



МНОГОКОМПОНЕНТНАЯ МОДЕЛЬ ТЕОРИИ ТЕСТИРОВАНИЯ ДЛЯ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ С РАЗНОРОДНЫМИ ФОРМАМИ ЗАДАНИЙ

Родионов А. В., Торопов В. Д.

*Федеральное государственное бюджетное образовательное учреждение высшего образования
«Байкальский государственный университет», Иркутск, Российская Федерация,
e-mail: avr-v@yandex.ru*

В статье рассматривается задача анализа результатов промежуточной аттестации в высшем образовании в условиях, когда испытание имеет неоднородную структуру. Например, задание состоит из нескольких компонентов, тестовая часть с закрытыми типами вопросов сочетается с заданиями открытого типа, оцениваемыми преподавателем. Целью исследования является разработка и теоретическое обоснование многокомпонентной модели теории тестирования, объединяющей обобщенную модель частичного кредита для закрытых заданий и модель градуированного отклика с параметром строгости эксперта для открытых заданий в трехмерном латентном пространстве «знания – умения – навыки». Параметры модели оцениваются методом предельного максимального правдоподобия с реализацией через EM алгоритм. Идентифицируемость параметров и корректность их восстановления проверены методом имитационного моделирования. Апробация модели проведена на эмпирических данных зачета по дисциплине «Информационные технологии» с участием четырех преподавателей. Показано, что оценки параметров фасета преподавателя статистически значимо различаются, и это можно интерпретировать как возможные различия строгости оценивания при принятых модельных предположениях. Модель может служить инструментом для выявления потенциальных рэйтер-эффектов, требующих дополнительной экспертной проверки. Пригодность модели для описания эмпирических данных подтверждена критериями согласия как на уровне отдельных заданий, так и на уровне модели в целом. Результаты работы представляют интерес для специалистов в области педагогических измерений и разработчиков фондов оценочных средств в системе высшего образования.

Ключевые слова: теория тестирования, многокомпонентная модель, смешанное оценивание, частичное оценивание, строгость эксперта, высшее образование, промежуточная аттестация, латентные переменные

A MULTICOMPONENT ITEM RESPONSE THEORY MODEL FOR INTERIM ASSESSMENT WITH HETEROGENEOUS TASK FORMATS

Rodionov A. V., Toropov V. D.

*Federal State Budgetary Educational Institution of Higher Education
“Baikal State University”, Irkutsk, Russian Federation, e-mail: avr-v@yandex.ru*

This article examines the problem of analyzing midterm assessment results in higher education under conditions where the test has a heterogeneous structure. For example, a task consists of several components, with a test section with closed-ended questions combined with open-ended questions assessed by a teacher. The aim of the study is to develop and theoretically substantiate a multicomponent test theory model combining a generalized partial credit model for closed-ended questions and a graded response model with an expert strictness parameter for open-ended questions in a three-dimensional latent knowledge-skills-abilities space. The model parameters are estimated using the maximum likelihood method implemented through the EM algorithm. The identifiability of the parameters and the correctness of their recovery are verified using simulation modeling. The model was tested using empirical data from a test in information technology involving four teachers. It is shown that the estimates of the teacher facet parameters differ statistically significantly, which can be interpreted as potential differences in assessment strictness under the adopted model assumptions. The model can serve as a tool for identifying potential rater effects that require additional expert verification. The model's suitability for describing empirical data is confirmed by goodness-of-fit criteria at both the individual task level and the model as a whole. The results of this study are of interest to specialists in educational measurement and developers of assessment tools in higher education.

Keywords: item response theory, multi-component model, mixed assessment, partial credit, rater severity, higher education, interim assessment, latent variables

Введение

Современная теория тестирования (Item Response Theory, IRT) представляет собой семейство вероятностных моделей, связывающих наблюдаемый отклик испытуемого с латентной способностью и параметрами задания. Такие модели, как однопараметрическая модель Раша, двух- и трехпараметрические логистические модели, были разработаны преимущественно для стандартизированных дихотомических тестов и ориентированы на измерение одномерного конструкта. Расширение сферы применения IRT на полиномические отклики привело к появлению моделей градуированного отклика (GRM) [1] и обобщенной

метрическая логистические модели, были разработаны преимущественно для стандартизированных дихотомических тестов и ориентированы на измерение одномерного конструкта. Расширение сферы применения IRT на полиномические отклики привело к появлению моделей градуированного отклика (GRM) [1] и обобщенной

модели частичного кредита (GPCM) [2], а учет субъективности экспертного оценивания – к многофасетной модели Раша (MFRM) [3, 4]. В последние годы активно развиваются модели, ориентированные именно на учет строгости экспертов и неоднородность формата заданий: предложены спецификации обобщенной модели частичного кредита с параметром строгости и его дрейфом в процессе оценивания [5], байесовские варианты многофасетной модели Раша с марковским моделированием дрейфа строгости [6], многомерные модели теории тестирования для оценивания письменных работ [7]. Вопросы устойчивости оценивания политомических моделей при малых выборках исследуются с использованием имитационного моделирования [8]. Предложены спецификации для учета «центрированности» оценок в рамках иерархической рейтинговой модели [9], когнитивно-диагностические модели эффектов эксперта [10], а также иерархические фасетные модели, совместно анализирующие выставленные баллы и затраченное на оценивание время [11]. Развиваются и методы согласования в случае ручной и автоматической проверки ответов, в том числе подходы к выравниванию баллов в рейтинг-опосредованном оценивании [12].

