Измерение образовательного прогресса на основе когнитивных операций

Сергей Тарасов, Ирина Зуева, Денис Федерякин

Статья поступила в редакцию в марте 2023 г.

Тарасов Сергей Владимирович — аспирант, стажер-исследователь Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики». Адрес: 101000, Москва, Потаповский пер., 16, стр. 10. E-mail: svtarasov@hse.ru. ORCID: https://orcid.org/0000-0003-4151-115X (контактное лицо для переписки)

Зуева Ирина Олеговна — аспирант, аналитик Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики». E-mail: izueva@hse.ru

Федерякин Денис Александрович — научный сотрудник департамента экономического образования, Университет им. Иоганна Гутенберга (Майнц, Германия). E-mail: denis.federiakin@uni-mainz.de

Аннотация

Измерение образовательного прогресса остается нетривиальной методологической задачей даже при наличии множества описанных в литературе подходов к его концептуализации и моделированию. Рассматривается методологический подход к измерению образовательного прогресса в рамках современной теории тестирования, при этом традиционная концептуализация этого подхода расширяется за счет моделирования когнитивных операций. Показано, что синтез традиционных моделей для измерения образовательного прогресса с одной из самых популярных моделей современной теории тестирования — LLTM, позволяющей моделировать когнитивные операции, — существенно обогащает возможности интерпретации тестовых баллов учеников, сохраняя все достоинства традиционного подхода к измерению образовательного прогресса. Для иллюстрации предлагаемого подхода использована линейка тестов, применявшихся для мониторинга образовательного прогресса в математике в 8–9-х классах средней школы.

Ключевые слова

измерение образовательного прогресса, IRT, LLTM, когнитивные операции, диагностические критериально-ориентированные пороги

Для цитирования

Тарасов С.В., Зуева И.О., Федерякин Д.А. (2023) Измерение образовательного прогресса на основе когнитивных операций. *Вопросы образования / Educational Studies Moscow*, № 3, сс. 172–196. https://doi.org/10.17323/vo-2023-16902

Measuring Learning Progress Based on Cognitive Operations

Sergei Tarasov, Irina Zueva, Denis Federiakin

Sergei V. Tarasov — PhD Student, Research Assistant at the Center for Psychometrics and Measurement in Education, National Research University Higher School of Economics. Address: Bld. 10, 16 Potapovsky Ln, 101000 Moscow, Russian Federation. E-mail: svtarasov@hse.ru. ORCID: https://orcid.org/0000-0003-4151-115X (corresponding author)

Irina O. Zueva — PhD Student, Analyst at the Center for Psychometrics and Measurement in Education, Institute of Education, National Research University Higher School of Economics. E-mail: izueva@hse.ru

Denis A. Federiakin — Research Assistant at the Department of Economic Education, Johannes Gutenberg University (Mainz). E-mail: denis.federiakin@uni-mainz.de

Abstract Measuring students' growth and change is considered one of the main ways for evidence-based development of educational systems. However, it is a non-trivial methodological task, despite the numerous approaches available for its conceptualization and statistical realization. In this article, we describe the main features of measuring students' growth and change using Item Response Theory (IRT) in detail. We then expand this approach to allow for the modeling of cognitive operations with the Linear Logistic Test Model (LLTM). We show that the synthesis of traditional IRT models for measuring growth and change with LLTM significantly enriches the interpretability of ability estimates while preserving the advantages of the traditional approach. To illustrate this approach, we use a set of monitoring tests to measure educational progress in mathematics in secondary school.

Keywords measurement of progress, IRT, LLTM, vertical alignment, cognitive operations, diagnostic thresholds

For citing Tarasov S.V., Zueva I.O., Federiakin D.A. (2023) Izmerenie obrazovatel´nogo progress ana osnove kognitivnykh operatsiy [Measuring Learning Progress Based on Cognitive Operations]. *Voprosy obrazovaniya / Educational Studies Moscow*, no 3, pp. 172–196. https://doi.org/10.17323/vo-2023-16902

Измерение образовательного прогресса — один из основных способов доказательного развития образования. Образовательный прогресс дает основания судить о том, какие образовательные практики и политики положительно ассоциированы с повышением уровня образовательных достижений у учащихся [Slavin, 2005]. Однако создание условий, необходимых для того, чтобы доказательно говорить об измерении образовательного прогресса, представляет собой нетривиальную методологическую задачу. Методологи предложили немало методов измерения прогресса [Саляхутдинова, Федерякин, 2022]. При всем их разнообразии в образовании доминирующим подходом остается современная теория тестирования благодаря ее интуитивной понятности и относительной технической простоте.

В отличие от классической теории тестирования, работающей с суммой первичных баллов за задания, современная теория тестирования (*Item Response Theory*, IRT) [Linden van der, 2018] подразумевает математическое разделение параметров заданий и респондентов. Из-за того что IRT обращается с этими параметрами как с независимыми — или, по крайней мере, отдельными, — этот подход к образовательным измерениям открывает много возможностей, которые нельзя реализовать в классическом подходе. IRT не только включает в математическую модель для оценки уровня способностей респондентов множество дополнительных факторов и аспектов оценки, но и позволяет доказательно говорить об измерении образовательного прогресса. С использованием IRT можно моделировать когнитивные структуры, задействованные в процессе решения заданий, например с помощью оценивания трудностей различных когнитивных операций [Fischer, 1973], — а значит, выяснять, какие когнитивные операции респондент освоил, а какие нет.

Цель данной работы состоит в том, чтобы проиллюстрировать новый подход к измерению образовательного прогресса с помощью когнитивных операций, задействованных в процессе решения заданий теста, в рамках парадигмы IRT. В статье мы прежде всего обосновываем невозможность измерения образовательного прогресса без использования специальных психометрических методов. Далее мы описываем один из подходов к измерению образовательного прогресса в IRT и обсуждаем его соотношение с другими способами измерения прогресса. Затем мы рассматриваем возможности, которые предоставляет IRT для анализа когнитивных операций, заложенных в тест, и синтезируем эти две области IRT, чтобы описать методологию измерения образовательного прогресса на основе когнитивных операций. И наконец, мы приводим пример реализации предложенного подхода на данных мониторинга индивидуального образовательного прогресса в математике учащихся 8–9-х классов и вписываем результаты этого исследования в более широкий исследовательский контекст.

1. Проблемы с измерением образовательного прогресса в классических подходах

Чтобы измерение образовательного прогресса было возможным, необходимы следующие условия:

- наличие нескольких (минимум двух) замеров одной и той же способности на определенной временной дистанции;
- наличие якорных заданий в разных тестах;
- применение специальных методов психометрического моделирования.

Первое условие очевидно. Выполнение второго и третьего условий — именно то, за что подвергаются критике попытки измерить образовательный прогресс в классической теории тестирования.

Наличие якорных заданий дает возможность проверить, действительно ли разные тесты все еще измеряют одну и ту же способность [Waterbury, DeMars, 2021]. Якорными называются абсолютно одинаковые задания, включенные в разные тесты (замеры). Если образовательный прогресс измеряется на длительной дистанции и проводится больше двух замеров, то якорные задания в разных парах замеров могут быть разными, т.е. между первым и вторым замером будут одни якорные задания, а между вторым и третьим — другие; причем между первым и третьим замером может не быть якорных заданий. Дело в том, что содержание образования меняется, поэтому темы из первого замера могут не пересекаться с третьим замером, в то время как между соседними замерами всегда должна находиться область пересечения. Таким образом, постепенное изменение содержания тестов позволяет одновременно сохранить непрерывность шкалы способности и обеспечить гибкость ее интерпретации.

