СРАВНЕНИЕ МОДЕЛЕЙ ТЕОРИЙ IRT И MIRT ДЛЯ ОЦЕНКИ КОМПЕТЕНЦИЙ СТУДЕНТОВ

COMPARISON OF IRT AND MIRT THEORY MODELS IN GRADUATE COMPETENCES ESTIMATION

Yu. Dyadkin

Summary. The article is dedicated to Item Response Theory (IRT) that is supposed to be an approach to graduate competences estimation. Graduate competences are treated as latent parameters that bring information of what professional skill level a graduate is on. Data of academic achievements are suggested to base on to get this task solved. The author describes pros and cons of different models used within IRT which allows to make conclusions of rational application of MIRT (Multidimensional Item Response Theory) models in represented sphere. Particularly Multidimensional Within-Item Partial Credit Model is suggested to use to implement author approach.

Keywords: competences, competency evaluation, Item Response Theory, IRT, MIRT, MPCM model.

Дядькин Юрий Алексеевич
Аспирант, ФГБО ВО «Байкальский государственный университет»;
Старший преподаватель, ФГБОУ ВО «Иркутский государственный университет» г. Иркутск
dyadkin_ua@inbox.ru

Аннотация. в статье рассматривается применение современной теории тестирования Item Response Theory (IRT) как одной из подходов к оценке латентных параметров — компетенций выпускника вуза на основе данных педагогических измерений. В частности, используются данные итоговой успеваемости. Указаны достоинства и недостатки применения моделей данной теории. Обосновывается переход от использования моделей IRT к моделям теории MIRT (Multidimensional Item Response Theory). Предлагается подход к оцениванию компетенций в рамках теории MIRT с применением модели Multidimensional Within-Item Partial Credit Model.

Ключевые слова: компетенции, оценка компетенций, современная теория тестирования, IRT, MIRT, Multidimensional Within-Item Partial Credit Model.

С введением компетентностного подхода наметился переход в образовании результатов обучения с позиции квалификации к компетентностной. В силу действующего в нынешнем образовании федерального государственного образовательного стандарта (ФГОС), требующего компетентностного подхода к построению образовательного процесса, уровень образованности обучающихся должен оцениваться как количественными, так и качественными показателями. В роли качественных показателей выступает уровень сформированности той или иной компетенции [9].

Профессиональный стандарт по различным направлениям определяет набор компетенций, которыми должен обладать выпускник, т.е. группа профессиональных компетенций (ПК), общепрофессиональных компетенций (ОПК), общекультурных компетенций (ОК) и др. [11].

Под компетенцией понимается обладание каждым индивидом способностью и умением выполнять определенные трудовые функции [8]. В свою очередь под компетентностью понимается интегральная характеристика индивидуума, обладающего множеством компетенций [8].

Таким образом, перед образовательными учреждениями встает задача оценки компетенций студентов на основе обработки результатов существующих педагогических измерений.

Педагогическое измерение — процесс отображения числами интересующих качеств личности [6]. Лорд и Новик определяют педагогическое измерение как такое присвоение чисел, которое верно отражает взаимное расположение испытуемых на числовой шкале в зависимости от их уровня измеряемого качества [2].

Целью педагогического измерения, по мнению В. С. Аванесова, является определение количества интересующего латентного свойства личности (меру интересующего признака), присущего данному испытуемому, а результатом педагогического измерения — некоторая числовая величина, позволяющая установить числовое соотношение между испытуемыми по изучаемому свойству [5].

Компетенции обучающегося формируются путем изучения соответствующих дисциплин, каждая из которых направлена на формирование определенного числа этих компетенций. По каждой дисциплине обучающийся получает первичный балл, который не может быть интерпретирован как итоговая оценка компетенции ввиду того, что этот первичный балл является отражением интегрированной характеристики, содержащую оценку, как правило, не одной компетенции.

