УДК 004.89

А.Ю. СОКОЛОВ, О.Г. МОЛЧАНОВА

Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Украина

МЕТОДЫ ОЦЕНИВАНИЯ РЕЗУЛЬТАТОВ ТЕСТИРОВАНИЯ В АВТОМАТИЗИРОВАННЫХ СИСТЕМАХ ОБУЧЕНИЯ

Целью работы является исследование различных методов оценивания, применяемых в современных системах тестирования. В работе рассматриваются методы, основанные на классической теории тестирования и теории IRT (item response theory). Было показано, что в результате применения классической теории тестирования и современной теории IRT можно улучшить показатели сопоставимости результатов тестирования. Предложены подходы к улучшению качества оценивания уровней подготовленности испытуемых как функции разнообразия оценочных баллов. В качестве программных средств использованы пакеты статистической обработки SPSS, анализа тестов WinSteps, пакет MatLab.

Ключевые слова: *классическая теория тестов, сложность задания, способность, латентные параметры, логит.*

Введение

Развитие информационных технологий привнесло значительный вклад в методы обучения и оценивания спудеев (это понятие характеризует учащихся всех категорий) [1]. Появились обучающие программы и интерактивные тесты, различные мониторинговые исследования, государственные программы по автоматизированному оцениванию знаний. Тестирование активно используется в дистанционном обучении, при внедрении Болонской системы для самостоятельного обучения студентов. Сфера применения автоматизированного тестирования расширяется и на производство, где управление персоналом трансформируется в непрерывный процесс повышения квалификации (разумеется, с последующим тестированием и оцениванием обучаемых). Особенностью таких систем является то, что в процессе обучения и оценивания миссия преподавателя значительно сужается, и результаты оцениваются автоматически. Это обусловлено как необходимостью одновременного оценивания большого количества испытуемых, так и самой идеологией автоматизированных средств обучения – самостоятельное обучение и независимое оценивание.

Одной из важных задач является сопоставимость результатов различных тестов, ранжирование спудеев по уровню знаний, формирование итоговых оценок по наборам тестов.

Использование так называемых «сырых» баллов, то есть суммарных оценок за успешное выполнение заданий, получаемых в результате выполнения теста, возможно весьма в ограниченном объеме (если тестирование ограничено только выявлением уровня знаний по конкретной теме и не интегрируется с другими результатами).

Эффективность тестовых оценок зависит не только от качества теста, но и от методов сравнения и интерпретации первичных («сырых») баллов испытуемой группы [2]. Поэтому важным представляется анализ известных методов сравнения и интеграции оценок различных тестов, а также исследования качества оценивания группы спудеев, понимая под *качеством методики оценивания разнообразие оценочных баллов*. Именно этим вопросам посвящена настоящая статья.

1. Оценивание в классической теории тестирования

В основе классической теории тестирования лежит идея преобразования «сырых» баллов в единую шкалу на основании анализа типа исходной информации. В соответствии с типологией педагогических измерений, виды шкал удобно представить в виде иерархии, предложенной С. Стивенсом [2] (рис.1).

В зависимости от целей анализа тестирования и характеристик исходных данных применяют ту или иную методику оценивания. В основном различают два типа оценивания – ранговое и критериальное.

В задаче ранжирования необходимо расположить спудеев, не делая вывод о том, насколько один лучше другого. Для этой цели вполне достаточно только качественных признаков. Критериальное же оценивание, в основе которого лежит сравнимость результатов, возможно только при использовании количественных шкал.

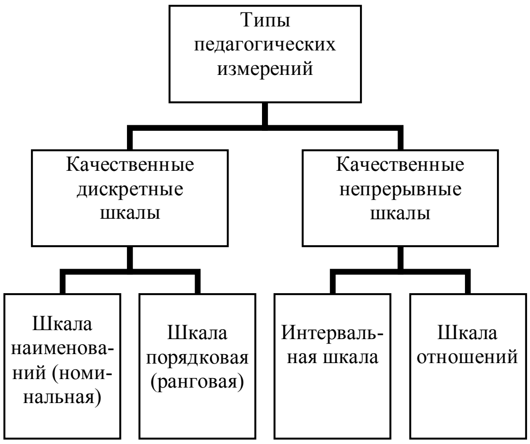

Рис. 1. Типология шкал по С. Стивенсу [2]

Несмотря на длительный (почти столетний) период использования, классическая теория тестирования и рекомендованные в рамках теории линейные преобразования «сырых» баллов повышают сопоставимость результатов спудеев, однако не меняют природу порядковой шкалы. К таким преобразованиям относятся шкалы, рассмотренные далее.