Когда претест и посттест являются абсолютно одинаковыми [Dimitrov, Rumrill, 2003], все задания могут рассматриваться как якорные, что гарантирует полную сопоставимость тестовых баллов даже без применения психометрического моделирования. Однако если образовательный прогресс измеряется на длительной дистанции, при предъявлении ученикам одного и того же множества заданий, например, в начале 1-го и в конце 4-го класса возникнут потолочные эффекты и достоверное измерение образовательного прогресса будет подавлено статистическими артефактами. Таким образом, одинаковые претест и посттест можно применять только для оценки образовательного прогресса на краткосрочной дистанции, например в течение учебной четверти, и только с помощью двух замеров подряд, потому что при большем количестве замеров становится невозможно игнорировать эффект знакомства респондентов с заданиями. На практике такое измерение образовательного прогресса встречается при проведении краткосрочных экспериментальных интервенций, когда требуется сравнить образовательные достижения учащихся, чтобы оценить эффекты новых образовательных практик и политик.

Когда якорных заданий в следующих один за другим замерах нет, происходит «разрыв» шкалы. Если претест и посттест состоят из непересекающихся множеств заданий, то с точки зрения психометрики такая ситуация аналогична использованию в претесте и посттесте тестов по разным предметам. Даже

если претест и посттест разработаны в одной и той же теоретической рамке и спецификации, результаты повторного тестирования совершенно не поддаются интерпретации, так как одни и те же баллы, полученные в двух тестах, не говорят об одинаковой способности ученика. Только наличие якорных заданий обеспечивает сопоставимость оценок способности, полученных в разных замерах. В следующем разделе статьи будет рассмотрен еще один аргумент в пользу наличия якорных заданий — математический.

Для измерения образовательного прогресса недостаточно просто включить якорные задания в соседние замеры. Чтобы сравнивать тестовые баллы из соседних замеров, необходимо обеспечить единство интерпретации этих тестовых баллов, т.е. добиться, чтобы один и тот же тестовый балл соответствовал одной и той же истинной способности респондента [Loyd, Hoover, 1980]. Единства в оценивании результатов двух разных замеров невозможно достичь без применения психометрического моделирования в IRT, так как в классической теории тестирования наблюдаемый балл респондента может интерпретироваться только внутри одного теста относительно конкретного набора заданий. Рассмотрим случай, когда из 30 заданий в претесте и 30 заданий в посттесте 10 являются якорными, а 20 заданий — уникальными для каждого замера. Если респондент набрал условные 15 баллов в претесте и 10 в посттесте, это совсем не обязательно означает отрицательный образовательный прогресс. Возможно, дело в том, что уникальные задания в посттесте были труднее, чем в претесте, — именно так обычно и конструируют задания при разработке инструментов для измерения образовательного прогресса. В классическом подходе сравнение тестовых баллов, полученных на хоть сколько-нибудь отличающихся друг от друга множествах заданий, невозможно. То есть одно и то же количество первичных баллов в претесте и посттесте не соответствует одному и тому же уровню способности. Таким образом, единства интерпретации шкалы невозможно достичь без психометрического моделирования.

2. Применение современной теории тестирования для измерения образовательного прогресса

Одна из самых сложных проблем в измерении образовательного прогресса — его концептуализация. Достаточно понятна процедура измерения прогресса, например, в физкультуре: высота прыжка или количество попаданий баскетбольным мячом в корзину представляют собой элементарные наблюдаемые метрики, интерпретация которых совершенно одинакова как для учеников начальной школы, так и для взрослых людей. Однако концептуализировать когнитивный прогресс чрезвычайно трудно, потому что это латентный процесс. Его обычно не

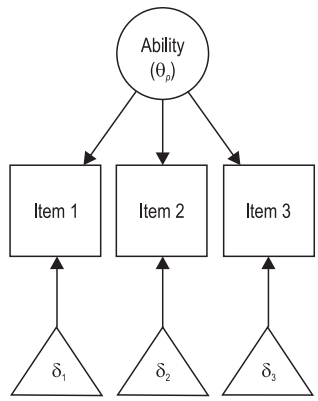
удается наблюдать в поведении учеников, а если и удается, то его количественное измерение составляет большую проблему. В частности, для преодоления этих трудностей и разрабатывался математический аппарат IRT [Sontag, 1984].

Ключевой постулат IRT — математическое разделение параметров респондента и задания. Пример — формула дихотомической модели Раша:

$$P(U_{pi} = 1) = \frac{\exp(\theta_p - \delta_i)}{1 + \exp(\theta_p - \delta_i)}, \quad (1)$$

где U_{pi} — ответ респондента p ϵ (1, 2, ..., P) на задание i ϵ (1, 2, ..., I); $P(U_{pi} = 1)$ — вероятность того, что этот ответ будет равен единице (вероятность правильного ответа); θ_p — оценка латентной способности респондента p на логит-шкале (шкале логарифмических шансов); δ_i — оценка латентной трудности задания i на логит-шкале. Эту модель можно представить с помощью путевой диаграммы (рис. 1).

Рис. 1. **Путевая диаграмма модели Раша**



Примечание: *Ability* — способность, *Item* — задание, кругом обозначена оцениваемая способность респондентов (латентная переменная), квадратами — ответы на задания теста (наблюдаемые переменные), треугольниками — параметры заданий (трудности). Стрелками с одним направлением обозначены регрессионные зависимости («эффект применяется к...»).

В моделях IRT способности респондентов и трудности заданий располагаются и сравниваются на одной метрической шкале. Параметры заданий и параметры респондентов оцениваются отдельно во всех моделях IRT, но только в моделях Раша они сравниваются напрямую без дополнительных модификаций, что позволяет достичь полного разделения этих параметров. Модель из уравнения (1) допускает, что вероятность ответа респондента на задание зависит от того, как соотносятся

уровень способности респондента и уровень трудности здания. Если уровень трудности задания выше, чем уровень способности респондента, вероятность правильного ответа будет меньше 50%. Если уровень способности респондента превышает трудность задания, вероятность правильного ответа больше 50%. Такой способ формулирования процесса ответа на математическом языке обладает интересными математическими свойствами [Andersen, 1977], лежащими в основе современных подходов к измерениям в социальных науках. Однако для целей измерения образовательного прогресса мы заинтересованы только в одном из них: в разделении вариативности ответов «внутри» респондента и между респондентами. Параметр θ_p описывает различия между респондентами, ранжируя их от самых слабых до самых сильных, а параметр δ_i отражает различия «внутри» респондентов, ранжируя задания от самых простых до самых трудных. Различия заданий по параметру δ_i показывают, насколько для любого респондента одно задание труднее другого, — отсюда и интерпретация этого параметра именно «внутри» респондента. Если одно задание труднее другого, оно будет труднее для всех респондентов, т.е. вариация «внутри» респондентов одинаковая (δ_i применяется ко всем респондентам), а все различия между ними вынесены в другой параметр модели — θ_p. И тогда вероятность наблюдать правильный ответ в разных заданиях служит прокси для измерения параметра интереса — θ_p.

Одну из первых моделей для измерения лонгитюдных изменений предложил Э. Андерсен [Andersen, 1985]. Математически она представляет собой многомерное расширение модели Раша на случай с якорными заданиями между замерами:

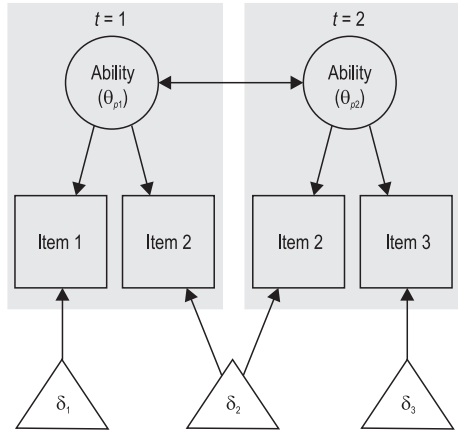
$$P(U_{pit}=1)=\frac{\exp(\theta_{pt}-\delta_i)}{1+\exp(\theta_{pt}-\delta_i)}, \quad (2)$$

где U_{pit} — ответ респондента p на задание i в момент замера $t \in (1, 2, ..., T)$; θ_{pt} — оценка латентной способности респондента p в момент замера t на логит-шкале; δ_i — оценка латентной трудности задания i на логит-шкале.