В то же время сложившаяся практика проведения промежуточной аттестации в российских образовательных организациях высшего образования формирует ряд требований, которые не в полной мере учтены каждым из указанных классов моделей по отдельности. Современные федеральные государственные образовательные стандарты ориентированы на компетентностный подход, при котором результат освоения образовательной программы описывается через совокупность формируемых у обучающегося компетенций. При этом компетенции не противопоставляются традиционной триаде «знания – умения – навыки», а содержательно опираются на нее: знания образуют теоретическую основу деятельности, умения характеризуют способность применять эти знания при решении учебных и профессионально ориентированных задач, а навыки отражают степень освоения устойчивых способов выполнения действий. Типичный экзаменационный билет, как правило, включает три взаимосвязанных компонента: тестовую часть с закрытыми заданиями различных форматов, допускающими автоматизированную проверку; задание, ориентированное на оценку умений обучающегося; задание, позволяющее выявить уровень сформированности практических

навыков. Последние два компонента обычно оцениваются экспертным способом на основе рубрик, что связано с неизбежным элементом субъективности при интерпретации результата и выставлении итоговой оценки.

Применение рассмотренных IRT-моделей к такой структуре экзамена имеет ряд ограничений. Во-первых, экзаменационные задания неоднородны – задания закрытого типа преимущественно проверяют воспроизведение и распознавание знаний, тогда как задания открытого типа направлены на оценку умений и навыков, связанных с самостоятельным решением задач. Во-вторых, частичный балл формируется по-разному: в закрытых заданиях он складывается из отдельных элементов ответа, а в открытых определяется экспертом целостно по шкале критериев. В-третьих, экспертное оценивание неизбежно содержит субъективный компонент, поскольку разные преподаватели могут отличаться степенью строгости. Отдельные модели учитывают эти особенности лишь частично. MFRM позволяет корректировать влияние эксперта, но не описывает неоднородность компонентов экзамена. GPCM и GRM применимы к заданиям с частичным баллом и упорядоченными категориями, однако не объединяют разные механизмы оценивания.

Цель исследования – разработка и теоретическое обоснование многокомпонентной IRT-модели оценивания результатов промежуточной аттестации, объединяющей GPCM для закрытых заданий и GRM с параметром экспертной строгости для открытых заданий.

Материалы и методы исследования

Формализацию модели начнем с описания латентной структуры. Каждому студенту $i = 1, \dots, N$ сопоставим вектор латентных способностей:

$$\theta_i = (\theta_i^K, \theta_i^U, \theta_i^N),$$

где θ_i^K – латентная способность по компоненту знаний; θ_i^U – латентная способность по компоненту умений; θ_i^N – латентная способность по компоненту навыков.

$$\theta_i \sim \mathcal{N}_3(0, \Sigma),$$

где Σ – ковариационная матрица латентных способностей.

Экзамен состоит из трех компонентов $k \in \{K, U, N\}$, где K – знания (закрытые задания с автоматической проверкой); U – умения (открытые задания); N – навыки (открытые задания).

Обозначим: $j = 1, \dots, J_k$ – индекс задания в компоненте k , где J_k – количество заданий;
 m_{jk} – максимальный балл за отдельное задание j в компоненте k ;
 a_{jk} – параметр дискриминативности;
 $b_{jk}^{c,c}$ – пороговый параметр трудности c задания j в компоненте k ;
 w_{jk} – весовой коэффициент задания j в компоненте k , определяемый как отношение максимального балла за задание к суммарному максимальному баллу компонента;
 δ_r – параметр строгости эксперта (преподавателя) r для открытых заданий.

Для заданий закрытого типа (компонент K) мы опираемся на обобщенную модель частичного кредита (GPCM), дополнив ее весовыми коэффициентами, чтобы учесть разный вклад заданий в итоговый балл компонента. Пусть $X_{ij}^K \in \{0, 1, \dots, m_{jk}\}$ наблюдаемый балл студента i за задание j компонента знаний. Тогда вероятность получения категории c :

$$P(X_{ij}^K = c | \theta_i^K, a_{jk}, b_{jk,s}) = \frac{\exp\left(\sum_{s=1}^c a_{jk}(\theta_i^K - b_{jk,s})\right)}{\sum_{h=0}^{m_{jk}} \exp\left(\sum_{s=1}^h a_{jk}(\theta_i^K - b_{jk,s})\right)}, c = 0, \dots, m_{jk},$$

где $\sum_{s=1}^0 (\cdot) \equiv 0$.

Таким образом,

$$X_{ij}^K | \theta_i^K \sim \text{GPCM}(a_{jk}, b_{jk,1}, \dots, b_{jk,m_{jk}}).$$

Для заданий открытого типа (компоненты U и N) применим модифицированную модель градуированного отклика (GRM), дополненную параметром экспертной строгости. При экспертной проверке практических работ преподаватель фактически принимает решение по упорядоченной шкале качества. Кумулятивная логика GRM естественно отражает именно такую последовательную природу суждения: $\{\text{«не приступил»} < \text{«частично верно»} < \text{«в основном верно»} < \text{«полностью верно»}\}$. Кумулятивная вероятность получения балла не ниже порога c :

$$P^*(X_{ijk} \geq c | \theta_i^k, a_{jk}, b_{jk,c}, \delta_r) = \frac{1}{1 + \exp(-a_{jk}(\theta_i^k - b_{jk,c} - \delta_r))}, c = 1, \dots, m_{jk}.$$

Вероятность конкретной категории c :

$$P(X_{ijk} = c) = P^*(X_{ijk} \geq c) - P^*(X_{ijk} \geq c + 1),$$

при условиях $P^*(X_{ijk} \geq 0) = 1, P^*(X_{ijk} \geq m_{jk} + 1) = 0$.

