл по тесту мы использовали одномерную модель Раша [Wright, Stone, 1979]. Мы исследовали размерность теста с помощью метода главных компонент (МГК), примененного к стандартизированным модельным остаткам, представляющим собой нормированные уклонения наблюдаемого ответа респондента от его математического ожидания согласно используемой модели [Linacre, 1998; Smith, 2002]. Таким образом, применяя МГК к остаткам, мы декомпозируем дисперсию, необъясненную с помощью модели, на отдельные компоненты. Теоретически, если предположение об одномерности теста поддерживается, корреляции между остатками должны быть близки к нулю. В этом случае в результате применения МГК не будут выделены компоненты, систематически объясняющие больше дисперсии, чем остальные. Соответственно распределение объясненной компонентами дисперсии будет близко к равномерному. Также принято считать, что, если собственное значение компонента, объясняющего наибольшее

количество дисперсии, меньше 2, то распределение остатков является случайным шумом и тест можно считать одномерным [Linacre, 2021]. В противном случае можно говорить о наличии второй размерности в тесте.

Согласие данных с моделью на уровне отдельных заданий оценивалось с помощью невзвешенной (*OutFit*) и информационно-взвешенной (*InFit*) среднеквадратических (MnSq) статистик согласия [Linacre, 2002]. Эти статистики также основаны на анализе стандартизированных остатков [Wright, Masters, 1990] и имеют математическое ожидание, равное 1. В литературе продуктивным для измерений называется интервал значений статистик согласия (0,75; 1,33), который и рассматривался в данном исследовании в качестве допустимого для этих статистик [Adams, Khoo, 1996].

Наконец, психометрическая состоятельность тестового балла по всему тесту оценивалась через определение его надежности и ошибки измерения.

Для проверки гипотезы о важности навыков чтения при решении задач по математике мы использовали модели с внутренней многомерностью заданий (*within-item multidimensionality*) [Adams, Wilson, Wang, 1997]. Эти модели подразумевают, что для выполнения заданий необходимо иметь несколько ненаблюдаемых способностей (например, по математике и по чтению), линейная комбинация которых задает действительную вероятность правильного ответа.

Для анализа навыков чтения выделены 15 заданий с максимальной текстовой нагрузкой в условиях. Далее была построена неполная бифакторная модель Раша, которая допускает, что математическая грамотность измеряется всеми заданиями теста, а выделенная группа заданий оценивает также навыки чтения. Эта модель может быть рассмотрена как расширенная тестлет-модель Раша (*extended Rasch testlet model*) [Paek et al., 2009] из класса косоугольных бифакторных моделей современной теории тестирования, в которой корреляция двух ненаблюдаемых способностей может быть оценена напрямую. Таким образом, основная размерность, нагружающая все задания, может быть интерпретирована как базовая математическая грамотность, из которой изъят вклад чтения. В то же время дополнительная способность, нагружающая выбранные 15 заданий, может быть проинтерпретирована как навыки чтения, которые проявляются в тесте базовой математической грамотности.

Наконец, для проверки гипотезы о возможности сообщать баллы по отдельным субшкалам мы использовали многомерные модели Раша без кросс-нагрузок (*between-item multidimensional models*) [Adams, Wilson, Wang, 1997]. Они подразумевают, что каждое задание нагружено только одной ненаблюдаемой способностью и при этом не используют допущение о наличии общего фактора. Выбор в пользу модели без общего фактора обусловлен крити-

кой, которой подвергаются модели с общим фактором — в частности, за то, что тестовые баллы по субшкалам, оцененные в рамках этих моделей, непригодны для практического использования. Для бифакторных моделей это означает низкую надежность баллов по субшкалам и ограничения, наложенные на матрицу дисперсий-ковариаций ненаблюдаемых переменных, которые затрудняют интерпретацию всех полученных баллов [Bonifay, Lane, Reise, 2017]. В то же время модели с факторами более высокого уровня вообще не предполагают использование баллов по субшкалам, поскольку в них основная информация о компонентах конструкта уже описывается общим фактором.

Все описанные многомерные модели сравнивались по их согласию с данными с одномерной моделью Раша, которая принята за базовую. Согласие данных с моделями (*Global Fit*) оценивалось с помощью информационного критерия Акаике (AIC) [Akaike, 1974]) и байесовского информационного критерия (*Bayesian information criterion*, BIC) [Schwarz, 1978]. Эти информационные критерии вводят пенальти в оценку согласия модели с данными за дополнительные параметры (AIC) с учетом размера выборки (BIC). Эти индексы показывают относительное согласие моделей с данными: меньшие их значения указывают на предпочтительную модель. Согласие данных с моделью на уровне отдельных заданий (*Local Fit*) оценивалось с помощью невзвешенной (*OutFit*) и информационно-взвешенной (*InFit*) статистик согласия, описанных выше.

Для оценки надежности во всех моделях мы использовали *Expected-a-Posteriori* (EAP) метод оценки параметров респондентов [Bock, Mislevy, 1982]. Метод EAP особенно эффективен в случаях многомерных инструментов измерения, поскольку привлекает информацию о многомерном распределении респондентов и весь профиль ответов для оценки способности по каждой из размерностей. Применение этого метода в данном случае оправданно, поскольку для данного инструмента не предполагается сбор данных по отдельным размерностям без использования остальных. Таким образом, использование многомерного моделирования в этом направлении адресует к другому использованию коллатеральной информации. В этом случае для каждой из субшкал ответы по всем остальным субшкалам (вместе с матрицей корреляций ненаблюдаемых размерностей) являются коллатеральной информацией [Wu, Tam, Jen, 2016]. Результатом становится повышение надежности измерений. Стандартное отклонение оцененного апостериорного распределения может быть принято за стандартную ошибку измерения способности респондента. Отношение этой дисперсии ошибки к оцененной дисперсии латентной способности задает итоговую оценку надежности [Adams, 2005].