Путевая диаграмма этой модели представлена на рис. 2.

В отличие от модели Раша, модель Андерсена является многомерной: она содержит столько размерностей, по которым различаются респонденты, сколько было проведено замеров. По каждой из размерностей одни и те же респонденты ранжированы по-разному. При этом от простых многомерных моделей модель Андерсена отличается фиксацией параметров повторяющихся заданий, нагружающих разные размерности респондентов (предъявленных в разные замеры), к одним и тем же значениям. Таким образом, якорные задания имеют

Рис. 2. **Путевая диаграмма модели Андерсена**



Примечание (в дополнение к примечаниям к рис. 1): Двусторонней стрелкой обозначена корреляция, закрашенными областями обозначены разные замеры.

один и тот же уровень трудности во всех замерах — и это условие составляет основу для концептуализации образовательного прогресса. В современной теории тестирования образовательный прогресс понимается как изменение вероятности решения якорных заданий от одного замера к другому. Тот факт, что параметры заданий отделены от параметров респондентов в уравнении модели, позволяет использовать θ_p в простой модели Раша как условную агрегацию вероятности решения заданий на уровень респондента. Изменения этой вероятности, переведенные на логит-шкалу, показывают, как меняется способность респондента. Использование якорных заданий для этой модели — критически необходимое условие сопоставимости шкал соседних замеров: трудность якорных заданий оценивается относительно распределения выборки в первом замере (и не меняется), а во втором (и/или в последующих замерах) уже относительно этих заданий оценивается положение выборки на той же шкале, что и в первом замере. Таким способом обеспечивается соответствие одного и того же значения параметров θ_{pt} для всех замеров одному и тому же уровню способности, что позволяет доказательно говорить об образовательном прогрессе.

Параллельно с отслеживанием образовательного прогресса эта модель позволяет анализировать психометрические характеристики как отдельных замеров, так и всего множества заданий, чтобы убедиться, что измерительный инструментарий соответствует необходимым стандартам качества, а также дает возможность делать индивидуальные и/или групповые выводы о респондентах.

3. Другие возможности измерения образовательного прогресса в современной теории тестирования

Один из наиболее популярных способов измерения прогресса — вертикальное выравнивание [Sontag, 1984; Саляхутдинова, Федерякин, 2022]. В отличие от модели Андерсена, вертикальное выравнивание допускает, что, несмотря на повторяющиеся замеры, множество заданий, используемых для измерения прогресса, является одномерным. То есть здесь отдельные замеры не рассматриваются как отдельные размерности. Такой подход существенно упрощает психометрическое моделирование, используемое для вертикального выравнивания, но предъявляет гораздо более высокие требования к качеству измерительного инструментария: он должен быть нечувствителен к тому, что ответы одного и того же респондента на якорные задания скоррелированы в разные моменты времени, т.е. высока вероятность того, что в каждый следующий замер респондент ответит на якорный вопрос так же, как ответил, столкнувшись с этим вопросом первый раз (или похожим образом). Модель Андерсена учитывает эту вероятность, позволяя оценкам способности из разных замеров коррелировать. При этом вертикальное выравнивание, как метод пост-хок размещения оценок респондентов на одной шкале, может не предполагать установления пороговых баллов. Соответственно этот метод может не предусматривать обогащения интерпретации баллов содержанием образовательного прогресса.

В IRT есть модели для измерения прогресса, в основе которых лежит внутренняя многомерность заданий, в этом случае в заданиях из последующих замеров проявляются также уровни способности из всех предыдущих замеров: например, модель С. Эмбретсон для роста и измерения (*model for measuring growth and change*) [Embretson, 1991]) или экспланаторная полиномиальная модель латентного роста (*explanatory polynomial model of latent growth*) [Wilson, Zheng, McGuire, 2012]). Они позволяют моделировать специфические изменения в латентной способности, в то время как модель Андерсена отражает просто положение респондента на единой шкале. Соответственно в модели Андерсена, чтобы вычислить именно прогресс, необходимо произвести дополнительные операции: вычесть из оценки способности респондента в последующий замер оценку способности в предыдущий. Однако внутренняя многомерность заданий делает модели IRT численно менее стабильными, что затрудняет их практическое применение в случаях, когда после каждого замера респондентам нужно давать обратную связь и сообщать оценки их способности. В частности, в моделях с внутренней многомерностью заданий добавление последующих замеров приводит к некоторому пересчету оценок способности во все предыдущие замеры. Эти изменения относительно неболь-

шие и находятся в пределах ошибки измерения (статистически незначимы), однако респондентам бывает трудно объяснить, почему их балл изменился после нового замера. Модель Андерсена же, как модель с многомерностью между заданиями, характеризуется большей численной стабильностью, что определило ее использование в данной работе.

Безусловно, существуют способы измерения образовательного прогресса, не относящиеся к IRT [Саляхутдинова, Федерякин, 2022], но здесь мы ограничиваемся рассмотрением только этой парадигмы. Все описанные подходы предполагают использование коллатеральной информации о респондентах. Под коллатеральной понимается информация, которая теоретически не обоснована, но которую удобно собирать в компьютерном формате и применять для повышения точности оценивания при психометрической обработке данных с сохранением оригинальной интерпретации параметров модели [Федерякин, Угланова, Скрябин, 2021].

4. Анализ когнитивных операций в современной теории тестирования

Одно из допущений моделей, созданных на основе модели Раша, состоит в том, что у каждого задания есть свой параметр — δ_i (см. уравнения 1–2). Поэтому весь анализ и интерпретация результатов тестирования проводятся на уровне заданий. Однако иногда исследователей интересует интерпретация результатов тестирования относительно не заданий, а стоящих за ними когнитивных операций. Когда теоретическая рамка теста содержит описание когнитивных операций, которые провоцируются заданиями, связь заданий с этими когнитивными операциями может быть отражена в когнитивной карте теста, которую называют *Q*-матрицей. *Q*-матрица представляет собой таблицу с заданиями теста в строках и когнитивными операциями в столбцах, а ячейки показывают, как соотносятся гипотетические когнитивные операции с заданиями: 0 — если когнитивная операция не задействована в задании, 1 — если задействована. Наивный пример когнитивной карты теста по арифметике из трех заданий приведен в табл. 1.

Таблица 1. **Пример когнитивной карты теста**

Задание	Когнитивные операции		
	Сложение	Умножение	Порядок действий
$2 + 3 \times 2$	1	1	1
$4 \times 2 + 1$	1	1	0
$2 + 3 + 5$	1	0	0

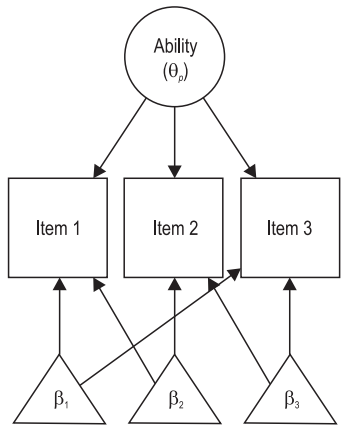
Специально для такого случая разработана (как частный случай модели Раша) одна из самых влиятельных моделей IRT — *Linear Logistic Test Model* (LLTM) [Fischer, 1973; 1995]:

$$P(U_{pi}=1)=\frac{\exp(\theta_{pt}-\sum_{k=1}^{K}\beta_k q_{ik})}{1+\exp(\theta_p-\sum_{k=1}^{K}\beta_k q_{ik})}, \quad (3)$$

где U_{pi} — ответ респондента p на задание i; θ_p — оценка латентной способности респондента p на логит-шкале; β_k — трудность когнитивной операции k ϵ (1, 2, ..., K) на логит-шкале; q_{ik} — элемент Q-матрицы из строки i и столбца k, обозначающий, необходима ли когнитивная операция k для решения задания i.