То есть

$$X_{ijk} | \theta_i^k, \delta_r \sim \text{GRM}(a_{jk}, b_{jk,1}, \dots, b_{jk,m_{jk}}, \delta_r), k \in \{U, N\}.$$

Тогда полное правдоподобие наблюдений для одного студента записывается в виде произведения по всем заданиям трех компонентов:

$$P(X_i | \theta_i, \Omega) = \prod_{j=1}^{J_K} P^{\text{GPCM}}(X_{ij}^K | \theta_i^K, a_{jk}, b_{jk}) \prod_{k \in U, N} \prod_{j=1}^{J_k} P^{\text{GRM}}(X_{ijk} | \theta_i^k, a_{jk}, b_{jk}, \delta_r),$$

где $\Omega = \{a_{jk}, b_{jk,c}, w_{jk}, \delta_r\}$.

При переходе к полной выборке в модель дополнительно включается плотность $\varphi_3(\theta_i | 0, \Sigma)$, задающая трехмерное нормальное распределение латентных способностей обучающегося с нулевым математическим ожиданием и ковариационной матрицей Σ .

Для оценки параметров модели воспользуемся методом предельного максимального правдоподобия (MML) [13], рассматривая латентные способности как ненаблюдаемые случайные величины, по которым выполняется интегрирование. Вычислительно эта задача решается итеративным ЕМ-алгоритмом.

Пусть имеются наблюдения $X = \{X_i\}$ для $i = 1, \dots, N$ студентов. Вектор всех параметров заданий и экспертов обозначим $\Omega = \{a_{jk}, b_{jk,c}, \delta_r\}$. Условная функция правдоподобия для одного студента при фиксированном θ_i :

$$L(\Omega | X_i, \theta_i) = \prod_{j=1}^{J_k} P^{\text{GPCM}}(X_{ij}^K | \theta_i^K, a_{jk}, b_{jk}) \prod_{k \in U, N} \prod_{j=1}^{J_k} P^{\text{GRM}}(X_{ijk} | \theta_i^k, a_{jk}, b_{jk}, \delta_r).$$

Функция правдоподобия получается интегрированием по латентным способностям:

$$L^*(\Omega, \Sigma | X) = \prod_{i=1}^N \int L(\Omega | X_i, \theta_i) \varphi_3(\theta_i | 0, \Sigma) d\theta_i.$$

Логарифм функции правдоподобия:

$$\ell^*(\Omega, \Sigma) = \sum_{i=1}^N \ln \int L(\Omega | X_i, \theta_i) \varphi_3(\theta_i | 0, \Sigma) d\theta_i.$$

ЕМ алгоритм на каждом шаге попеременно уточняет оценки $\hat{\Omega}$ и $\hat{\Sigma}$. Содержательно Е-шаг отвечает на вопрос «где, вероятнее всего, расположен каждый студент в латентном пространстве при текущих параметрах». На итерации (t) при текущих оценках параметров $\Omega^{(t)}$, $\Sigma^{(t)}$ для каждого студента i вычисляется апостериорная плотность латентных способностей по формуле Байеса:

$$g(\theta_i | X_i, \Omega^{(t)}, \Sigma^{(t)}) = \frac{L(\Omega^{(t)} | X_i, \theta_i) \varphi_3(\theta_i | 0, \Sigma^{(t)})}{\int L(\Omega^{(t)} | X_i, \theta) \varphi_3(\theta | 0, \Sigma^{(t)}) d\theta}.$$

Ожидаемое полное логарифмическое правдоподобие (Q -функция), максимизируемое на M -шаге, имеет вид

$$Q(\Omega, \Sigma | \Omega^{(t)}, \Sigma^{(t)}) = \sum_{i=1}^N \int [\ln L(\Omega | X_i, \theta_i) + \ln \varphi_3(\theta_i | 0, \Sigma)] g(\theta_i | X_i, \Omega^{(t)}, \Sigma^{(t)}) d\theta_i.$$

Здесь проявляется удобное свойство выбранной структуры: благодаря аддитивности логарифма Q -функция декомпозируется на три слагаемых, соответствующих компонентам K , U , N , плюс слагаемое для ковариационной матрицы

$$Q = Q^K(\Omega_K) + Q^U(\Omega_U, \delta) + Q^N(\Omega_N, \delta) + Q_\Sigma(\Sigma),$$

где

$$Q^K(\Omega_K) = \sum_i \sum_j \int \ln P^{\text{GPCM}}(X_{ij}^K | \theta_i^K, a_{jk}, b_{jk}) g(\theta_i | X_i) d\theta_i,$$

$$Q^U(\Omega_U, \delta) = \sum_i \sum_j \int \ln P^{\text{GRM}}(X_{ij}^U | \theta_i^U, a_{jU}, b_{jU}, \delta_r) g(\theta_i | X_i) d\theta_i,$$

$$(\text{и аналогично для } Q^N), \text{ а также } Q_\Sigma(\Sigma) = \sum_i \int \ln \varphi_3(\theta_i | 0, \Sigma) g(\theta_i | X_i) d\theta_i.$$

Интеграл по $\theta_i = (\theta_i^K, \theta_i^U, \theta_i^N)$ не имеет аналитического решения, поэтому его приходится аппроксимировать численно. Используем квадратуру Гаусса – Эрмита [13]. Пусть $\{\theta_q, w_q\}$, $q = 1, \dots, Q$ – набор узлов и весов одномерной квадратуры, а Q^3 – размер полной трехмерной сетки. Тогда

$$\int f(\theta) \varphi_3(\theta | 0, \Sigma) d\theta \approx \sum_{q^1, q^2, q^3} f(\theta_{q^1 q^2 q^3}) w_{q^1} w_{q^2} w_{q^3},$$

где $\theta_{q^1 q^2 q^3} = (\theta_{q^1}, \theta_{q^2}, \theta_{q^3})$ – узел сетки. При ненулевых корреляциях между компонентами Σ применяется разложение Холецкого $\Sigma = LL'$ и узлы трансформируются $\theta_q \rightarrow Lz_q$, где z_q – стандартные узлы. На практике можно взять 15–21 узлов на каждое измерение, что при трех измерениях дает $Q^3 \approx 3375$ – 9261 узлов.