Все использованные модели можно рассматривать как частные случаи многомерной мультиномиальной логит-модели смешанных эффектов (*Multidimensional random coefficients multinomial*

Таблица 3. **Анализ согласия заданий с одномерной моделью Раша**

Блок	Задание	Количество ответов	Оценка трудности	Стандартная ошибка	InFit MnSq	OutFit MnSq
Пространственные представления	I01	6041	–1,18	0,03	0,97	0,96
	I02	5975	–0,56	0,03	1,03	1,04
	I03	5997	–0,23	0,03	1,05	1,06
	I04	5430	0,37	0,03	1,04	1,06
	I05	4987	0,69	0,03	1,00	0,99
	I06	4125	1,19	0,04	1,08	1,14
	I07	3098	1,94	0,05	0,98	0,97
Измерения величин	I08	5843	–1,51	0,03	1,07	1,13
	I09	5860	–1,47	0,03	1,03	1,05
	I10	5811	–1,69	0,04	0,95	0,88
	I11	5626	–0,76	0,03	1,02	1,01
	I12	5375	–0,65	0,03	0,93	0,89
	I13	5080	–0,59	0,03	0,95	0,92
Закономерности	I14	5560	–2,41	0,05	0,95	0,87
	I15	5473	–2,06	0,04	1,02	1,04
	I16	5340	–1,42	0,04	0,96	0,92
	I17	5009	–1,16	0,03	1,00	1,03
	I18	4755	–0,56	0,03	1,06	1,07
	I19	4398	–0,12	0,03	1,03	1,04
Моделирование	I20	4603	–0,62	0,03	0,98	0,98
	I21	4263	–0,41	0,03	0,89	0,86
	I22	3826	–0,37	0,04	1,10	1,14
	I23	3013	1,78	0,05	1,04	1,11
	I24	2423	0,72	0,05	1,06	1,12
	I25	1702	0,73	0,05	1,00	1,01
Работа с информацией	I26	2808	–2,31	0,06	0,98	0,98
	I27	2469	–2,20	0,06	0,93	0,83
	I28	2320	–1,35	0,05	0,89	0,83
	I29	1969	–1,60	0,06	0,88	0,76
	I30	1708	1,05	0,06	0,95	0,93

logit model) [Adams, Wilson, Wang, 1997]. Для оценки всех моделей использовался квази-Монте-Карло-алгоритм, внедренный в пакет TAM v. 3.5–19 [Robitzsch, Kiefer, Wu, 2020] для программного обеспечения R v. 3.6.2. Все модели были оценены методом максимального маргинального правдоподобия, накладывающим параметрическое допущение о нормальности распределения на параметры респондентов [Bock, Aitkin, 1981]. Все модели были идентифицированы фиксацией среднего выборки по каждой размерности, равной нулю. Это особенно важно для идентификации моделей с внутренней многомерностью заданий. Для анализа размерности применялся пакет psych v. 1.9.12.31 [Revelle, 2020]. Для вычисления остатков использовались оценки параметров респондентов, полученные с помощью метода *Warm's weighted maximum likelihood* [Warm, 1989]. Именно этот метод выбран для вычисления остатков ввиду наименьшего смещения полученных оценок по сравнению с другими методами, популярными в современной теории тестирования.

4. Результаты

4.1. Проверка гипотезы о возможности сообщать общий балл по тесту

Результаты анализа согласия заданий теста с одномерной моделью Раша приведены в табл. 3. Все задания находятся в удовлетворительном согласии с моделью: значения статистик согласия всех заданий не выходят за допустимые пределы.

В результате анализа стандартизированных остатков методом главных компонент обнаружено, что собственное значение первого компонента равняется 1,45, что соответствует 4,2% дисперсии остатков. Собственные значения следующих четырех компонент находятся в промежутке (1,15, 1,20), и распределение дисперсии среди компонент практически равномерное — около 4%. Следовательно, одномерная модель в достаточной мере описывает распределение вероятностей ответа и тест можно считать одномерным.

EAP-надежность измерения математической грамотности из одномерной модели составила 0,76 (при дисперсии 0,93). Классическая надежность теста (альфа Кронбаха) высокая — 0,81.

Таким образом, на основании проведенного анализа тест может рассматриваться как одномерный, даже несмотря на разные способы группирования заданий, — а значит, по результатам тестирования можно сообщать потребителю один общий тестовый балл математической грамотности, который будет обладать хорошими психометрическими характеристиками.

Одномерная модель Раша рассматривалась в качестве базовой модели для сравнения со всеми остальными моделями.

4.2. Проверка гипотезы о вкладе навыков чтения

В табл. 4 приведены результаты сравнения одномерной модели Раша с моделью Раша, откалиброванной для оценки вклада навыка чтения в вероятность выполнения заданий.

Таблица 4. **Сравнение базовой модели и модели для анализа вклада чтения**

Модель	Логарифмическое правдоподобие	Выборка	Количество параметров	AIC	BIC
Одномерная	144 255,6	6078	31	144 318	144 526
С выделенным чтением	142 638,5		33	142 705	142 926

По согласию с данными модель с выделенным чтением выглядит предпочтительнее. Тем не менее оцененная дисперсия чтения составила 0,02, что в 52,35 раза меньше, чем оцененная дисперсия математической грамотности — 0,89. Похожая разница найдена и в оценке надежностей — в 41,83 раза: надежность математической грамотности составила 0,75, надежность чтения — 0,01. Следовательно, респонденты не различаются по дополнительной способности, которую образуют выделенные 15 заданий. Далее, линейная корреляция Пирсона размерности чтения с размерностью математической грамотности незначимо отличается от нуля (r = 0,01, p > 0,05 согласно t-тесту для корреляции Пирсона). Этот результат не согласуется с предыдущими исследованиями (например, [Grimm, 2008]). Объяснение, возможно, состоит в низкой дисперсии и, как следствие, низкой надежности измерения чтения: при таких оценках надежности и дисперсии различия в навыках чтения между респондентами полностью подавлены случайными флуктуациями. На основе проведенного анализа можно заключить, что в рассматриваемом тесте вклад чтения может иметь место, как и в других инструментах, однако он настолько мал, что неидентифицируем в принципе.

4.3. Проверка гипотезы о возможности сообщать баллы по отдельным субшкалам

В табл. 5 приведены результаты сравнения одномерной модели Раша с моделями Раша, откалиброванными для подтверждения спроектированных тематических областей и групп когнитивных операций.

Приведенные в табл. 5 данные свидетельствуют, что любая из двух конкурентных моделей подходит данным лучше, чем одномерная модель. Следовательно, заложенная при разработке заданий таксономия действительно провоцирует респондентов на ожидаемое поведение.

Результаты анализа матриц дисперсии-ковариации и надежностей каждой из размерностей для использованных моделей приведены в табл. 6 и 7.