Путевая диаграмма модели LLTM показана на рис. 3.

Рис. 3. **Пример путевой диаграммы LLTM**



Таким образом, LLTM моделирует не индивидуальные трудности заданий, а трудности когнитивных операций, стоящих за этими заданиями. В LLTM часто добавляют не только когнитивные операции, но и другие атрибуты заданий, например формат или способ презентации информации в задании [Rolfes, Roth, Schnotz, 2018]. Но в данной работе мы концентрируемся только на изучении эффектов когнитивных операций.

В LLTM трудности заданий (δ_i из уравнения 1) допускаются равными сумме трудностей задействованных в задании когнитивных операций:

$$\delta_i=\sum_{k=1}^{K}\beta_k q_{ik}. \quad (4)$$

Ключевое отличие LLTM от модели Раша состоит в том, что в LLTM параметры, оцениваемые со стороны заданий, не уникальны, а «делятся» заданиями на основе общности когнитивных операций между ними. При этом, чтобы LLTM могла быть

оценена, когнитивных операций не должно быть больше, чем заданий. В отличие от обычной модели Раша, LLTM позволяет делать вывод о респондентах в терминах когнитивных операций, сравнивая оценки способности с уровнем трудности этих операций. По аналогии с оригинальной интерпретацией модели Раша, если уровень трудности когнитивной операции выше, чем уровень способности респондента, вероятность ее освоения этим респондентом меньше 50%, если уровень способности респондента превышает уровень трудности когнитивной операции, вероятность ее освоения больше 50%. Такой подход позволяет существенно обогатить интерпретацию тестовых баллов: вместо балла на логит-шкале, трансформированного на какую-то другую шкалу, результатом тестирования становится информация о том, что респондент уже умеет или чего еще не умеет [Lee, 2016]. Таким образом, применение LLTM служит основой для доказательного установления диагностических пороговых баллов на шкале способности: оцененный уровень трудности когнитивных операций становится диагностическим порогом, который можно напрямую сравнить со способностями респондентов.

LLTM допускает прогрессию (иерархию) в освоении когнитивных операций: более простая когнитивная операция будет более простой для всех респондентов, и нет никакой возможности двинуться дальше — к освоению следующей, более трудной когнитивной операции, пока предыдущая не освоена. Модели когнитивной диагностики, в том числе лонгитюдные [Deonovic et al., 2019], и анализ латентных переходов [Nsowaa, 2018] позволяют «расслабить» это допущение (или проверить его), однако они относятся к радикально другой статистической парадигме: они не ранжируют респондентов по способности, а сразу классифицируют их по нескольким основаниям — по степени освоения каждой из когнитивных операций, и их лонгитюдные расширения для измерения прогресса сложнее в применении.

Необходимым условием эффективного использования LLTM является качественная *Q*-матрица. Если она составлена неправильно, LLTM не только будет подходить данным хуже Раш-модели, в этом случае результаты моделирования с большой вероятностью будут содержать различные статистические артефакты [Baker, 1993; Macdonald, 2014]. На практике ситуация, когда LLTM подходит данным лучше, чем модель Раша, практически не встречается, но это не значит, что ее не стоит применять к данным. Теоретическая сила LLTM состоит в ее полезности для объяснения трудности когнитивных операций. При этом использование *Q*-матрицы для расчета LLTM адресует к использованию коллатеральной информации о заданиях [Федерякин, Угланова, Скрябин, 2021].

5. Методологические основы измерения образовательного прогресса на основе когнитивных операций

Для того чтобы проиллюстрировать измерение прогресса на основе когнитивных операций, мы подставляем уравнение (4) в уравнение (2), объединяя модель Андерсена и LLTM. Таким образом мы получаем модель LLTM — Андерсена (5) и противопоставляем ее в дальнейшей работе модели Раша — Андерсена (2):

$$P(U_{pit}=1)=\frac{\exp(\theta_{pt}-\sum_{k=1}^{K}\beta_k q_{ik})}{1+\exp(\theta_{pt}-\sum_{k=1}^{K}\beta_k q_{ik})}, \quad (5)$$

где U_{pit} — ответ респондента p на задание i в замер t; θ_{pt} — оценка латентной способности респондента p в замер t на логит-шкале; β_k — трудность когнитивной операции k на логит-шкале; q_{ik} — элемент Q-матрицы из строки i и столбца k.

Можно сказать, что модель LLTM — Андерсена выравнивает шкалы разных замеров не через трудности якорных заданий, а через трудности якорных когнитивных операций. Кроме выравнивания, модель LLTM — Андерсена обладает всеми достоинствами обычной LLTM, позволяя доказательно устанавливать диагностические пороговые баллы, но уже на вертикальной шкале сквозь все замеры. В этом отношении предлагаемая нами модель LLTM — Андерсена схожа с вертикально интерпретируемыми пороговыми баллами (*vertically moderated standard setting*, VMSS) [Cizek, 2013], однако она более состоятельна с математической точки зрения, поскольку не предполагает таких сильных теоретических допущений, как VMSS, где необходимым условием является равенство пороговых баллов, разделяющих респондентов на качественные группы и установленных внутри каждого теста по отдельности. В модели LLTM — Андерсена пороги на вертикальной шкале — это скорее следствие измерения прогресса, чем средство для него.

Таким образом, предлагаемый нами подход сочетает использование коллатеральной информации о заданиях (в виде Q-матрицы) и коллатеральной информации о респондентах (в виде замеров в разные моменты времени) для повышения надежности измерения при сохранении способа интерпретации способности из оригинальной Раш-модели. При этом возможно также привлечение коллатеральной информации о взаимодействии респондентов и заданий, например сведений о времени решения и/или о попытках решения.

6. Пример измерения образовательного прогресса на основе когнитивных операций

6.1. Инструмент

Для иллюстрации применения описанного подхода мы используем линейку тестов, разработанную в Центре психометрики и измерений в образовании Института образования НИУ «Высшая школа экономики» для мониторинга индивидуального образовательного прогресса учащихся средней школы. Мониторинг предполагает тестирование учащихся в начале и в конце учебного года. Первое тестирование проводится в конце сен-

тября — начале октября, когда учащиеся уже адаптировались к школе после летних каникул, но еще не начали активно изучать новый материал. Второе тестирование проводится в конце апреля — начале мая, когда основная учебная программа уже пройдена. В начале каждого учебного года используется тот же тест, что и в конце предыдущего учебного года.

В данном исследовании мы использовали данные, полученные в четырех волнах мониторинга по математике одной когорты учеников — от начала 8-го класса до конца 9-го класса. Так как тесты, предъявлявшиеся в конце 8-го класса и в начале 9-го класса, совпадают, базу данных составляют результаты выполнения школьниками трех уникальных тестов. Каждый тест состоял из 25 заданий с выбором одного варианта ответа из нескольких предложенных или коротким самостоятельно формулируемым ответом. Каждые два соседних теста содержали 4 якорных задания, а первый и третий тесты имели 2 якорных задания.

В заданиях представлены 12 разных когнитивных операций, причем в первом тесте их было 8, во втором — 11, а третий тест включает все 12. Примеры когнитивных операций, использованных в этой линейке тестов: навыки выполнения вычисления с рациональными числами, навык работы с координатной плоскостью. Весь набор когнитивных операций и их распределение между замерами представлены в табл. 2. Набор когнитивных операций для каждого задания определялся разработчиками теста и впоследствии был скорректирован экспертом в предметной области.