На Е-шаге для каждого студента i вычисляются апостериорные вероятности узлов квадратуры:

$$\hat{g}_i(\theta_q) = \frac{L(\Omega^{(i)} | X_i, \theta_q) \varphi_3(\theta_q)}{\sum_{q'} L(\Omega^{(i)} | X_i, \theta_{q'}) \varphi_3(\theta_{q'})}.$$

Затем для каждого задания j компонента k и каждой категории c вычисляются ожидаемые «достаточные статистики»

$$\bar{r}_{jk,c}(\theta_q) = \sum_i I(X_{ijk} = c) \hat{g}_i(\theta_q)$$

и ожидаемое количество студентов в узле θ_q – $\bar{n}(\theta_q) = \sum_i \hat{g}_i(\theta_q)$.

M -шаг заключается в пересчете параметров заданий посредством раздельной максимизации каждого слагаемого Q -функции. Далее выполняется обновление параметров GPCM (компонент K). Для каждого задания j компонент знаний максимизируется:

$$Q_j^K(a_{jk}, b_{jk}) = \sum_q \sum_{c=0}^{m_{jk}} \bar{r}_{jk,c}(\theta_q) \ln P^{\text{GPCM}}(c | \theta_q^K, a_{jk}, b_{jk}) \rightarrow \max_{a_{jk}, b_{jk}}.$$

Максимизация выполняется методом Ньютона – Рафсона с использованием матрицы Гессияна соответствующего слагаемого.

Обновление параметров GRM (компоненты U, N) – для каждого задания j компонентов умений или навыков максимизируется:

$$Q_{jk}(a_{jk}, b_{jk}, \delta_r) = \sum_q \sum_{c=0}^{m_{jk}} \bar{r}_{jk,c}(\theta_q) \ln P^{\text{GRM}}(c | \theta_q^k, a_{jk}, b_{jk}, \delta_r) \rightarrow \max.$$

Ковариационная матрица Σ обновляется на основании апостериорных ожиданий с использованием квадратуры

$$\Sigma^{(t+1)} = \frac{1}{N} \sum_i \sum_q \hat{g}_i(\theta_q) \theta_q \theta_q'.$$

Поскольку полученная на этом шаге матрица в общем случае не удовлетворяет условиям идентифицируемости ($\sigma_{kk} = 1$), то выполняется перенормировка

$$\tilde{\Sigma}_{kl} = \frac{\Sigma_{kl}}{\sqrt{\Sigma_{kk} \cdot \Sigma_{ll}}},$$

что фактически переводит ковариационную матрицу в корреляционную. Соответственно корректируются параметры заданий

$$b_{jk,c} \rightarrow \frac{b_{jk,c}}{\sqrt{\Sigma_{kk}}} \quad \text{и} \quad a_{jk} \rightarrow a_{jk} \sqrt{\Sigma_{kk}}.$$

Стандартные ошибки получают из наблюдаемой информационной матрицы Фишера, а поскольку речь идет об MML-оценках с проинтегрированными латентными переменными, то целесообразно применить тождество Луи [14], выражающее

наблюдаемую информацию через score-функцию полных данных:

$$I(\hat{\Omega}) = -\frac{\partial^2 \ell^*}{\partial \Omega^2} = \sum_i \left[E_i(\tilde{S}_i \tilde{S}_i') - E_i(\tilde{S}_i) E_i(\tilde{S}_i') \right],$$

где $\tilde{S}_i = \frac{\partial \ln L(\Omega | X_i, \theta_i)}{\partial \Omega}$ – вектор score-

функции для полных данных, а $E_i(\cdot)$ – ожидание по апостериорному распределению $g(\theta_i | X_i, \hat{\Omega})$.

Стандартные ошибки вычисляются по формуле

$$SE(\hat{\Omega}) = \sqrt{\text{diag}(I^{-1}(\hat{\Omega}))}.$$

Для обеспечения идентифицируемости модели вводятся следующие ограничения:

$$\mu = 0; \sigma_{kk} = 1 \text{ для всех } k; \sum_r \delta_r = 0.$$

Для анализа соответствия отдельных заданий параметрам модели применяются два среднеквадратичных показателя: взвешенный показатель Infit и невзвешенный показатель Outfit [15]. Показатель Outfit характеризует общую величину отклонений

наблюдаемых ответов от модельных ожиданий и для задания j определяется как среднее значение квадратов стандартизованных остатков

$$Outfit_j = \frac{1}{N} \sum_i z_{ij}^2.$$

Показатель Infit учитывает вес каждого наблюдения и потому является более чувствительным к отклонениям в той области шкалы, где задание наиболее информативно. Он вычисляется по формуле

$$Infit_j = \frac{\sum_i W_{ij} z_{ij}^2}{\sum_i W_{ij}},$$

где $z_{ij} = \frac{X_{ij} - E[X_{ij}]}{\sqrt{Var(X_{ij})}}$, а весовой коэф-

фициент определяется как $W_{ij} = Var(X_{ij})$. Значения Infit и Outfit, близкие к единице, свидетельствуют о согласовании эмпирических данных с модельными предпосылками. В прикладных исследованиях в качестве приемлемого обычно рассматривается диапазон от 0,7 до 1,3.