Представленные в табл. 6 показатели позволяют заключить, что, во-первых, все размерности обладают надежностью, достаточной для мониторингового использования инструмента. Несмотря на малочисленность заданий, применение теста базовой мате-

Таблица 5. **Сравнение одномерной модели Раша с моделями Раша для тематических областей и когнитивных операций**

Модель	Логарифмическое правдоподобие	Выборка	Количество параметров	AIC	BIC
Одномерная	144 255,6	6078	31	144 318	144 526
Тематические области	143 875,7		45	143 966	144 268
Когнитивные операции	143 965,4		36	144 037	144 279

Таблица 6. **Надежности, дисперсии и корреляции для модели с тематическими областями**

Размерность (тематическая область)	Пространственные представления	Измерения	Закономерности	Моделирование	Работа с информацией
Пространственные представления		0,85	0,80	0,83	0,80
Измерения			0,85	0,90	0,83
Закономерности				0,86	0,84
Моделирование					0,83
Дисперсия	0,89	1,23	1,12	1,06	2,95
Надежность	0,68	0,71	0,67	0,68	0,63
Число заданий	7	6	6	6	5

матической грамотности для мониторингов возможно благодаря используемому методу оценки надежности и многомерности инструмента. Во-вторых, все тематические области коррелируют друг с другом приблизительно одинаково — на уровне 0,8–0,9, добавляя аргументы в пользу одномерности теста, даже несмотря на то что многомерная модель статистически лучше подходит данным, чем одномерная. Эти результаты означают, что общий фактор математической грамотности одинаково проявляется при выполнении заданий из любой тематической области.

Для модели с группами когнитивных операций выводы аналогичны (табл. 7): размерности обладают надежностью, достаточной для мониторингового использования инструмента. При этом стоит отметить размерность «рассуждение», состоящую всего из четырех дихотомических заданий. Такое малое количество заданий фактически делает сырые тестовые баллы по этой шкале неиспользуемыми, в отличие от баллов по этой латентной размерности. Ре-

Таблица 7. **Надежности, дисперсии и корреляции для модели с группами когнитивных операций**

Размерность (группа когнитивных операций)	Знание	Применение	Рассуждение
Знание		0,95	0,85
Применение			0,85
Дисперсия	1,37	0,82	0,60
Надежность	0,75	0,74	0,61
Число заданий	12	14	4

зультаты анализа таблицы корреляций когнитивных операций также поддерживают гипотезу о существенной одномерности теста.

5. Заключение

В социальных науках приобретают все большую популярность композитные инструменты измерения, которые призваны выдавать как единый тестовый балл, так и баллы по субшкалам. Одной из стратегий применения результатов по таким инструментам может являться использование сырых тестовых баллов [Wilson, Gochyyev, 2020]. Однако для этого необходимы психометрические исследования, направленные на изучение того, как много информации сырые тестовые баллы дают в дополнение к общему тестовому баллу [Haberman, 2005]. При этом в большинстве случаев сырые тестовые баллы психометрически невозможно использовать, в частности из-за низкой их надежности [Haberman, Sinharay, 2010].

Другой, более популярной стратегией является применение сложных психометрических моделей, интерпретация которых зачастую неясна практикам [Bonifay, Lane, Reise, 2017]. Это касается в первую очередь бифакторных моделей, для оценки параметров которых требуются сильные допущения, затрудняющие интерпретацию получаемых тестовых баллов [Wilson, Gochyyev, 2020]. Даже несмотря на недавний прогресс в области косоугольных бифакторных моделей, выработка единой традиции интерпретации и дальнейшее психометрическое изучение этих моделей все еще далеки от завершения [Kanonire, Federiakin, Uglanova, 2020]. Другой возможностью является использование моделей с факторами более высокого порядка [Gignac, 2008]. В этих моделях индикаторами общего фактора являются баллы по субшкалам. Однако такие модели вообще не предполагают использования баллов по субшкалам, что ограничивает их полезность для практиков, нисколько не умаляя их академической полезности.

Потребность практиков в баллах по субшкалам чрезвычайно высока, потому что эта информация позволяет дать заключение

не только о том, каков уровень достижений респондента по конструкту, но и о том, как именно он его достиг. Чтобы удовлетворить эту потребность, исследователи, проводящие международные сравнения школьного образования, используют тот факт, что средний балл одного респондента по всем субшкалам из многомерной модели равняется его баллу из одномерной модели с учетом их независимого линейного перевода на шкалы с одними и теми же численными параметрами (например, средним значением 500 и стандартным отклонением 100) [Foy, Yin, 2015]. Это позволяет исследователям не ограничивать интерпретацию одной-единственной моделью и избежать использования тяжело запараметризованных психометрических моделей [Brandt, Duckor, Wilson, 2014].

К похожей стратегии мы прибегли, чтобы сообщить пользователям результаты по математическому тесту PROGRESS-ML. Для обоснования возможности использования общего балла по тесту мы провели анализ одномерности теста. Для этого использованы метод главных компонент, примененный к модельным остаткам, а также статистики согласия заданий с моделью. Мы показали, что этот тест может быть использован в качестве одномерного инструмента измерения и, следовательно, единый балл базовой математической грамотности можно докладывать пользователям результатов.

Далее мы оценивали возможность сообщать баллы по субшкалам в дополнение к общему тестовому баллу. Поскольку математическая грамотность — сложный многокомпонентный конструкт, тестовые баллы по его компонентам обладают высокой ценностью для практиков и пользователей результатов. Поэтому для повышения прикладной ценности результатов тестирования согласно требованиям «Стандартов психологического и образовательного тестирования» [American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014] мы проверили возможность использования баллов по субшкалам теста. Для этого мы выбрали подход, который наиболее подходит данному инструменту измерения, — рекалибровка данных в рамках других моделей — и показали, что полученные баллы по компонентам конструкта обладают психометрической состоятельностью и их можно сообщать пользователям.

Таким образом, главным результатом тестирования является общий балл респондента. Однако повторные рекалибровки данных с учетом различных тематических областей и разных групп когнитивных операций, необходимых для решения заданий, позволяют установить, каким способом целевой тестовый балл был набран. Сутью такого использования результатов является декомпозиция общего тестового балла на компоненты, которые его составляют. При этом информация о корреляциях субшкал позволяет использовать даже относительно маленькие шкалы (например,

шкала «рассуждение» из модели для когнитивных операций, состоящая всего из четырех заданий) с достаточно высокой надежностью.