Следовательно, напрямую измерить способности обучающегося (скрытые/латентные характеристики) не представляется возможным, их можно измерить косвенным образом, посредством проявления индикаторов.

Для упрощения решения поставленной задачи можно использовать оценки заданий, которые связаны только с одной компетенцией. Однако такой статистикой не обладает большинство учебных заведений. Необходимо предложить подход, при котором можно произвести оценку компетенций, основываясь на итоговых баллах по некоторым дисциплинам.

Для оценивания компетенций необходим инструмент, применимый к оцениванию компетенций.

В силу того, что мы рассматриваем итоговое задание по некоторой дисциплине как средство педагогического измерения, а результатом измерения является оценка, обычно выражаемая в числовом виде, то изначально следует рассмотреть вопрос об измерительных шкалах.

Рассмотрим традиционную шкалу оценивания. Нельзя сказать, что разница уровней знаний обучающихся, получившего оценку «5» и получившего оценку «4», такая же, как для получившего «4» и «3». Исходя из вышесказанного, можем сделать вывод о принадлежности подобных шкал оценивания к порядковому типу. Из этого можем сделать вывод, что нельзя вычислить среднюю оценку и прочие показатели, использующие неопределённые операции для данного типа шкал.

Результаты педагогических измерений необходимо подвергнуть обработке с использованием математических методов для решения важнейшей задачи — получения количественной оценки наблюдаемых педагогических явлений.

Одним из подходов, позволяющих оценить латентные качества индивидуума, является применение современной теории тестирования — IRT (Item Response Theory), предметом которой является оценка вероятности дать правильный ответ испытуемым на задания различной категории сложности [10].

Кроме этого, задействовав математический аппарат IRT, можно осуществить переход от порядковой шкалы к интервальной, оценить латентные черты обучающихся.

Одномерные модели современной теории тестирования (Unidimensional Item Response Theory Models, модели UIRT) — это множество моделей, которые используются для оценки таких параметров как сложность тестового задания и уровень подготовленности респондента. В основе этих моделей лежит идея о том, что процесс выполнения тестовых заданий обучающимися может быть представлен математическим выражением, включающим единственный параметр. Этот параметр описывает некоторую латентную характеристику тестируемого (уровень его подготовленности для решения данного задания). Иными словами, IRT основывается на предположении о том, что существует связь между латентными параметрами испытуемого и наблюдаемой оценкой по итогам тестирования.

Базовым представлением такого рода моделей служит следующее выражение [4]:

$$P(U = u \mid \theta) = f(\theta, \eta, u) \quad (1)$$

где θ — единственный параметр, описывающий характеристику экзаменуемого, η — вектор параметров, описывающий характеристики тестового задания, U — величина, представляющая счет (балл, оценку) за тестовое задание, u — вероятное значение величины U, f — функция, которая описывает связь между параметрами и вероятностью ответа $P(U = u)$.

Данную теорию тестирования также можно применять для оценки компетенций выпускников учебных заведений различного типа.

Тогда параметры модели можно интерпретировать следующим образом: θ — параметр, описывающий компетенцию выпускника; η — вектор параметров, описывающий характеристики дисциплины; U — величина, представляющая итоговый балл за освоение дисциплины. Соответственно, u — вероятное значение величины U, f — функция, которая описывает связь между параметрами и вероятностью получения итогового балла по соответствующей дисциплине $P(U = u)$.

Далее описание моделей будет производиться с учетом данной интерпретации параметров.

Основной интерес представляет θ_i — латентный параметр i-го испытуемого. Необходимой для расчётов переменной является латентный параметр δ_j, характеризующий трудность j-го задания теста. Для того, чтобы оценить результаты работы теста, необходимо на основе наблюдаемых оценок испытуемых сделать вывод о значении латентных параметров θ_i и δ_j.