Таблица 2. **Распределение когнитивных операций между замерами**

Когнитивная операция	Замер 1	Замеры 2 и 3	Замер 4
Навыки выполнения вычислений с целыми числами	21	21	19
Навыки выполнения вычислений с рациональными числами	8	6	8
Навыки построения и применения математических моделей	7	6	8
Вспоминание фактов/формул и их последующее применение	11	18	19
Навык преобразования алгебраических выражений	10	7	6
Навыки решения уравнений	5	3	5
Навыки представления и считывания информации из текстового описания	2	5	3
Навыки выстраивания цепочки решения в несколько шагов	8	4	3
Навыки выполнения вычислений с вещественными числами	0	4	1
Навыки работы с координатной плоскостью	0	3	5
Навыки решения систем уравнений	0	2	1
Навыки решения неравенств и систем неравенств	0	0	4

6.2. Выборка и сбор данных

Выборка состоит из 472 учеников, которые принимали участие хотя бы в одном из четырех тестирований в ходе мониторинга. Из них 267 учеников принимали участие во всех четырех замерах, 121 — в трех (пропустили один замер), 53 — в двух (пропустили два замера) и 31 — только в одном. По замерам распределение равномерное: 394 ученика протестированы в начале 8-го класса, 383 — в конце 8-го класса, 388 — в начале 9-го класса и 395 — в конце 9-го класса; т.е. ни на одном из этапов с выборкой не происходило существенных изменений. С географической точки зрения выборка гомогенна, так как все ученики обучаются в школах одного и того же района одного из регионов России. Тестирования проводились в 2020–2022 гг.

Ученики проходили тестирование на персональных компьютерах, предоставленных образовательными учреждениями, на которых им требовалось авторизоваться, чтобы получить доступ к системе тестирования и тестовым материалам. Соблюдение процедуры тестирования, включая контроль списывания, а также безопасность тестовых материалов обеспечивали местные наблюдатели.

Для обработки результатов тестирования использовалась библиотека TAM v.3.7-16 для языка программирования R v.4.1.0.

6.3. Результаты

На общих данных по всем четырем замерам построены две модели: модель Раша — Андерсена, которая с учетом якорных заданий содержит 67 индивидуальных параметров трудности заданий, и модель LLTM — Андерсена, в которой 12 параметров трудности когнитивных операций. Для проверки согласия моделей с данными мы используем статистики глобального согласия: информационный критерий Акаике (*Akaike Information Criterion*, AIC) [Akaike, 1974] и информационный критерий Шварца (*Bayesian Information Criterion*, BIC) [Gideon, 1978]. Эти индексы описывают общее согласие моделей с данными, вводя пенальти за дополнительные параметры (AIC) с учетом размера выборки (BIC). Судя по статистикам согласия, модель Раша — Андерсена подходит нашим данным лучше модели LLTM — Андерсена (табл. 3) — возможно, за счет того, что она точнее описывает трудности заданий из-за большего количества оцениваемых параметров. EAP-надежность для всех замеров [Adams, 2005] незначительно выше в модели Раша — Андерсена.

В обеих моделях средняя способность в первом замере была зафиксирована равной нулю, а в остальных замерах она является оцениваемым параметром, и при этом отслеживается ее изменение по сравнению с первым замером. Во втором замере наблюдается увеличение средней способности, в третьем — ее спад. Такая динамика представляет собой относитель-

Таблица 3. **Сравнение моделей Раша — Андерсена и LLTM — Андерсена**

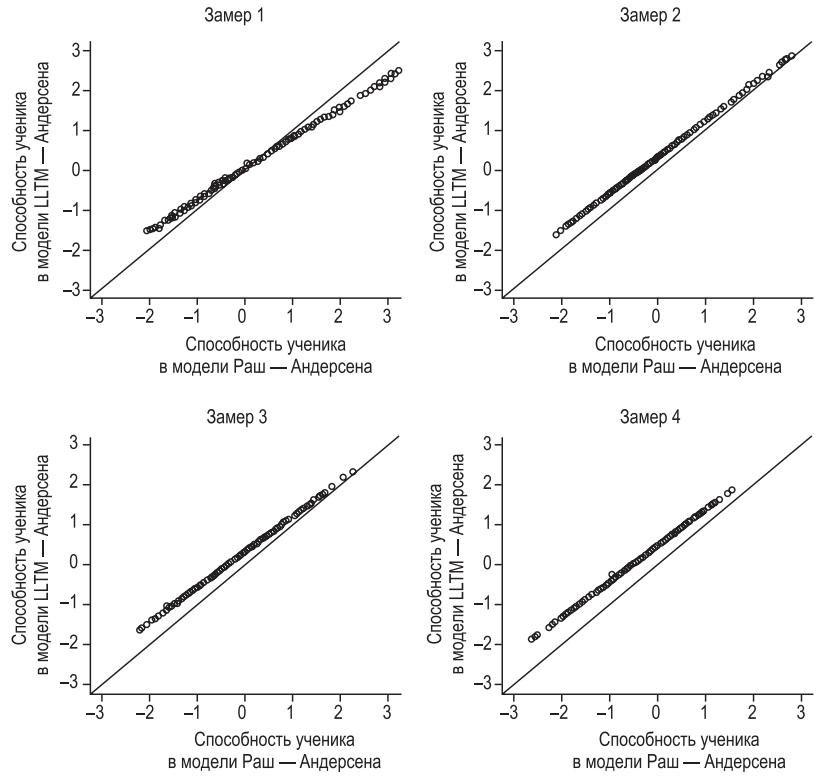
Статистика		Раш — Андерсен	LLTM — Андерсен
Глобальное согласие	AIC	41 828	45 748
	BIC	42 160	45 851
Надежность	Замер 1	0,787	0,767
	Замер 2	0,779	0,772
	Замер 3	0,767	0,760
	Замер 4	0,688	0,673
Средняя способность учеников	Замер 1	0	0
	Замер 2	0,060	0,218
	Замер 3	–0,221	–0,031
	Замер 4	–0,272	0,074
Средняя способность 267 учеников*	Замер 1	0,054	0,076
	Замер 2	–0,026	0,305
	Замер 3	–0,308	0,060
	Замер 4	–0,431	0,104
Средняя трудность заданий	Замер 1	0,88	0,74
	Замер 2	0,62	0,74
	Замер 3	0,62	0,74
	Замер 4	0,53	0,80

* Ученики, участвовавшие во всех четырех замерах.

но хорошо изученный эффект летних каникул: уровень подготовленности учеников школ «проседает» после длительного перерыва в занятиях [Cooper et al., 1996]. Однако в четвертом замере в модели Раша — Андерсена спад средней способности продолжается, а в модели LLTM — Андерсена происходит ее рост. Дело в том, что, несмотря на схожесть, эти модели оценивают способность по-разному: относительно заданий и относительно когнитивных операций. В последнем замере средняя трудность заданий в модели LLTM — Андерсена, полученная как сумма трудностей задействованных в заданиях когнитивных операций, на 0,27 логита выше, чем средняя трудность заданий в модели Раша — Андерсена. Таким образом, трудность заданий последнего замера в модели Раша — Андерсена может быть занижена, так как в данном замере в мониторинг введены новые содержательные области предмета, а задания по этим областям были простыми относительно якорных заданий. Модель LLTM — Андерсена учитывает, что в этих заданиях задействуются более сложные когнитивные операции, уровень трудности которых моделируется на основе заданий всех четырех замеров. Такое расхождение в динамике средней способности, оцененной в разных моделях, означает, что выводы, сделанные на основе одной из них, возможно, невалидны. Мы считаем, что результаты, полученные в рамках модели LLTM — Андер-

сена, более точно отражают реальность. При этом обе модели оценивают способность на одном и том же материале, и корреляции показателей способностей учеников из одного и того же замера, полученных в разных моделях, выше 0,99. Однако оценки способностей учеников для второго, третьего и четвертого замеров в модели Раша — Андерсена ниже, чем оценки в модели LLTM — Андерсена. Причем от замера к замеру эта разница растет (рис. 4).

Рис. 4. **Связь оценок способностей в двух моделях для всех замеров**



Примечание: Диагональная линия показывает численное соотношение способностей 1:1.