Для политомических заданий дополнительно используется обобщенная статистика [16] $S - X^2$, позволяющая учитывать оценку латентной способности испытуемых. Данный показатель основан на сравнении наблюдаемых и ожидаемых частот баллов в группах респондентов, сформированных по суммарному тестовому результату. Для задания j статистика имеет вид

$$S - X_j^2 = \sum_j \frac{N_g (O_{jg} | E_{jg})^2}{E_{jg} (1 | E_{jg})},$$

где O_{jg} обозначает наблюдаемую долю ответов в балльной группе g , а E_{jg} – соответствующую модельную долю.

На уровне модели в целом применяется ограниченно-информационная статистика согласия $M2$ [17]. На ее основе определяется показатель корневого среднеквадратичного приближения

$$RMSEA = \sqrt{\frac{\max(\frac{M2 - df}{N - 1} | 0)}{df}}.$$

Значения RMSEA [17] ниже 0,06 обычно интерпретируются как свидетельство хорошего согласия модели с данными.

Результаты исследования и их обсуждение

Для оценки параметров модели разработан программный модуль на языке Python (библиотеки NumPy, SciPy, Pandas), реализующий вычисление вероятностей GPCM и GRM, апостериорных весов квадратурных узлов, Q-функции и ее градиентов. Прежде чем применять предложенную модель к эмпирическим данным, необходимо убедиться в корректности восстановления ее параметров в контролируемых условиях. Особое значение имеет идентифицируемость параметров экспертной строгости δ_r в ситуации, когда каждый преподаватель, как правило, оценивает работы только закрепленных за ним учебных групп. В таких условиях индивидуальная строгость эксперта может быть смешана с реальным уровнем подготовки соответствующей группы.

Для проверки возможности разделения этих эффектов был проведен численный эксперимент методом Монте-Карло. Его цель состояла в том, чтобы установить, позволяет ли модель за счет общего автоматически проверяемого компонента K отличить влияние строгости преподавателя от различий в уровне подготовленности обучающихся. Сгенерировано 500 наборов данных ($N = 200$, девять учебных групп по три на каждого из трех преподавателей) с истинными значениями строгости $\delta_1 = 0,30$, $\delta_2 = -0,45$, $\delta_3 = 0,15$ ($\sum_r \delta_r = 0$) и реалистичной неоднородностью групп ($\pm 0,15$ SD по уровню подготовки). Компонент K моделировался 30 политомическими – заданиями (18 дихотомическими, 6 с $m = 3$, 6 с $m = 4$), компоненты U и N – одним заданием с семью градациями каждый.

Латентные уровни подготовленности и параметры заданий генерировались из стандартных распределений, а параметры строгости преподавателей задавались фиксированными значениями с нулевой суммой: $\theta_i \sim N_3(0, 1)$, $\beta_j \sim N_3(0, 1)$, $\delta_1 = 0,30$, $\delta_2 = -0,45$, $\delta_3 = 0,15$ ($\sum_r \delta_r = 0$), $\tau_1 < \tau_2 < \dots < \tau_m$, где θ_i – латентный уровень обучающегося, β_j – трудность (положение) задания, δ_r – параметр строгости (фасета) преподавателя, τ_k – пороги категорий политомического задания. Порождающая модель для смежных категорий имеет вид

$$\ln \left(\frac{P(Y_{jr} | k)}{P(Y_{jr} = k - 1)} \right) = a_j(\theta_i | \beta_j | \delta_r | \tau_k),$$

где a_j – параметр дискриминации задания. Дискриминации a_j при генерации задава-

лись из логнормального распределения с медианой около единицы (умеренная дискриминация), а пороги категорий τ_k – как упорядоченные приращения, обеспечивающие условие $\tau_1 < \tau_2 < \dots < \tau_m$. Условия эксперимента: $N = 200$ обучающихся, 30 заданий закрытого блока (1–4 категории) и открытые задания (до 7 категорий), 3 преподавателя, девять учебных групп; выполнено 500 повторов Монте-Карло. Инициализация: $a_j = 1$, β_j и пороги – равномерная сетка, $\delta_r = 0$, ковариационная матрица – единичная. Критерий сходимости – относительное изменение логарифма маргинального правдоподобия $|\Delta \ln L| < 10^{-4}$ при максимуме 150 итераций, несошедшиеся запуски исключались из агрегирования, их доля в проведенном эксперименте равна нулю.

Результаты симуляции. Для каждого из 500 повторов вычислялись оценки параметров экспертной строгости δ_r , оценки параметров заданий (a, b), их стандартные ошибки и эмпирическое покрытие 95 %-х доверительных интервалов. Сводные результаты по параметрам экспертной строгости представлены в табл. 1.

Как следует из табл. 1, смещение оценок параметров δ_r не превышает 0,02, а значения среднеквадратичной ошибки (RMSE) близки к эмпирическому стандартному отклонению оценок. Это указывает на отсутствие выраженной систематической ошибки при оценивании экспертной строгости. Эмпирическое покрытие 95 %-ных довери-

тельных интервалов находится в пределах 93,4–94,2 %, то есть в целом соответствует номинальному уровню. Незначительное снижение на 0,8–1,6 процентных пункта можно считать ожидаемым для методов, основанных на асимптотической нормальности при конечном объеме выборки. Оно не свидетельствует о наличии устойчивых проблем в оценивании стандартных ошибок.

Таким образом, при наличии общего автоматически проверяемого компонента K , играющего роль якоря для многомерной калибровки, модель обеспечивает корректную идентифицируемость δ_r даже при полностью разделенных экспертах и устойчива к реалистичной неоднородности групп. В отсутствие якорного компонента схема с непересекающимися экспертами требует перекрестной проверки части работ.