Решение этой задачи в 1961 г. показал датский статистик Георг Раш, предложив представить соотношение между латентными параметрами в виде разности $\theta - \delta$ при условии, что эти параметры будут оценены в одной шкале [3]. Для этого Г. Раш ввёл интервальную шкалу, использующую в качестве единицы измерения один логит.

Если разность $\theta_i - \delta_j$ составляет 1 логит, то вероятность верного выполнения i-ым испытуемым j-го задания равна 0,73.

Рассмотрим геометрическую интерпретацию описанной выше модели. Абсолютное значение разности $|\theta_i - \delta_j|$ показывает, на каком расстоянии уровень подготовленности θ_i испытуемого находится от уровня сложности задания δ_j. В случае, если эта величина $\theta_i - \delta_j$ имеет большое отрицательное значение, испытуемый не сможет справиться с данным заданием, т.к. имеет слишком низкий уровень подготовленности для этого задания. В случае, когда значение этой величины велико и положительно, это означает, что испытуемый справиться с заданием легко, т.к. уровень его подготовленности очень высок по сравнению со сложностью самого задания. Таким образом, наибольший интерес представляют ситуации, когда абсолютное значение разности приближено к нулю, т.е. измеряемая латентная характеристика близка по значению к уровню трудности задания.

Все модели UIRT основываются на следующих положениях:

- вероятность получения зачета по дисциплине, которое может быть оценено как «зачтено» или «не зачтено», возрастает, если возрастает параметр θ;
- исключается взаимодействие обучающихся друг с другом во время выполнения зачетного задания;
- зачетные задания должны быть независимы друг от друга;
- ответ любого обучающегося зависит исключительно только от единственного параметра θ (компетенции обучающегося) и вектора η (характеристики тестового задания).

Частным случаем UIRT-моделей являются модели, применяемые для заданий с двумя категориями оценки,— «верное» (обозначается единицей), «неверно» (обозначается нулем). Такие модели называются дихотомическими. Верный ответ считается признаком более высокого уровня сформированности изучаемой черты (характеристики) испытуемого, нежели неверный. Дихотомические модели в свою очередь разделяются на классы в зависимости от числа параметров, характеризующих тестовые задания. Наиболее распространены следующие классы подобных моделей: 1) однопараметрические; 2) двухпараметрические; 3) трехпараметрические.

Однопараметрическая модель Г. Раша, известная в зарубежной литературе как «1 Parametric Logistic Latent Trait Model» (1PL) описывает вероятность успеха испытуемого как функцию, зависящую только от одного параметра $\theta_i - \delta_j$. Модель Раша имеет следующий вид [3]:

$$P_j(\theta) = \frac{e^{(\theta-\delta_j)}}{1+e^{(\theta-\delta_j)}}$$

$$P_i(\delta) = \frac{e^{(\theta_i-\delta)}}{1+e^{(\theta_i-\delta)}} \quad (2)$$

где θ и δ независимые переменные для первой и второй функции соответственно.

Второй класс логистических функций представляет двухпараметрическая модель (2PL) А. Бирнбаума. Её математическая запись выглядит так:

$$P_j(\theta) = \frac{e^{a_j(\theta-\delta_j)}}{1+e^{a_j(\theta-\delta_j)}}$$

$$P_i(\delta) = \frac{e^{a_j(\theta_i-\delta)}}{1+e^{a_j(\theta_i-\delta)}} \quad (3)$$

где a_j — параметр, характеризующий дифференцирующую способность j-го задания (item discrimination parametr).

Третий класс представляет трёхпараметрическая модель (3PL) А. Бирнбаума:

$$P_j(\theta) = c_j + (1-c_j)\frac{e^{a_j(\theta-\delta_j)}}{1+e^{a_j(\theta-\delta_j)}}$$

$$P_i(\delta) = c_j + (1-c_j)\frac{e^{a_j(\theta_i-\delta)}}{1+e^{a_j(\theta_i-\delta)}} \quad (4)$$

где c_j — параметр угадывания правильного ответа на j-ое задание.