Корреляция трудностей заданий из разных моделей равна 0,56. В модели LLTM — Андерсена трудность заданий определена исходя из суммы трудностей задействованных когнитивных операций, в то время как модель Раша — Андерсона оценивает параметр трудности каждого задания напрямую. В принципе разницу между трудностями заданий из двух моделей можно объяснить *Q*-матрицей, которая не включает всех факторов, формирующих трудность заданий, кроме задействованных когнитивных операций. Такими факторами могут быть, например,

формат заданий или ориентация — теоретическая или практическая. Таким образом, оценка трудности заданий в модели LLTM — Андерсена лишена вклада данных факторов, не относящихся к предметной сложности заданий, а полученные в этой модели оценки способностей и прогресса будут характеризовать когнитивные операции.

В результате для модели LLTM — Андерсена мы получили оценки уровня трудности когнитивных операций (табл. 4). Соотнося способность ученика с уровнем трудности когнитивных операций, мы можем сделать вывод о том, какие операции ученик освоил. Например, ученик со способностью 0,5 логита, вероятно, уже освоил навыки выполнения вычислений с целыми числами, но еще не освоил навыки выполнения вычислений с рациональными числами.

Таблица 4. Т**рудности когнитивных операций из модели LLTM — Андерсена**

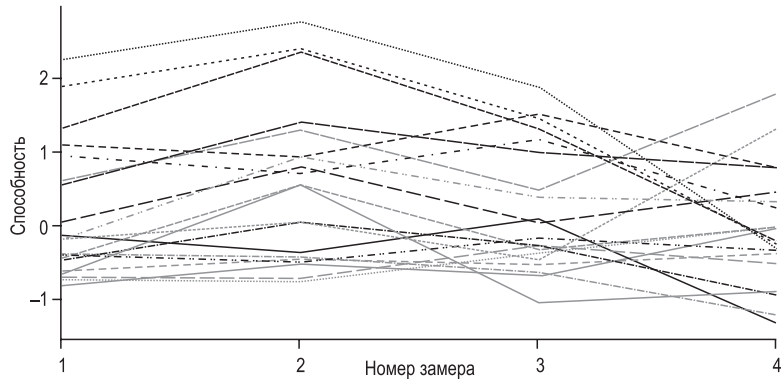
Когнитивная операция	Трудность
Навыки выполнения вычислений с целыми числами	0,026
Навыки выполнения вычислений с рациональными числами	0,647
Навыки построения и применения математических моделей	0,169
Вспоминание фактов/формул и их последующее применение	0,401
Навык преобразования алгебраических выражений	0,143
Навыки решения уравнений	0,426
Навыки представления и считывания информации из текстового описания	–0,503
Навыки выстраивания цепочки решения в несколько шагов	0,575
Навыки выполнения вычислений с вещественными числами	0,021
Навыки работы с координатной плоскостью	0,501
Навыки решения систем уравнений	1,095
Навыки решения неравенств и систем неравенств	–0,360

Полученная иерархия трудностей когнитивных операций по большей части соответствует изначальным предположениям. Трудность около нуля (0,026) имеют задания, в которых необходимо выполнить вычисления с целыми числами, а если требуются вычисления с рациональными числами, трудность задания повышается на 0,647. Минимальную трудность представляет операция «навыки представления и считывания информации из текстового описания», т.е. ученики в первую очередь осваивают данную когнитивную операцию, и наличие в составе задания операции считывания информации из текстового описания снижает трудность задания на 0,503. Если же

задание предполагает выстраивание цепочки решения в несколько шагов, его трудность возрастает на 0,575. Ожидания не подтвердились в отношении двух операций, а именно «навыков выполнения вычислений с вещественными числами» и «навыков решения неравенств и систем неравенств». Оценки их трудности оказались ниже, чем можно было ожидать. Возможно, причина состоит в малом количестве заданий, включающих данные операции (табл. 2). Соответствующая содержательная область только появилась в учебной программе и представлена в мониторинге только базовыми заданиями. Простота этих заданий в рамках своей содержательной области и определила низкие показатели трудности данных когнитивных операций.

Модель LLTM — Андерсена позволяет отслеживать прогресс каждого ученика. На рис. 5 приведены траектории изменения способности 20 случайно выбранных учеников, которые участвовали во всех замерах.

Рис. 5. **Траектории образовательного прогресса учеников**



7. Обсуждение и выводы

Мы рассмотрели способ измерения образовательного прогресса с помощью когнитивных операций, освоение которых проверяется в тесте. Для этой цели применены специальные психометрические методы IRT, а именно модель LLTM, совмещенная с моделью Андерсена для моделирования повторяющихся замеров. В отличие от классической модели Раша — Андерсена, ориентированной на оценку трудности заданий, модель LLTM — Андерсена направлена на анализ трудностей когнитивных операций как составляющих этих заданий. При этом в модели LLTM — Андерсена наличие якорных заданий в последовательных замерах не является обязательным, в отличие от модели Раша — Андерсена. В этом отношении модель LLTM — Андерсена аналогична техникам лонгитюдного моделирования в моделях когнитивной диагностики [Yu, Zhan, Chen, 2023].

Если *Q*-матрица составлена идеально и трудности когнитивных операций, задействованных в заданиях, абсолютно стабильны друг относительно друга и полностью описывают разброс трудностей заданий, то можно использовать непересекающиеся наборы заданий в разных замерах, поскольку в модели LLTM — Андерсена нет индивидуальных трудностей заданий. Однако в реальности составление *Q*-матрицы всегда вносит существенные ограничения в оцениваемые параметры. Соответственно, использование непересекающихся наборов заданий потребует очень сильных допущений, которые могут не выдерживаться в данных.

Использование логики LLTM в моделировании образовательного прогресса означает необходимость учета всех ограничений LLTM. В частности, появляется допущение, что разброс трудностей заданий достаточно хорошо описывается разбросом трудностей когнитивных операций. В реальности такого практически не бывает, из-за чего LLTM никогда не подходит данным лучше Раш-модели и преимущества LLTM перед Раш-моделью являются интерпретационными, а не статистическими. Качество результатов в LLTM очень сильно зависит от качества использованной *Q*-матрицы, что дополнительно актуализирует исследования по валидизации или автоматическому составлению *Q*-матрицы [Sun et al., 2014]. При этом, однако, важно сохранить интерпретируемость LLTM как ее основное достоинство.

Модели LLTM — Андерсена пригодны для анализа лонгитюдной измерительной инвариантности [Vandenberg, Lance, 2000], который не был проведен в данной работе, что является одним из ее ограничений. Анализ измерительной инвариантности — необходимый этап любого лонгитюдного моделирования, предназначенный для подтверждения валидности выводов. Он призван доказать, что психометрические характеристики заданий не меняются со временем. Поскольку в модели LLTM — Андерсена нет индивидуальных характеристик заданий, классические анализы измерительной инвариантности (*Differential Item Functioning* (DIF) анализы) в ней невозможны. Оценить измерительную инвариантность когнитивных операций можно с помощью анализа относительной стабильности их трудностей [Bechger, Maris, 2015]. Для таких анализов необходимо откалибровать LLTM отдельно в каждом замере (без установления общей шкалы), найти разницу между трудностью каждой когнитивной операции и каждой другой, а затем сравнить эти разницы между разными замерами. Если смещения этих разниц несущественны, можно заключить, что трудности когнитивных операций стабильны и все изменения в способности респондентов будут отражены только в изменении оцен-

ки этой способности от одного замера к другому. В силу того, что целью данной статьи было лишь проиллюстрировать и обосновать первое применение такого подхода к измерению образовательного прогресса, мы не проводили полный перечень анализов, необходимых для обеспечения валидности результатов. В этом состоит одно из главных ограничений данного исследования.