Дополнительно мы оценили вклад чтения в вероятность решения заданий. Для этого применялись экспертные оценки нагруженности заданий чтением, что позволило выделить потенциальную вторую размерность, оценить ее дисперсию и корреляцию с основной размерностью. Обнаружено, что чтение не вносит значимого вклада в вероятность решения заданий. Опробованный подход обладает большим потенциалом для генерализации и может быть применен в других инструментах для анализа вклада посторонних размерностей.

Таким образом, в статье описаны психометрические свойства инструмента измерения математической базовой грамотности PROGRESS-ML. В результате трех этапов исследования мы показали, что 1) данный тест может быть использован в качестве одномерного инструмента измерения, что дает возможность сообщать пользователю общий тестовый балл по тесту; 2) чтение не вносит значимого вклада в вероятность решения заданий; 3) полученные баллы по отдельным компонентам теста могут самостоятельно сообщаться пользователям результатов тестов в дополнение к общему баллу.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19–29–14110.

Литература

1. Козлов В. В., Кондаков А. М. (ред.) (2009) Фундаментальное ядро содержания общего образования. М.: Просвещение. https://kpfu.ru/docs/F1999935214/fundamentalnoe.yadro.pdf
2. Фрумин И. Д., Добрякова М. С., Баранников К. А., Реморенко И. М. (2018) Универсальные компетентности и новая грамотность: чему учить сегодня для успеха завтра. Предварительные выводы международного доклада о тенденциях трансформации школьного образования. М.: НИУ ВШЭ. https://ioe.hse.ru/data/2018/07/12/1151646087/2_19.pdf
3. Adams R. J. (2005) Reliability as a Measurement Design Effect // Studies in Educational Evaluation. Vol. 31. No 2–3. P. 162–172. doi:https://doi.org/10.1016/j.stueduc.2005.05.008.
4. Adams R. J., Khoo S. T. (1996) Quest. Melbourne, Australia: Australian Council for Educational Research.
5. Adams R. J., Wilson M., Wang W. C. (1997) The Multidimensional Random Coefficients Multinomial Logit Model // Applied Psychological Measurement. Vol. 21. No 1. P. 1–23. doi:https://doi.org/10.1177/0146621697211001.
6. Akaike H. (1974) A New Look at the Statistical Model Identification // IEEE Transactions on Automatic Control. Vol. 19. No 6. P. 716–723. doi:https://doi.org/10.1109/TAC.1974.1100705.
7. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014) Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

8. Benoit L., Lehalle H., Molina M., Tijus C., Jouen F. (2013) Young Children's Mapping between Arrays, Number Words, and Digits // Cognition. Vol. 129. No 1. P. 95–101. doi:https://doi.org/10.1016/j.cognition.2013.06.005.
9. Bloom B. S. (ed.) (1956) Taxonomy of Educational Objectives: The Classification of Educational Goals. New York: Longmans, Green and Company.
10. Bock R. D., Aitkin M. (1981) Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm // Psychometrika. Vol. 46. No 4. P. 443–459. doi:https://doi.org/10.1007/BF02293801.
11. Bock R. D., Mislevy R. J. (1982) Adaptive EAP Estimation of Ability in a Microcomputer Environment // Applied Psychological Measurement. Vol. 6. No 4. P. 431–444. doi:https://doi.org/10.1177/014662168200600405.
12. Bonifay W., Lane S. P., Reise S. P. (2017) Three Concerns with Applying a Bifactor Model as a Structure of Psychopathology // Clinical Psychological Science. Vol. 5. No 1. P. 184–186. doi:https://doi.org/10.1177/2167702616657069.
13. Brandt S., Duckor B., Wilson M. (2014) A Utility Based Validation Study for the Dimensionality of the Performance Assessment for California Teachers (PACT). Paper presented at Annual Conference of the American Educational Research Association. https://www.researchgate.net/publication/281645866_A_Utility_Based_Validation_Study_for_the_Dimensionality_of_the_Performance_Assessment_for_California_Teachers_PACT
14. Casey D. P. (1978) Failing Students: A Strategy of Error Analysis // P. Costello (ed.) Aspects of Motivation. Melbourne: Mathematical Association of Victoria. P. 295–306.
15. Chen Q., Li J. (2014) Association between Individual Differences in Non-Symbolic Number Acuity and Math Performance: A Meta-Analysis // Acta Psychologica. Vol. 148. May. P. 163–172. doi:https://doi.org/10.1016/j.actpsy.2014.01.016.
16. Clements M. A.K. (1980) Analyzing Children's Errors on Written Mathematical Tasks // Educational Studies in Mathematics. Vol. 11. No 1. P. 1–21. doi:https://doi.org/10.2307/3482042.
17. Clements M. A., Ellerton N. (1996) The Newman Procedure for Analysing Errors on Written Mathematical Tasks. https://compasstech.com.au/ARNOLD/PAGES/newman.htm
18. Cummins D. D., Kintsch W., Reusser K., Weimer R. (1988) The Role of Understanding in Solving Word Problems // Cognitive Psychology. Vol. 20. No 4. P. 405–438. doi:https://doi.org/10.1016/0010-0285(88)90011-4.
19. Davey T., Hirsch T. M. (1991) Concurrent and Consecutive Estimates of Examinee Ability Profiles. Paper presented at the Annual Meeting of the Psychometric Society (New Brunswick, NJ).
20. De Boeck P., Wilson M. (2004) Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach. New York, NY: Springer Science & Business Media.
21. Foy P., Yin L. (2015) Scaling the TIMSS2015 Achievement Data // M. O. Martin, I.V.S. Mullis, M. Hooper (eds) Methods and Procedures in TIMSS2015. Chestnut Hill, MA: Boston College. P. 13.1–13.62. https://timssandpirls.bc.edu/publications/timss/2015-methods/chapter-13.html
22. Gelman R., Butterworth B. (2005) Number and Language: How Are They Related? // Trends in Cognitive Sciences. Vol. 9. No 1. P. 6–10. doi:https://doi.org/10.1016/j.tics.2004.11.004.
23. Gignac G. E. (2008) Higher-Order Models versus Direct Hierarchical Models: g as Superordinate or Breadth Factor? // Psychology Science Quarterly. Vol. 50. No 1. P. 21–43. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.531.4178&rep=rep1&type=pdf
24. Grimm K. J. (2008) Longitudinal Associations between Reading and Mathematics Achievement // Developmental Neuropsychology. Vol. 33. No 3. P. 410–426. doi:https://doi.org/10.1080/87565640801982486.