Однако, описанные выше модели применимы только в случае использования дихотометрической оценки тестирования (верно-неверно). На практике чаще присутствуют промежуточные варианты оценивания. Возникает необходимость дополнения моделей, с целью получения возможности работы с другими вариантами оценок. Данная проблема может быть решена с помощью моделей, применимых к шкалам множественных категорий. К таким моделям относятся:

- модель частичного оценивания или Partial Credit Model (PCM).
- модель рейтингового оценивания или Ratings Scale Model (RSM).

В PCM-модели функция вероятности получения оценки k (k ∈ (0, …, N), где N — число уровней в шкале оценки) i-м испытуемым за j-е задание имеет следующий вид:

$$P(x_{ij} = k|\theta_i) = \frac{e^{\sum_{u=0}^{k}(\theta_i-\delta_{ju})}}{\sum_{v=0}^{m_j} e^{\sum_{u=0}^{v}(\theta_i-\delta_u)}} \quad (5)$$

Таблица 1. Формирование матрицы W по дисциплине i

Компетенция Категория	Θ_1	Θ_2	Θ_3	Θ_4	...	Θ_l
k=0	1	0	0	1		1
k=1	1	0	0	1		1
k=2	1	0	0	1		1

где k — оценка за j-е задание, m_j — максимальная полученная оценка за j-ое задание, δ_{ju} — пороговый параметр, который определяет сложность достижения u-го пункта шкалы.

Модель PCM «работает» с заданиями, которые имеют возрастающую сложность. Подразумевается, что более сложная часть соответствует более высокой оценке [7].

В RSM-модели напротив, предполагается, что каждый u-й пункт достижим с одинаковой сложностью. В такой модели вероятность получения оценки k i-м испытуемым за j-ое задание задаётся формулой:

$$P(x_{ij} = k|\theta_i) = \frac{e^{\sum_{u=1}^{k}(\theta_i-(\delta_j+\tau_u))}}{\sum_{v=1}^{k} e^{\sum_{u=0}^{v}(\theta_i-(\delta_j+\tau_u))}} \quad (6)$$

где δ_j — параметр трудности j-го задания, τ_u — параметр, определяющий сложность достижения каждого u-го пункта шкалы (общий для всех заданий) [1].

Для оценки параметров θ_i и δ_j можно применить классический метод максимального правдоподобия.

Существуют и более сложные модели, и их вариации, однако все модели теории IRT имеют ряд ограничений, при которых их можно использовать, а именно:

- наличие латентных параметров, недоступных для непосредственного наблюдения;
- наличие индикаторов, связанных с латентными параметрами, которые можно наблюдать;
- латентный параметр обязательно должен быть одномерным, т.е. можно оценить только одну переменную.

Обращая внимание на последнее ограничение можно сказать, что данные модели позволяют оценивать только один параметр, в нашем случае только одну компетенцию при выполнении итогового задания по дисциплине. Однако на практике это не так.

Как уже было отмечено выше, за формирование одной компетенции отвечает множество дисциплин. А при выполнении итогового задания по той или иной дисциплине, обучающемуся необходимо проявлять не одну, а зачастую, множество компетенций, которые должны быть сформированы при изучении данной дисциплины.

Именно этим и обусловлен переход от использования одномерных моделей теории IRT к многомерным моделям теории MIRT. Где под многомерностью понимается снятие ограничения моделей IRT, связанного с тем, что можно сразу оценить несколько латентных переменных, что подразумевает оценку нескольких компетенций при выполнении одного итогового (экзаменационного/зачетного) задания.

С учетом рассмотренных выше моделей IRT целесообразно выбрать модель теории MIRT, учитывающую следующие параметры:

- задания могут быть различной сложности;
- задания связаны с несколькими компетенциями;
- шкала оценивания должна быть политомической;
- модель должна иметь параметры, связывающие дисциплины с матрицей компетенций учебного плана.