Стабильными от замера к замеру в модели LLTM — Андерсена должны оставаться не только параметры заданий (когнитивных операций), но и сама когнитивная карта якорных заданий. Одно и то же задание не может менять состав своих когнитивных операций от одного замера к другому. Например, если задание в первом замере относилось к навыкам выполнения вычислений с целыми числами и навыкам выстраивания цепочки решения в несколько шагов, то и во всех последующих замерах оно должно быть классифицировано точно так же. В таком случае в качестве когнитивных операций не может использоваться какой-либо уровень сложности заданий (базовый или повышенный), так как он может меняться в зависимости от класса: задание, которое представляло повышенную сложность в начале 8-го класса, стало заданием базовой сложности в конце 9-го класса. Соответственно, для составления *Q*-матрицы необходимо использовать только «стабильные» когнитивные операции, интерпретация и проявление которых не меняются. В противном случае они не могут служить порогами на вертикальной шкале интерпретационно и элементами *Q*-матрицы математически.

Одним из главных ограничений предложенного подхода является использование для вынесения суждений о респондентах порогов, которыми в модели LLTM — Андерсена можно считать оцененные трудности когнитивных операций. На том основании, что респондент преодолевает порог, т.е. его способность оказывается выше трудности определенной когнитивной операции, делается вывод об освоении этим респондентом данной когнитивной операции. При принятии решений с высокими ставками в большинстве тестирований пороговые баллы используются с обязательным соблюдением специальных процедур их установления, а именно с участием в этих процедурах представителей всех групп пользователей результатов тестирования: учеников, родителей, учителей, образовательных администраторов и т.д. [Cizek, Bunch, 2007]. Такие процедуры необходимы для минимизации потенциально нежелательных социальных последствий использования, в том числе и некорректного, этих пороговых баллов [Messick, 1998; Hubley, Zumbo, 2011]. Однако рассматриваемые в данной работе пороги являются побочным продуктом предложенной нами методологии,

это пороги исключительно диагностические и предназначены для принятия решений в тестировании с низкими ставками. Такие пороги можно называть критериально-ориентированными диагностическими порогами, потому что они выведены из психометрических параметров заданий, а не на основе распределения групп респондентов, как нормативно ориентированные пороги. Использование данных порогов для принятия решений с высокими ставками влечет за собой риск нежелательных или непредвиденных социальных последствий.

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

Литература

1. Саляхутдинова Д.Р., Федерякин Д.А. (2022) Способы связывания шкал для измерения образовательного прогресса в разных парадигмах анализа данных образовательного тестирования. *Отечественная и зарубежная педагогика*, т. 1, № 3, сс. 98–111. https://doi.org/10.24412/2224-0772-2022-84-98-111
2. Федерякин Д.А., Угланова И.Л., Скрябин М.А. (2021) Новые источники информации в компьютерном тестировании. *Вестник Томского государственного университета*, № 465, сс. 179–187. https://doi.org/10.17223/15617793/465/24
3. Adams R.J. (2005) Reliability as a Measurement Design Effect. *Studies in Educational Evaluation*, vol. 31, no 2, pp. 162–172. https://doi.org/10.1016/j.stueduc.2005.05.008
4. Akaike H. (1974) A New Look at the Statistical Model Identification*. IEEE Transactionson Automatic Control*, vol. 19, no 6, pp. 716–723. https://doi.org/10.1109/TAC.1974.1100705
5. Andersen E.B. (1985) Estimating Latent Correlations between Repeated Testings. *Psychometrika*, vol. 50, March, pp. 3–16. https://doi.org/10.1007/BF02294143
6. Andersen E.B. (1977) Sufficient Statistics and Latent Trait Models. *Psychometrika*, vol. 42, March, pp. 69–81. https://doi.org/10.1007/BF02293746
7. Baker F.B. (1993) Sensitivity of the Linear Logistic Test Model to Misspecification of the Weight Matrix. *Applied Psychological Measurement*, vol. 17, no 3, pp. 201–210. https://doi.org/10.1177/014662169301700301
8. Bechger T.M., Maris G. (2015) A Statistical Test for Differential Item Pair Functioning. *Psychometrika*, vol. 80, June, pp. 317–340. https://doi.org/10.1007/s11336-014-9408-y
9. Cizek G.J. (ed.) (2013) *Vertically Moderated Standard Setting: A Special Issue of Applied Measurement in Education*. New York, NY: Routledge. https://doi.org/10.4324/97813150459008
10. Cizek G.J., Bunch M.B. (2007) *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests.* Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781412985918
11. Cooper H., Nye B., Charlton K., Lindsay J., Greathouse S. (1996) The Effects of Summer Vacation on Achievement Test Scores: A Narrative and Meta-Analytic Review. *Review of Educational Research*, vol. 66, no 3, pp. 227–268. https://doi.org/10.2307/1170523

12. Deonovic B., Chopade P., Yudelson M., de la Torre J., von Davier A.A. (2019) Application of Cognitive Diagnostic Models to Learning and Assessment Systems. *Handbook of Diagnostic Classification Models: Models and Model Extensions, Applications, Software Packages* (eds M. von Davier, Y.-S. Lee), Cham: Springer, pp. 437–460. https://doi.org/10.1007/978-3-030-05584-4_21
13. Dimitrov D.M., Rumrill Jr. P.D. (2003) Pretest-Posttest Designs and Measurement of Change. *Work*, vol. 20, no 2, pp. 159–165.
14. Embretson S.E. (1991) A Multidimensional Latent Trait Model for Measuring Learning and Change. *Psychometrika*, vol. 56, September, pp. 495–515. https://doi.org/10.1007/BF02294487
15. Fischer G.H. (1995) The Linear Logistic Test Model. *Rasch Models* (eds G.H. Fischer, I.W. Molenaar), New York, NY: Springer, pp. 131–155. https://doi.org/10.1007/978-1-4612-4230-7_8
16. 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
17. Gideon S. (1978) Estimating the Dimension of a Model. *The Annals of Statistics*, vol. 6, no 2, pp. 461–464. https://doi.org/10.1214/aos/1176344136
18. Hubley A.M., Zumbo B.D. (2011) Validity and the Consequences of Test Interpretation and Use. *Social Indicators Research*, vol. 103, no 2, pp. 219–230. https://doi.org/10.1007/s11205-011-9843-4
19. Lee H.K. (2016) *An Application of Item Response Theory to Investigate the Validity of a Learning Progression for Number Sense* (PhD Thesis). Berkeley, CA: University of California.
20. Linden van der W. J. (2018) *Handbook of Item Response Theory: Three Volume Set*. Boca Raton, FL: Chapman and Hall/CRC. https://doi.org/10.1201/9781315119144
21. Loyd B.H., Hoover H.D. (1980) Vertical Equating Using the Rasch Model. *Journal of Educational Measurement*, vol. 17, no 3, pp. 179–193. https://doi.org/10.1111/j.1745-3984.1980.tb00825.x
22. Macdonald G.T. (2014) *The Performance of the Linear Logistic Test Model When the Q-Matrix Is Misspecified: A Simulation Study* (PhD Thesis). Tampa, FL: University of South Florida.
23. Messick S. (1998) Test Validity: A Matter of Consequence. *Social Indicators Research*, vol. 45, November, pp. 35–44. https://doi.org/10.1023/A:1006964925094
24. Nsowaa B. (2018) *The Ordered Latent Transition Analysis Model for the Measurement of Learning* (PhD Thesis). New York, NY: Columbia University.
25. Rolfes T., Roth J., Schnotz W. (2018) Effects of Tables, Bar Charts, and Graphs on Solving Function Tasks. *Journal für Mathematik-Didaktik,* vol. 39, no 1, pp. 97–125. http://dx.doi.org/10.1007/s13138-017-0124-x
26. Slavin R.E. (2005) *Evidence-Based Reform: Advancing the Education of Students at Risk. Report Prepared for Renewing Our Schools, Securing Our Future*. Available at: https://goo.su/vYeO (accessed 20 July 2023).
27. Sontag L.M. (1984) *Vertical Equating Methods: A Comparative Study of Their Efficacy*. New York, NY: Columbia University.
28. Sun Y., Ye S., Inoue S., Sun Y. (2014) Alternating Recursive Method for Q-matrix Learning. Proceedings of the *7th International Conference on Educational Data Mining* (London, July 4–7, 2014), pp. 14–20.
29. Vandenberg R.J., Lance C.E. (2000) A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, vol. 3, no 1, pp. 4–70. https://doi.org/10.1177/109442810031002
30. Waterbury G.T., DeMars C.E. (2021) Anchors Aweigh: How the Choice of Anchor Items Affects the Vertical Scaling of 3PL Data with the Rasch Model.