25. Haberman S. J. (2005) When Can Subscores Have Value? ETS RR-05-08. Princeton, NJ: ETS. doi:https://doi.org/10.1002/j.2333-8504.2005.tb01985.x
26. Haberman S. J., Sinharay S. (2010) Reporting of subscores using multidimensional item response theory // Psychometrika. Vol. 75. No. 2. P. 209–227. doi: https://doi.org/10.1007/s11336-010-9158-4.
27. Harris K. M., Martin M. O. (eds) (2013) TIMSS2015 Assessment Framework. https://timssandpirls.bc.edu/timss2015/downloads/T15_Frameworks_Full_Book.pdf
28. Jablonka E. (2003) Mathematical Literacy // A. J. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick, F.K.S. Leung (eds) Second International Handbook of Mathematics Education. Dordrecht, Netherlands: Kluwer Academic. P. 75–102. doi:https://doi.org/10.5951/mtlt.2019.0397.
29. Jordan N. C., Hanich L. B., Kaplan D. (2003) A Longitudinal Study of Mathematical Competencies in Children with Specific Mathematics Difficulties versus Children with Comorbid Mathematics and Reading Difficulties // Child Development. Vol. 74. No 3. P. 834–850. doi:https://doi.org/10.1111/1467-8624.00571.
30. Joyner R. E., Wagner R. K. (2020) Co-Occurrence of Reading Disabilities and Math Disabilities: A Meta-Analysis // Scientific Studies of Reading: The Official Journal of the Society for the Scientific Study of Reading. Vol. 24. No 1. P. 14–22. doi:https://doi.org/10.1080/10888438.2019.1593420.
31. Kamata A. (2001) Item Analysis by the Hierarchical Generalized Linear Model // Journal of Educational Measurement. Vol. 38. No 1. P. 79–93. doi:https://doi.org/10.1111/j.1745-3984.2001.tb01117.x.
32. Kanonire T., Federiakin D. A., Uglanova I. L. (2020) Multicomponent Framework for Students' Subjective Well-Being in Elementary School // School Psychology Quarterly. Vol. 35. No 5. P. 321–331. doi:https://doi.org/10.1037/spq0000397.
33. Kolkman M. E., Kroesbergen E. H., Leseman P. P.M. (2013) Early Numerical Development and the Role of Non-Symbolic and Symbolic Skills // Learning and Instruction. Vol. 25. June. P. 95–103. doi:https://doi.org/10.1016/j.learninstruc.2012.12.001.
34. Libertus M. E., Odic D., Feigenson L., Halberda J. (2016) The Precision of Mapping between Number Words and the Approximate Number System Predicts Children's Formal Math Abilities // Journal of Experimental Child Psychology. Vol. 150. October. P. 207–226. doi:https://doi.org/10.1016/j.jecp.2016.06.003.
35. Linacre J. M. (1998) Structure in Rasch Residuals: Why Principal Components Analysis // Rasch Measurement Transactions. Vol. 12. No 2. P. 636. https://www.rasch.org/rmt/rmt122m.htm
36. Linacre J. M. (2002) What Do Infit and Outfit, Mean-Square and Standardized Mean // Rasch Measurement Transactions. Vol. 16. No 2. P. 878. https://www.rasch.org/rmt/rmt162f.htm
37. Linacre J. M. (2021) A User's Guide to Winsteps and Ministep: Rasch-Model Computer Programs. Program Manual 4.8.0. https://www.winsteps.com/manuals.htm
38. Linden W. J. van der (ed.) (2018) Handbook of Item Response Theory. Boca Raton, FL: CRC.
39. Mansolf M., Reise S. P. (2017) When and Why the Second-Order and Bifactor Models Are Distinguishable // Intelligence. Vol. 61. February. P. 120–129. doi: https://doi.org/10.1016/j.intell.2017.01.012.
40. Newman M. A. (1977) An Analysis of Sixth-Grade Pupils' Errors on Written Mathematical Tasks // M. A. Clements, J. Foyster (eds) Research in Mathematics Education in Australia. Melbourne: Swinburne College. Vol. 1. P. 239–258.
41. OECD (2013) PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy. Paris: OECD. doi: https://doi.org/10.1787/9789264190511-en.
42. OECD (2019) PISA 2018 Assessment and Analytical Framework. Paris: OECD. doi: https://doi.org/10.1017/CBO9781107415324.004.