Рассмотрим применение модели на реальных данных на примере зачета по дисциплине «Информационные технологии». Согласно рабочей программе и ФОС, зачет проводится по 100-балльной шкале и включает три компонента: тест на знания (40 баллов), задание на умения (30 баллов) и задание на навыки (30 баллов). Выборка составила 226 студентов 1-го курса 10 учебных групп.

Компонент «Знания» состоит из 30 закрытых заданий (18 с единственным верным ответом; 6 с множественным выбором; 6 на сопоставление) и проверяется автоматически в Moodle.

Таблица 1

Восстановление параметров экспертной строгости

Параметр	Истинное	Среднее	SD	Смещение	RMSE	95 % ДИ
δ_1	+0,30	+0,29	0,09	–0,010	0,091	94,2 %
δ_2	–0,45	–0,43	0,11	+0,020	0,112	93,4 %
δ_3	+0,15	+0,14	0,10	–0,010	0,100	93,8 %
Среднее	–	–	0,10	0,013	0,101	93,8 %

Примечание: составлена авторами на основе полученных данных в ходе исследования.

Таблица 2

Описательная статистика баллов по компонентам зачета

Компонент	Max	M	SD	Min	Max набр.	α
Знания (K)	40	26,1	7,8	4	40	0,76
Умения (U)	30	18,9	8,2	0	30	–
Навыки (N)	30	19,2	8,5	0	30	–
Итого	100	63,2	18,2	11	97	–

Примечание: составлена авторами на основе полученных данных в ходе исследования.

Компоненты «Умения» и «Навыки» представлены по одному практическому заданию в офисном ПО (работа в табличном и текстовом процессорах соответственно), каждое максимумом 30 баллов, проверяется преподавателем по рубрике с семью градациями (0, 5, 10, 15, 20, 25, 30). Десять групп распределены между четырьмя преподавателями неравномерно: $r = 1$ – группы 1–3 ($n_1 = 66$), $r = 2$ – группы 4–6 ($n_2 = 69$), $r = 3$ – группы 7–8 ($n_3 = 46$), $r = 4$ – группы 9–10 ($n_4 = 45$). Средний суммарный балл за зачет составил 63,2 ($SD = 18,2$), медиана 65 баллов. Описательная статистика баллов по компонентам зачета приведена в табл. 2.

Коэффициент альфа Кронбаха для 30 заданий компонента K составил $\alpha = 0,76$, что свидетельствует о приемлемой внутренней согласованности теста. Для компонент U и N , содержащих по одному заданию, расчет α не применим.

Корреляция Пирсона между компонентами составила: $r(K, U) = 0,51$, $r(K, N) = 0,46$, $r(U, N) = 0,63$. Наибольшая связь отмечается между компонентами умений и навыков, а более низкие корреляции с компонентом знаний согласуются с представлением о теоретическом знании и практическом владении как о частично различных конструктах.

Для оценки параметров модели был использован метод MML с 15 узлами квадратуры Гаусса – Эрмита. ЕМ-алгоритм сошелся за 108 итераций (критерий: $|d\ell^*| < 10^{-4}$). Предельное логарифмическое правдоподобие при сходимости составило $\ell^* = -19\,783,9$. Параметры заданий компо-

нента знаний представлены в табл. 3. Средняя дискриминативность заданий составила $\bar{a} = 1,08$ ($SD = 0,41$), что соответствует умеренной дифференцирующей способности. Наибольшую дискриминативность показали задания на сопоставление ($\bar{a} = 1,29$), наименьшую – задания с единственным выбором ($\bar{a} = 0,94$). Средняя трудность варьировалась от $\bar{b} = -1,58$ (наиболее легкое задание) до $\bar{b} = 1,82$ (наиболее трудное). Два задания продемонстрировали низкую дискриминативность ($a < 0,5$), что может свидетельствовать о необходимости переработки таких заданий. Пригодность модели оценивалась по критериям согласованности с данными. В среднем задания демонстрируют приемлемое соответствие IRT-модели, однако диагностика на уровне отдельных заданий выявила два задания с повышенными значениями Outfit и статистически значимым $S - X^2$, что указывает на необходимость их содержательного пересмотра (табл. 4).

На уровне модели в целом получены значения, соответствующие хорошему согласию: $RMSEA = 0,031$ ($< 0,06$), $SRMSR = 0,047$ ($< 0,08$); статистика согласия $M2$ не показала значимого расхождения модельной и эмпирической структур ассоциаций ($p > 0,05$). Анализ согласия на уровне испытуемых не выявил аномальной доли нетипичных индивидуальных профилей. Совокупность этих результатов подтверждает, что предложенная спецификация адекватно описывает структуру наблюдаемых данных по всему диапазону способностей.

Таблица 3

Параметры заданий компонента знаний (GPCM)

Тип задания	Кол-во	m_j	$\bar{a}$ (SD)	$\bar{b}$ (SD)	Infit M	Outfit M
Единств. выбор	18	1	0,94 (0,33)	-0,08 (0,97)	1,03	1,06
Множеств. выбор	6	3	1,14 (0,30)	0,19 (0,72)	0,91	0,97
Сопоставление	6	4	1,29 (0,45)	0,32 (0,61)	0,94	0,96
Все задания	30	—	1,08 (0,41)	0,06 (0,89)	0,99	1,02

Примечание: составлена авторами на основе полученных данных в ходе исследования.

Таблица 4

Значения статистик по отдельным заданиям

№	Компонент	Тип задания	Кат.	a (SE)	b / пороги (SE)	Infit	Outfit	$S - X^2$ (df)	$p / p(BH)$	Решение
21	К	множественный выбор	3	—	—	1,25	1,41	95,0 (20)	$< 0,001 / < 0,001$	пересмотр
27	К	сопоставление	4	—	—	1,20	1,37	66,4 (27)	$0,002 / 0,006$	пересмотр

Примечание: составлена авторами на основе полученных данных в ходе исследования.