Educational Assessment, vol. 26, no 3, pp. 175–197. https://doi.org/10.1080/10627197.2020.1858782

31. Wilson M., Zheng X., McGuire L. (2012) Formulating Latent Growth Using an Explanatory Item Response Model Approach. *Journal of Applied Measurement*, vol. 13, no 1, pp. 1–22.
32. Yu X., Zhan P., Chen Q. (2023) Don't Worry about the Anchor-Item Setting in Longitudinal Learning Diagnostic Assessments. *Frontiers in Psychology*, vol. 14, February, Article no 1112463. https://doi.org/10.3389/fpsyg.2023.1112463

References

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

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

Andersen E.B. (1985) Estimating Latent Correlations between Repeated Testings. *Psychometrika*, vol. 50, March, pp. 3–16. https://doi.org/10.1007/BF02294143

Andersen E.B. (1977) Sufficient Statistics and Latent Trait Models. *Psychometrika*, vol. 42, March, pp. 69–81. https://doi.org/10.1007/BF02293746

Baker F.B. (1993) Sensitivity of the Linear Logistic Test Model to Misspecification of the Weight Matrix. *Applied Psychological Measurement*, vol. 17, no 3, pp. 201–210. https://doi.org/10.1177/014662169301700301

Bechger T.M., Maris G. (2015) A Statistical Test for Differential Item Pair Functioning. *Psychometrika*, vol. 80, June, pp. 317–340. https://doi.org/10.1007/s11336-014-9408-y

Cizek G.J. (ed.) (2013) *Vertically Moderated Standard Setting: A Special Issue of Applied Measurement in Education*. New York, NY: Routledge. https://doi.org/10.4324/97813150459008

Cizek G.J., Bunch M.B. (2007) *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests.* Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781412985918

Cooper H., Nye B., Charlton K., Lindsay J., Greathouse S. (1996) The Effects of Summer Vacation on Achievement Test Scores: A Narrative and Meta-Analytic Review. *Review of Educational Research*, vol. 66, no 3, pp. 227–268. https://doi.org/10.2307/1170523

Deonovic B., Chopade P., Yudelson M., de la Torre J., von Davier A.A. (2019) Application of Cognitive Diagnostic Models to Learning and Assessment Systems. *Handbook of Diagnostic Classification Models: Models and Model Extensions, Applications, Software Packages* (eds M. von Davier, Y.-S. Lee), Cham: Springer, pp. 437–460. https://doi.org/10.1007/978-3-030-05584-4_21

Dimitrov D.M., Rumrill Jr. P.D. (2003) Pretest-Posttest Designs and Measurement of Change. *Work*, vol. 20, no 2, pp. 159–165.

Embretson S.E. (1991) A Multidimensional Latent Trait Model for Measuring Learning and Change. *Psychometrika*, vol. 56, September, pp. 495–515. https://doi.org/10.1007/BF02294487

Federiakin D.A. Uglanova I.L., Skryabin M.A. (2021) Novye istochniki informatsii v komp'yuternom testirovanii [New Sources of Information in Computerized Testing]. *Tomsk State University Journal*, no 465, pp. 179–187. https://doi.org/10.17223/15617793/465/24

Fischer G.H. (1995) The Linear Logistic Test Model. *Rasch Models* (eds G.H. Fischer, I.W. Molenaar), New York, NY: Springer, pp. 131–155. https://doi.org/10.1007/978-1-4612-4230-7_8

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

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

Hubley A.M., Zumbo B.D. (2011) Validity and the Consequences of Test Interpretation and Use. *Social Indicators Research*, vol. 103, no 2, pp. 219–230. https://doi.org/10.1007/s11205-011-9843-4

Lee H.K. (2016) *An Application of Item Response Theory to Investigate the Validity of a Learning Progression for Number Sense* (PhD Thesis), Berkeley, CA: University of California.

Linden van der W. J. (2018) *Handbook of Item Response Theory: Three Volume Set*. Boca Raton, FL: Chapman and Hall/CRC. https://doi.org/10.1201/9781315119144

Loyd B.H., Hoover H.D. (1980) Vertical Equating Using the Rasch Model. *Journal of Educational Measurement*, vol. 17, no 3, pp. 179–193. https://doi.org/10.1111/j.1745-3984.1980.tb00825.x

Macdonald G.T. (2014) *The Performance of the Linear Logistic Test Model When the Q-Matrix Is Misspecified: A Simulation Study* (PhD Thesis). Tampa, FL: University of South Florida.

Messick S. (1998) Test Validity: A Matter of Consequence. *Social Indicators Research*, vol. 45, November, pp. 35–44. https://doi.org/10.1023/A:1006964925094

Nsowaa B. (2018) *The Ordered Latent Transition Analysis Model for the Measurement of Learning* (PhD Thesis). New York, NY: Columbia University.

Rolfes T., Roth J., Schnotz W. (2018) Effects of Tables, Bar Charts, and Graphs on Solving Function Tasks. *Journal für Mathematik-Didaktik,* vol. 39, no 1, pp. 97–125. http://dx.doi.org/10.1007/s13138-017-0124-x

Salyakhutdinova D.R., Federiakin D.A. (2022) Sposoby svyazyvaniya shkal dlya izmereniya obrazovatel'nogo progressa v raznykh paradigmakh analiza dannykh obrazovatel'nogo testirovaniya [Methods of Linking Scales for Measuring Educational Progress in Different Paradigms of Educational Testing Data Analysis]. *Domestic and Foreign Pedagogy*, vol. 1, no 3, pp. 98–111. https://doi.org/10.24412/2224-0772-2022-84-98-111

Slavin R.E. (2005) *Evidence-Based Reform: Advancing the Education of Students at Risk. Report Prepared for Renewing Our Schools, Securing Our Future*. Available at: https://goo.su/vYeO (accessed 20 July 2023).

Sontag L.M. (1984) *Vertical Equating Methods: A Comparative Study of Their Efficacy*. New York, NY: Columbia University.

Sun Y., Ye S., Inoue S., Sun Y. (2014) Alternating Recursive Method for Q-matrix Learning. Proceedings of the *7th International Conference on Educational Data Mining* (London, July 4–7, 2014), pp. 14–20.

Vandenberg R.J., Lance C.E. (2000) A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, vol. 3, no 1, pp. 4–70. https://doi.org/10.1177/109442810031002

Waterbury G.T., DeMars C.E. (2021) Anchors Aweigh: How the Choice of Anchor Items Affects the Vertical Scaling of 3PL Data with the Rasch Model. *Educational Assessment*, vol. 26, no 3, pp. 175–197. https://doi.org/10.1080/10627197.2020.185878231

Wilson M., Zheng X., McGuire L. (2012) Formulating Latent Growth Using an Explanatory Item Response Model Approach. *Journal of Applied Measurement*, vol. 13, no 1, pp. 1–22.

Yu X., Zhan P., Chen Q. (2023) Don't Worry about the Anchor-Item Setting in Longitudinal Learning Diagnostic Assessments. *Frontiers in Psychology*, vol. 14, February, Article no 1112463. https://doi.org/10.3389/fpsyg.2023.1112463