43. Paek I., Yon H., Wilson M., Kang T. (2009) Random Parameter Structure and the Testlet Model: Extension of the Rasch Testlet Model // Journal of Applied Measurement. Vol. 10. No 4. P. 394–407. https://bearcenter.berkeley.edu/sites/default/files/Wilson%20%2327.pdf
44. Peng P., Namkung J., Barnes M., Sun C. (2016) A Meta-Analysis of Mathematics and Working Memory: Moderating Effects of Working Memory Domain, Type of Mathematics Skill, and Sample Characteristics // Journal of Educational Psychology. Vol. 108. No 4. P. 455–473. doi:https://doi.org/10.1037/edu0000079.
45. Purpura D.J., Baroody A.J., Lonigan C.J. (2013) The Transition from Informal to Formal Mathematical Knowledge: Mediation by Numeral Knowledge // Journal of Educational Psychology. Vol. 105. No 2. P. 453–464. doi:https://doi.org/10.1037/a0031753.
46. Rasch G. (1993) Probabilistic Models for Some Intelligence and Attainment Tests. Chicago, IL: MESA.
47. Reckase M.D. (2009) Multidimensional Item Response Theory. New York: Springer-Verlag.
48. Reise S.P. (2012) The Rediscovery of Bifactor Measurement Models // Multivariate Behavioral Research. Vol. 47. No 5. P. 667–696. doi:https://doi.org/10.1080/00273171.2012.715555.
49. Revelle W. (2020) Package «Psych» — Version: 1.9.12.31. https://cran.r-project.org/web/packages/psych/index.html
50. Rijmen F. (2010) Formal Relations and an Empirical Comparison among the Bi-Factor, the Testlet, and a Second-Order Multidimensional IRT Model // Journal of Educational Measurement. Vol. 47. No 3. P. 361–372. doi:https://doi.org/10.1111/j.1745–3984.2010.00118.x.
51. Robitzsch A., Kiefer T., Wu M. (2020) Package «TAM». Test Analysis Modules — Version: 3.5–19. https://cran.r-project.org/web/packages/TAM/index.html
52. Schmid J., Leiman J.M. (1957) The Development of Hierarchical Factor Solutions // Psychometrika. Vol. 22. No 1. P. 53–61. doi:https://doi.org/10.1007/BF02289209.
53. Schneider M., Beeres K., Coban L., Merz S., Schmidt S.S., Stricker J., de Smedt B. (2017) Associations of Non-Symbolic and Symbolic Numerical Magnitude Processing with Mathematical Competence: A Meta-Analysis // Developmental Science. Vol. 20. May. No e12372. doi:https://doi.org/10.1111/desc.12372.
54. Schwarz G. (1978) Estimating the Dimension of a Model // The Annals of Statistics. Vol. 6. No 2. P. 461–464. doi: http://dx.doi.org/10.1214/aos/1176344136.
55. Sinharay S., Puhan G., Haberman S.J. (2011) An NCME Instructional Module on Subscores // Educational Measurement: Issues and Practice. Vol. 30. No 3. P. 29–40. doi:https://doi.org/10.1111/j.1745–3992.2011.00208.x.
56. Smith E.V. (2002) Detecting and Evaluating the Impact of Multidimensionality Using Item Fit Statistics and Principal Component Analysis of Residuals // Journal of Applied Measurement. Vol. 3. No 2. P. 205–231.
57. Toll S.W.M., van Viersen S., Kroesbergen E.H., van Luit J.E.H. (2015) The Development of (Non-)Symbolic Comparison Skills throughout Kindergarten and their Relations with Basic Mathematical Skills // Learning and Individual Differences. Vol. 38. February. P. 10–17. doi:https://doi.org/10.1016/j.lindif.2014.12.006.
58. Wang W., Chen P., Cheng Y. (2004) Improving Measurement Precision of Test Batteries Using Multidimensional Item Response Models // Psychological Methods. Vol. 9. No 1. P. 116–136. doi: https://doi.org/10.1037/1082–989X.9.1.116.
59. Warm T.A. (1989) Weighted Likelihood Estimation of Ability in Item Response Theory // Psychometrika. Vol. 54. No 3. P. 427–450. doi:https://doi.org/10.1007/BF02294627.
60. Wilson M.R. (2005) Constructing Measures: An Item Response Modeling Approach. New Jersey: Routledge.

61. Wilson M., Gochyyev P. (2020) Having Your Cake and Eating It too: Multiple Dimensions and a Composite // Measurement. Vol. 151. November. No 107247. doi:https://doi.org/10.1016/j.measurement.2019.107247.
62. Wright B. D., Masters G. N. (1990) Computation of OUTFIT and INFIT Statistics // Rasch Measurement Transaction. Vol. 3. No 4. P. 84–85. https://www.rasch.org/rmt/rmt34e.htm
63. Wright B. D., Stone M. H. (1979) Best Test Design. Chicago, IL: MESA.
64. Wu M. (2010) Comparing the Similarities and Differences of PISA 2003 and TIMSS. OECD Education Working Papers No 32. doi:https://doi.org/10.1787/5km4psnm13nx-en.
65. Wu M., Tam H. P., Jen T. H. (2016) Multidimensional IRT Models in Book: Educational Measurement for Applied Researchers. Theory into Practice. Singapore: Springer.

References

Adams R. J. (2005) Reliability as a Measurement Design Effect. *Studies in Educational Evaluation*, vol. 31, no 2–3, pp. 162–172. doi:https://doi.org/10.1016/j.stueduc.2005.05.008.

Adams R. J., Khoo S. T. (1996) *Quest*. Melbourne, Australia: Australian Council for Educational Research.

Adams R. J., Wilson M., Wang W. C. (1997) The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement*, vol. 21, no 1, pp. 1–23. doi:https://doi.org/10.1177/0146621697211001.

Akaike H. (1974) A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, vol. 19, no 6, pp. 716–723. doi: https://doi.org/10.1109/TAC.1974.1100705.

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

Benoit L., Lehalle H., Molina M., Tijus C., Jouen F. (2013) Young Children's Mapping between Arrays, Number Words, and Digits. *Cognition*, vol. 129, no 1, pp. 95–101. doi:https://doi.org/10.1016/j.cognition.2013.06.005.

Bloom B. S. (ed.) (1956) *Taxonomy of Educational Objectives: The Classification of Educational Goals*. New York: Longmans, Green and Company.

Bock R. D., Aitkin M. (1981) Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm. *Psychometrika*, vol. 46, no 4, pp. 443–459. doi:https://doi.org/10.1007/BF02293801.

Bock R. D., Mislevy R. J. (1982) Adaptive EAP Estimation of Ability in a Microcomputer Environment. *Applied Psychological Measurement*, vol. 6, no 4, pp. 431–444. doi:https://doi.org/10.1177/014662168200600405.

Bonifay W., Lane S. P., Reise S. P. (2017) Three Concerns with Applying a Bifactor Model as a Structure of Psychopathology. *Clinical Psychological Science*, vol. 5, no 1, pp. 184–186. doi:https://doi.org/10.1177/2167702616657069.

Brandt S., Duckor B., Wilson M. (2014) *A Utility Based Validation Study for the Dimensionality of the Performance Assessment for California Teachers (PACT)*. Paper presented at Annual Conference of the American Educational Research Association. Available at: https://www.researchgate.net/publication/281645866_A_Utility_Based_Validation_Study_for_the_Dimensionality_of_the_Performance_Assessment_for_California_Teachers_PACT (accessed 2 April 2021).

Casey D. P. (1978) Failing Students: A Strategy of Error Analysis. Aspects of Motivation (ed. P. Costello), Melbourne: Mathematical Association of Victoria, pp. 295–306.

Chen Q., Li J. (2014) Association between Individual Differences in Non-Symbolic Number Acuity and Math Performance: A Meta-Analysis. *Acta Psychologica*, vol. 148, May, pp. 163–172. doi:https://doi.org/10.1016/j.actpsy.2014.01.016.

Clements M. A.K. (1980) Analyzing Children's Errors on Written Mathematical Tasks. *Educational Studies in Mathematics*, vol. 11, no 1, pp. 1–21. doi:https://doi.org/10.2307/3482042.