Таблица 5

Оценки параметра фасета преподавателя

Препоэд.	δ_r	SE	z	p	p (Holm)
1	+0,33	0,07	4,71	< 0,001	< 0,001
2	-0,39	0,09	-4,33	< 0,001	< 0,001
3	+0,24	0,10	2,40	0,016	0,033
4	-0,18	0,10	-1,80	0,072	0,072

Примечание: составлена авторами на основе полученных данных в ходе исследования.

Параметры заданий компонентов умения и навыки. Дискриминативность задания на умения (работа в табличном процессоре) составила $a_U = 1,74$ (SE = 0,19), задания на навыки (создание составного документа) – $a_N = 1,91$ (SE = 0,22). Высокие значения дискриминативности закономерны: открытые задания с развернутым ответом и широкой шкалой оценивания (7 категорий) обладают значительно большей информативностью, чем отдельные закрытые вопросы.

Параметры строгости при ограничении $\Sigma \delta_r = 0$ представлены в табл. 5: наибольшее положительное значение параметра фасета получено для преподавателя 1, наибольшее отрицательное – для преподавателя 2. Разности $\delta_1 - \delta_2 = 0,72$ ($z = 6,31$, $p < 0,001$) и $\delta_3 - \delta_4 = 0,42$ ($z = 2,97$, $p < 0,01$) статистически значимы. В пересчете на балльную шкалу компонента умений (30 баллов) разница между крайними экспертами эквивалентна приблизительно 4,2 балла. Поскольку каждый преподаватель проверяет только свои группы, выявленные различия могли бы частично отражать различия в подготовке групп; однако общий компонент К (идентичная для всех автоматическая проверка) позволяет калибровать способности на единой шкале, а средние значения $\hat{\theta}_i^K$ по десяти группам близки (от -0,07 до 0,09, SD = 0,05), что согласуется с интерпретацией δ_r как возможных различий строгости при принятых модельных предположениях. Полученные различия следует рассматривать как статистические индикаторы возможных рэйтер-эффектов.

Результаты использования предложенной модели при анализе итогов промежуточной аттестации (зачета) по информационным технологиям позволяют сформулировать следующие выводы. Установленная латентная структура показывает, что связи между выделенными компонентами являются не настолько высокими, чтобы рассматривать их как проявление единого недифференцированного результата. Это означает, что объединение итогов зачета/экзамена в один общий балл способно

скрывать существенные различия в уровне сформированности знаний, умений и навыков обучающегося. Практическое значение имеют различия в дифференцирующей способности заданий различных форматов, представленные в табл. 3. Задания, предполагающие градуированную шкалу ответа, обладают большей информативностью относительно латентной способности испытуемого. Поэтому при ограниченном объеме тестовой части увеличение доли таких заданий может рассматриваться как один из способов повышения точности измерения и более надежного различения уровней подготовленности студентов. Особого внимания заслуживают выявленные статистически значимые различия параметра фасета преподавателя, интерпретируемые как возможные различия строгости оценивания при принятых модельных предположениях, поскольку она непосредственно связана с проблемой справедливости оценивания. В высших учебных заведениях преподаватели, как правило, закрепляются за определенными учебными группами, вследствие чего студент фактически не имеет право выбора, кому сдавать зачеты и экзамены. В такой ситуации итоговая оценка может зависеть не только от уровня подготовки обучающегося, но и от индивидуальной строгости назначенного преподавателя. Предложенная модель позволяет диагностировать подобную статистическую неоднородность на единой латентной шкале. Под корректировкой смещения здесь понимается не изменение первичных баллов и не автоматический пересмотр официальных отметок, а модельный расчет латентной оценки с учетом параметра фасета преподавателя

$$\hat{\theta}_i^{adj} = E(\theta_i | Y_i, \hat{\beta}, \hat{\tau}, \hat{\delta}),$$

где $\hat{\delta}$ – оцененный параметр фасета преподавателя. Главное ограничение исследования состоит в том, что перекрестной проверки открытых работ двумя независимыми преподавателями не проводилось. Поэтому оценки фасета преподавателя нельзя считать окончательно доказанным индивиду-

альным эффектом строгости. Корректнее трактовать их как статистические индикаторы возможных рэйтер-эффектов, полученные при принятых предпосылках модели. В практическом отношении модель следует рассматривать как диагностический инструмент для выявления потенциальных рэйтер-эффектов, требующих дополнительной экспертной проверки. Использование таких результатов для пересмотра оценок обучающихся возможно только при наличии утвержденной процедуры и независимой экспертизы.

Заключение

В настоящей работе предложена многокомпонентная IRT-модель оценивания результатов промежуточной аттестации. Эмпирической основой для формализации послужила нормативно-правовая база и фонды оценочных средств ФГБОУ ВО «Байкальский государственный университет». Отличительной чертой предложенной модели является возможность описания в рамках единой спецификации различных по форме и способу проверки элементов экзаменационной процедуры, включая автоматически оцениваемую тестовую часть и открытые задания, проверяемые экспертом.

Практическая проверка модели была проведена на материалах зачета по дисциплине «Информационные технологии» и показала возможность ее применения в реальных условиях образовательного процесса. Симуляционный эксперимент в контролируемых условиях показал возможность восстановления параметров фасета преподавателя при наличии общего автоматически проверяемого якорного компонента, а эмпирическая апробация выявила статистически значимые различия этих параметров, интерпретируемые как возможные различия строгости оценивания при принятых модельных предпосылках.

