Декомпозиция трудности заданий в тесте читательской грамотности

Алина Иванова, Инна Антипкина

Статья поступила в редакцию в марте 2023 г.

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

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

Аннотация

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

Ключевые слова

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

Для цитирования

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

Decomposing Difficulty of Reading Literacy Test Items

Alina Ivanova, Inna Antipkina

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

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

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

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

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

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

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

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

1. Читательские умения: простые и сложные

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

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

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

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

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

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

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

2. Методология

2.1. Инструмент исследования

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

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

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

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

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

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

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

2.2. Выборка

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

2.3. Метод анализа

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

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

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

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

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

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

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

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

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

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

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

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

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

2.4. *Q*-матрица

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

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

3. Результаты анализа

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Таблица 3. **Критерии сравнения моделей**

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

Нулевой моделью в нашем анализе является модель Раша.

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

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

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

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

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

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

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

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

4. Обсуждение результатов

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

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

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

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

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

5. Заключение

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

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

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

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

Благодарности

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

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

Литература

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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