Clements M. A., Ellerton N. (1996) *The Newman Procedure for Analysing Errors on Written Mathematical Tasks*. Available at: https://compasstech.com.au/ARNOLD/PAGES/newman.htm (accessed 2 April 2021).

Cummins D. D., Kintsch W., Reusser K., Weimer R. (1988) The Role of Understanding in Solving Word Problems. *Cognitive Psychology*, vol. 20, no 4, pp. 405–438. doi:https://doi.org/10.1016/0010-0285(88)90011-4.

Davey T., Hirsch T. M. (1991) *Concurrent and Consecutive Estimates of Examinee Ability Profiles*. Paper presented at the Annual Meeting of the Psychometric Society (New Brunswick, NJ).

De Boeck P., Wilson M. (2004) *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York, NY: Springer Science & Business Media.

Foy P., Yin L. (2015) Scaling the TIMSS2015 Achievement Data. *Methods and Procedures in TIMSS2015* (eds M. O. Martin, I.V.S. Mullis, M. Hooper), Chestnut Hill, MA: Boston College, pp. 13.1–13.62. Available at: https://timssandpirls.bc.edu/publications/timss/2015-methods/chapter-13.html (accessed 2 April 2021).

Froumin I. D., Dobryakova M. S., Barannikov K. A., Remorenko I. M. (2018) *Universalnye kompetentnosti i novaya gramotnost: chemu uchit segodnya dlya uspekha zavtra. Predvaritelnye vyvody mezhdunarodnogo doklada o tendentsiyakh transformatsii shkolnogo obrazovaniya* [Key Competences and New Literacy: From Slogans to School Reality. Preliminary Results of the International Report of Major Trends in the On-Going Transformation of School Education. A Summary for Discussion]. Moscow: HSE. Available at: https://ioe.hse.ru/data/2018/07/12/1151646087/2_19.pdf (accessed 2 April 2021).

Gelman R., Butterworth B. (2005) Number and Language: How Are They Related? *Trends in Cognitive Sciences*, vol. 9, no 1, pp. 6–10. doi:https://doi.org/10.1016/j.tics.2004.11.004.

Gignac G. E. (2008) Higher-Order Models versus Direct Hierarchical Models: g as Superordinate or Breadth Factor? *Psychology Science Quarterly*, vol. 50, no 1, pp. 21–43. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.531.4178&rep=rep1&type=pdf (accessed 2 April 2021).

Grimm K. J. (2008) Longitudinal Associations Between Reading and Mathematics Achievement. *Developmental Neuropsychology*, vol. 33, no 3, pp. 410–426. doi:https://doi.org/10.1080/87565640801982486.

Haberman S. J. (2005) *When Can Subscores Have Value? ETS RR-05–08*. Princeton, NJ: ETS. doi:https://doi.org/10.1002/j.2333–8504.2005.tb01985.x.

Haberman S. J., Sinharay S. (2010) Reporting of Subscores Using Multidimensional Item Response Theory. *Psychometrika*, vol. 75, no 2, pp. 209–227. doi:https://doi.org/10.1007/s11336-010-9158-4.

Harris K. M., Martin M. O. (eds) (2013) *TIMSS2015 Assessment Framework*. Available at: https://timssandpirls.bc.edu/timss2015/downloads/T15_Frameworks_Full_Book.pdf (accessed 2 April 2021).

Jablonka E. (2003) Mathematical Literacy. *Second International Handbook of Mathematics Education* (eds A. J. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick, F. K. S. Leung), Dordrecht, Netherlands: Kluwer Academic, pp. 75–102. doi:https://doi.org/10.5951/mtlt.2019.0397.

Jordan N. C., Hanich L. B., Kaplan D. (2003) A Longitudinal Study of Mathematical Competencies in Children with Specific Mathematics Difficulties versus Children with Comorbid Mathematics and Reading Difficulties. *Child Development*, vol. 74, no 3, pp. 834–850. doi:https://doi.org/10.1111/1467-8624.00571.

Joyner R. E., Wagner R. K. (2020) Co-Occurrence of Reading Disabilities and Math Disabilities: A Meta-Analysis. *Scientific Studies of Reading: The Official Journal of the Society for the Scientific Study of Reading*, vol. 24, no 1, pp. 14–22. doi:https://doi.org/10.1080/10888438.2019.1593420.

Kamata A. (2001) Item Analysis by the Hierarchical Generalized Linear Model. *Journal of Educational Measurement*, vol. 38, no 1, pp. 79–93. doi:https://doi.org/10.1111/j.1745-3984.2001.tb01117.x.

Kanonire T., Federiakin D. A., Uglanova I. L. (2020) Multicomponent Framework for Students' Subjective Well-Being in Elementary School. *School Psychology Quarterly*, vol. 35, no 5, pp. 321–331. doi:https://doi.org/10.1037/spq0000397.

Kolkman M. E., Kroesbergen E. H., Leseman P. P.M. (2013) Early Numerical Development and the Role of Non-Symbolic and Symbolic Skills. *Learning and Instruction*, vol. 25, June, pp. 95–103. doi:https://doi.org/10.1016/j.learninstruc.2012.12.001.

Kozlov V. V., Kondakov A. M. (2009) *Fundamentalnoe yadro soderzhaniya obshchego obrazovaniya* [The Fundamental Core of the Content of General Education]. Moscow: Prosveshchenie. Available at: https://kpfu.ru/docs/F1999935214/fundamentalnoe.yadro.pdf (accessed 2 April 2021).

Libertus M. E., Odic D., Feigenson L., Halberda J. (2016) The Precision of Mapping between Number Words and the Approximate Number System Predicts Children's Formal Math Abilities. *Journal of Experimental Child Psychology*, vol. 150, October, pp. 207–226. doi:https://doi.org/10.1016/j.jecp.2016.06.003.

Linacre J. M. (1998) Structure in Rasch Residuals: Why Principal Components Analysis. *Rasch Measurement Transactions*, vol. 12, no 2, pp. 636. Available at: https://www.rasch.org/rmt/rmt122m.htm (accessed 2 April 2021).

Linacre J. M. (2002) What Do Infit and Outfit, Mean-Square and Standardized Mean. *Rasch Measurement Transactions*, vol. 16, no 2, pp. 878. Available at: https://www.rasch.org/rmt/rmt162f.htm (accessed 2 April 2021).