Список литературы

1. Samejima F. Estimation of latent ability using a response pattern of graded scores // *Psychometrika*. 1969. Vol. 34. Is. S1. P. 1–97. DOI: 10.1007/BF03372160.
2. Muraki E. A generalized partial credit model: Application of an EM algorithm // *Applied Psychological Measurement*. 1992. Vol. 16. Is. 2. P. 159–176. DOI: 10.1177/014662169201600206.
3. Uto M. A multidimensional generalized many-facet Rasch model for rubric-based performance assessment // *Behaviormetrika*. 2021. Vol. 48. Is. 2. P. 425–457. URL: <https://link.springer.com/article/10.1007/s41237-021-00144-w> (дата обращения: 26.05.2026). DOI: 10.1007/s41237-021-00144-w.
4. Linacre J. M. Many-Facet Rasch Measurement. Chicago: MESA Press, 1989. 216 p. ISBN 0-941938-02-6.
5. Huang H.-Y. Modeling Rating Order Effects Under Item Response Theory Models for Rater-Mediated Assessments // *Applied Psychological Measurement*. 2023. Vol. 47. Is. 4. P. 312–327. URL: <https://journals.sagepub.com/doi/10.1177/01466216231174566> (дата обращения: 26.05.2026). DOI: 10.1177/01466216231174566.
6. Uto M. A Bayesian many-facet Rasch model with Markov modeling for rater severity drift // *Behavior Research Methods*. 2023. Vol. 55. Is. 7. P. 3910–3928. URL: <https://link.springer.com/article/10.3758/s13428-022-01997-z> (дата обращения: 26.05.2026). DOI: 10.3758/s13428-022-01997-z.
7. Miller F., Facke-Fornius E. Parallel Optimal Calibration of Mixed-Format Items for Achievement Tests // *Psychometrika*. 2024. Vol. 89. Is. 3. P. 903–928. URL: <https://link.springer.com/article/10.1007/s11336-024-09968-3> (дата обращения: 26.05.2026). DOI: 10.1007/s11336-024-09968-3.
8. Dai S., Vo T. T., Kehinde O. J., He H., Xue Y., Demir C., Wang X. Performance of Polytomous IRT Models with Rating Scale Data: An Investigation Over Sample Size, Instrument Length, and Missing Data // *Frontiers in Education*. 2021. Vol. 6. Art. 721963. URL: <https://www.frontiersin.org/articles/10.3389/educ.2021.721963/full> (дата обращения: 26.05.2026). DOI: 10.3389/educ.2021.721963.
9. Qiu X.-L., Chiu M. M., Wang W.-C., Chen P.-H. A new item response theory model for rater centrality using a hierarchical rater model approach // *Behavior Research Methods*. 2022. Vol. 54. Is. 4. P. 1854–1868. URL: <https://link.springer.com/article/10.3758/s13428-021-01699-y> (дата обращения: 26.05.2026). DOI: 10.3758/s13428-021-01699-y.
10. Li X., Wang W.-C., Xie Q. Cognitive Diagnostic Models for Rater Effects // *Frontiers in Psychology*. 2020. Vol. 11. Art. 525. URL: <https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00525/full> (дата обращения: 26.05.2026). DOI: 10.3389/fpsyg.2020.00525.
11. Jin K.-Y., Eckes T. Human ratings take time: A hierarchical facets model for the joint analysis of ratings and rating times // *Behavior Research Methods*. 2024. Vol. 56. Is. 4. P. 3535–3547. URL: <https://link.springer.com/article/10.3758/s13428-023-02165-7> (дата обращения: 26.05.2026). DOI: 10.3758/s13428-023-02259-2.
12. Wu Y.-F. IRT Observed-Score Equating for Rater-Mediated Assessments Using a Hierarchical Rater Model // *Journal of Educational Measurement*. 2025. Vol. 62. Is. 1. P. 5–31. URL: <https://onlinelibrary.wiley.com/doi/10.1111/jedm.12425> (дата обращения: 26.05.2026). DOI: 10.1111/jedm.12425.
13. Bock R. D., Aitkin M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm // *Psychometrika*. 1981. Vol. 46. Is. 4. P. 443–459. DOI: 10.1007/BF02293801.
14. Louis T. A. Finding the observed information matrix when using the EM algorithm // *Journal of the Royal Statistical Society. Series B*. 1982. Vol. 44. Is. 2. P. 226–233.
15. Müller M. Item fit statistics for Rasch analysis: can we trust them? // *Journal of Statistical Distributions and Applications*. 2020. Vol. 7. Art. 5. URL: <https://link.springer.com/article/10.1186/s40488-020-00108-7> (дата обращения: 25.06.2026). DOI: 10.1186/s40488-020-00108-7.
16. Ranger J., Brauer K. On the Generalized S-X²-Test of Item Fit: Some Variants, Residuals, and a Graphical Visualization // *Journal of Educational and Behavioral Statistics*. 2022. Vol. 47. Is. 2. P. 202–230. URL: <https://journals.sagepub.com/doi/10.3102/10769986211050304> (дата обращения: 26.05.2026). DOI: 10.3102/10769986211050304.
17. Xu J., Paek I., Xia Y. Investigating the behaviors of M2 and RMSEA2 in fitting a unidimensional model to multidimensional data // *Applied Psychological Measurement*. 2017. Vol. 41. Is. 8. P. 632–644. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC5978478/> (дата обращения: 25.06.2026). DOI: 10.1177/0146621617710464.

Конфликт интересов: Авторы заявляют об отсутствии конфликта интересов.

Conflict of interest: The authors declare that there is no conflict of interest.

Финансирование: Авторы заявляют об отсутствии внешнего финансирования.

Financing: The research was performed without external funding.