В соответствии с этими требованиями для оценки компетенций была выбрана модель Multidimensional Within-Item Partial Credit Model [4], которая имеет вид:

$$P(u_{ij} = k\,|\theta_j, b_i) = \frac{e^{\sum_{l=1}^{m}(\theta_{jl}-b_{ik})W_{ilk}}}{\sum_{r=0}^{K_i} e^{\sum_{l=1}^{m}(\theta_{jl}-b_{ir})W_{ilr}}} \quad (7)$$

где параметры модели можно интерпретировать следующим образом:

- u_{ij} — категория, которую достиг студент j по заданию i;
- k — первичный бал, полученный студентом;
- θ_i — вектор оценок компетенций студента j;
- b_i — оценка сложности задания i;
- m — количество оцениваемых компетенций с помощью сформированного набора заданий;
- θ_{jl} — оценка компетенции l студента j;
- b_{ilk} — сложность достижения категории k задания i в рамках компетенции l;
- W_{ilk} — матрица, характеризующая факт оценивания категории k задания i в рамках компетенции l.

Матрица W имеет непосредственную связь с матрицей компетенций образовательной программы и формируется для каждой дисциплины в отдельности. Например, если компетенция формируется в рамках данной дисциплины i, то в столбце с этой компетенцией l проставляются единицы, иначе — нули.

Таким образом, можно заключить, что для оценки компетенций предпочтительней использование моделей теории MIRT. В частности, модель Multidimensional Within-Item Partial Credit Model является наиболее «близкой» по наличию параметров и их интерпретации к задаче по оценке компетенций выпускника вуза на основе данных педагогических измерений. Данная модель позволяет:

- установить взаимосвязь между учебным планом и матрицей компетенций;
- производить оценку компетенций, принимая во внимание тот факт, что разные баллы, полученные обучающимся, за итоговое задание предполагают различный уровень сформированности компетенций;
- учитывать сложность получения той или иной оценки.

ЛИТЕРАТУРА

1. Daniel C. Furr. Rating scale and generalized rating scale models with latent regression.— 2017.; URL: http://mc-stan.org/users/documentation/case-studies/rsm_and_grsm.html#rating-scale-model-with-latent-regression (дата обращения: 15.04.2019).
2. Lord, F.M., Nivick, M. R. Statistical Theories of Mental Test Scores.— New York: Addison-Wesley Publ. Co., 2008.— 592 с.
3. Rasch, G. Probabilistic models for some intelligence and attainment tests.— Chicago: The University of Chicago Press, 1980.— 199 с.
4. Reckase, M. D. Multidimensional Item Response Theory.— London: Springer, 2009.— 364 с.
5. Аванесов, В. С. Основные направления развития педагогических измерений // Школьные технологии.— 2012.— № 1.— С. 157–174.
6. Аванесов, В. С. Педагогическое измерение латентных качеств // Педагогические измерения.— М.: Издательский дом «Народное образование», 2003.— С. 12–16.
7. Братищенко, В.В., Кешиков, К. А. Модель с латентными параметрами для оценивания компетенций студентов по данным текущей успеваемости // Известия Байкальского государственного университета.— 2016.— № Т. 26, № 5.— С. 811–817.
8. Маркова, А. К. Психология труда учителя: книга для учителя.— М.: Просвещение, 1993.— 192 с.
9. Михайленко Т. С. Компетентностный подход в оценивании качества результатов обучения студентов // Научно-методический электронный журнал Концепт.— 2014.— № S22.— С. 51–55.
10. Родионов, А. В. Разработка моделей, методов и программного обеспечения для оценки компетенций учащихся ВУЗов: дис. . . . канд.т.н. наук: 05.13.01.— Иркутск, 2015.— 228 с.
11. Федеральные государственные образовательные стандарты // ФГОС URL: https://fgos.ru/ (дата обращения: 15.04.2019).

Журнал «Современная наука: актуальные проблемы теории и практики»