Linacre J. M. (2021) A User's Guide to Winsteps and Ministep: Rasch-Model Computer Programs. Program Manual 4.8.0. Available at: https://www.winsteps.com/manuals.htm (accessed 2 April 2021).

Linden W. J. van der (ed.) (2018) *Handbook of Item Response Theory*. Boca Raton, FL: CRC.

Mansolf M., Reise S. P. (2017) When and Why the Second-Order and Bifactor Models Are Distinguishable. *Intelligence*, vol. 61, February, pp. 120–129. doi:https://doi.org/10.1016/j.intell.2017.01.012.

Newman M. A. (1977) An Analysis of Sixth-Grade Pupils' Errors on Written Mathematical Tasks. *Research in Mathematics Education in Australia* (eds M. A. Clements, J. Foyster), Melbourne: Swinburne College, vol. 1, pp. 239–258.

OECD (2013) *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. Paris: OECD. doi:https://doi.org/10.1787/9789264190511-en.

OECD (2019) *PISA 2018 Assessment and Analytical Framework*. Paris: OECD. doi:https://doi.org/10.1017/CBO9781107415324.004.

Paek I., Yon H., Wilson M., Kang T. (2009) Random Parameter Structure and the Testlet Model: Extension of the Rasch Testlet Model. *Journal of Applied Measurement*, vol. 10, no 4, pp. 394–407. Available at: https://bearcenter.berkeley.edu/sites/default/files/Wilson%20%2327.pdf (accessed 2 April 2021).

Peng P., Namkung J., Barnes M., Sun C. (2016) A Meta-Analysis of Mathematics and Working Memory: Moderating Effects of Working Memory Domain, Type of Mathematics Skill, and Sample Characteristics. *Journal of Educational Psychology*, vol. 108, no 4, pp. 455–473. doi:https://doi.org/10.1037/edu0000079.

Purpura D. J., Baroody A. J., Lonigan C. J. (2013) The Transition from Informal to Formal Mathematical Knowledge: Mediation by Numeral Knowledge. *Journal of Educational Psychology*, vol. 105, no 2, pp. 453–464. doi:https://doi.org/10.1037/a0031753.

Rasch G. (1993) *Probabilistic Models for Some Intelligence and Attainment Te*sts. Chicago, IL: MESA.

Reckase M. D. (2009) *Multidimensional Item Response Theory*. New York: Springer-Verlag.

Reise S. P. (2012) The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research*, vol. 47, no 5, pp. 667–696. doi:https://doi.org/10.1080/00273171.2012.715555.

Revelle W. (2020) *Package "Psych"—Version: 1.9.12.31.* Available at: https://cran.r-project.org/web/packages/psych/index.html (accessed 2 April 2021).

Rijmen F. (2010) Formal Relations and an Empirical Comparison among the Bi-Factor, the Testlet, and a Second-Order Multidimensional IRT Model. *Journal of Educational Measurement*, vol. 47, no 3, pp. 361–372. doi:https://doi.org/10.1111/j.1745–3984.2010.00118.x.

Robitzsch A., Kiefer T., Wu M. (2020) *Package "TAM". Test Analysis Modules — Version: 3.5–19.* Available at: https://cran.r-project.org/web/packages/TAM/index.html (accessed 2 April 2021).

Schmid J., Leiman J. M. (1957) The Development of Hierarchical Factor Solutions. *Psychometrika*, vol. 22, no 1, pp. 53–61. doi:https://doi.org/10.1007/BF02289209.

Schneider M., Beeres K., Coban L., Merz S., Schmidt S. S., Stricker J., de Smedt B. (2017) Associations of Non-Symbolic and Symbolic Numerical Magnitude Processing with Mathematical Competence: A Meta-Analysis. *Developmental Science*, vol. 20, May, no e12372. doi:https://doi.org/10.1111/desc.12372.

Schwarz G. (1978) Estimating the Dimension of a Model. *The Annals of Statistics*, vol. 6, no 2, pp. 461–464. doi:http://dx.doi.org/10.1214/aos/1176344136.

Sinharay S., Puhan G., Haberman S. J. (2011) An NCME Instructional Module on Subscores. *Educational Measurement: Issues and Practice*, vol. 30, no 3, pp. 29–40. doi:https://doi.org/10.1111/j.1745–3992.2011.00208.x.

Smith E. V. (2002) Detecting and Evaluating the Impact of Multidimensionality Using Item Fit Statistics and Principal Component Analysis of Residuals. *Journal of Applied Measurement*, vol. 3, no 2, pp. 205–231.

Toll S. W.M., van Viersen S., Kroesbergen E. H., van Luit J. E.H. (2015) The Development of (Non-)Symbolic Comparison Skills throughout Kindergarten and Their Relations with Basic Mathematical Skills. *Learning and Individual Differences*, vol. 38, February, pp. 10–17. doi:https://doi.org/10.1016/j.lindif.2014.12.006.

Wang W., Chen P., Cheng Y. (2004) Improving Measurement Precision of Test Batteries Using Multidimensional Item Response Models. *Psychological Methods*, vol. 9, no 1, pp. 116–136. doi:https://doi.org/10.1037/1082–989X.9.1.116.

Warm T. A. (1989) Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika*, vol. 54, no 3, pp. 427–450. doi:https://doi.org/10.1007/BF02294627.

Wilson M. R. (2005) *Constructing Measures: An Item Response Modeling Approach*. New Jersey: Routledge.

Wilson M., Gochyyev P. (2020) Having Your Cake and Eating It Too: Multiple Dimensions and a Composite. *Measurement*, vol. 151, November, no 107247. doi:https://doi.org/10.1016/j.measurement.2019.107247.

Wright B. D., Masters G. N. (1990) Computation of OUTFIT and INFIT Statistics. *Rasch Measurement Transaction*, vol. 3, no 4, pp. 84–85. Available at: https://www.rasch.org/rmt/rmt34e.htm (accessed 2 April 2021).

Wright B. D., Stone M. H. (1979) *Best Test Design*. Chicago, IL: MESA.

Wu M. (2010) *Comparing the Similarities and Differences of PISA 2003 and TIMSS. OECD Education Working Papers no 32*. doi:https://doi.org/10.1787/5km4psnm13nx-en.

Wu M., Tam H. P., Jen T. H. (2016) *Multidimensional IRT Models in Book: Educational Measurement for Applied Researchers. Theory into Practice*. Singapore: Springer.