В первую очередь это Z-шкала, в основе которой лежит преобразование «сырого» балла r

$$z = \frac{r - M}{\sigma},$$

где M и σ - математическое ожидание и среднеквадратическое отклонение генеральной совокупности соответственно. Эта шкала имеет ряд преимуществ, среди которых возможность сопоставимости отклонений в «сильной» и «слабой» группах.

На основе Z-шкалы построено большое количество других линейных преобразований – IQ - $(100 + 15z)$, CEEB - $(500 + 100z)$, Векслера - $(10 + 3z)$, Т-шкала - $(50 + 10z)$. Каждая из этих шкал используется в практических педагогических измерениях и отличается особенностями интерпретации.

Наилучшую сопоставимость результатов в рамках классической теории предлагает группа методов, основанных на процентильных преобразованиях «сырых» баллов. Так, для установления относительного положения спудея в группе испытуемых можно установить его ранг в процентилях – процентной доле испытуемых из группы, результаты которых не выше первичного показателя данного спудея. Несмотря на очевидное достоинство - возможность на ранговой шкале установить относительное положение спудея, сравнение разных выборок (то есть разных тестов, сессий по одному предмету и т.д.) затруднено тем, что процентильное распределение тесно связано с частотным распределением той выборки, на которой оно получено [3]. Поэтому даже процентильные балы трудно сравнивать между собой, если они получены по разным выборкам. Очевидное решение этой проблемы – стандартизация выборок, то есть приведение всех баллов к единой шкале. Для этого все эмпирические плотности распределения частот «сырых» баллов преобразуются к одному и тому же «эталонному» распределению – нормальному распределению с заданными математическим ожиданием и дисперсией. Обычно – это центрированное и нормированное нормальное распределение. Этот метод называется методом эквипроцентильной нормализации (ЭПН) и используется во многих странах для оценивания знаний. К примеру, на Украине метод предполагает перевод «сырых» баллов в шкалу [100,200], для чего используется эталонное распределение N(150,15).

Таким образом, в качестве основного метода в сравнительном анализе мы будем использовать метод ЭПН.

2. Оценивание в современной теории тестирования

Современная теория тестирования (IRT) базируется на исследовании взаимосвязи трудности заданий, подготовленности испытуемых и вероятности правильного ответа. Основная модель, отражающая эту связь, называется моделью Г. Раша [3].

Успех в решении задания имеет вероятностный характер. Пусть вероятность того, что участник тестирования верно решит задание,(вероятность успеха) определяется как функция уровня подготовленности участника a и уровня трудности задания d:

$$p = p(a, d) = \frac{a}{a + d} = \frac{a/d}{1 + a/d} = \frac{\lambda}{1 + \lambda}. \qquad (1)$$

Символ λ означает отношение латентных параметров подготовленности и трудности.

Формула (1) – это и есть модель Раша, в соответствии с которой вероятность успеха зависит не от каждого аргумента, а от их отношения. Исследуем некоторые свойства данной функции.

Единица измерения подготовленности и трудности – одинакова. Если одному заданию приписать единичную трудность $d_0 = 1$ (то же можно сделать для подготовленности $a_0 = 1$), тогда трудности всех заданий можно сравнивать по сложности с единичным, так же можно сравнивать и подготовленность (если трудность d задания меньше 1, оно в d раз легче единичного, если больше 1, то соответственно – в d раз труднее).

Очевидно, что $d, a, \lambda \in [0, \infty), p \in [0,1]$.

При $\lambda \to 0, \frac{a}{d} \to 0, p \to 0$, то есть абсолютно неподготовленный участник тестирования никогда не выполнит задание.

При $\lambda \to \infty,\ \frac{a}{d} \to \infty,\ p \to 1$ участник тестирования, уровень подготовленности которого во много раз превышает трудность задания, гарантированно его сдаст успешно.

Аргументы функции (1) нельзя измерить непосредственно, а вот значение функции, то есть вероятность, доступна для измерения по результатам тестирования. В этом и заключается основная идея теории IRT – по известной вероятности оценить сложность заданий и уровень подготовленности спудеев. Исходя из вида функции (1) очевидно, что данная задача является некорректной.

Обратная функция позволит определить только параметр λ по измеренному значению p, то есть

$$\lambda = \frac{p}{1-p} = \frac{p}{q} \tag{2}$$

и найти лишь отношение латентных параметров подготовленности и трудности.

Если же мы имеем эталонное задание с единичным значением трудности, тогда можно определить и соответствующее значение подготовленности, разместив его на шкале измерений. Это – вторая плодотворная идея IRT, позволяющая решать задачу стандартизации различных выборок.

2.1. Логистическая функция Раша

На практике удобно выражать аргументы уровня подготовленности и трудности не в линейном, а в логарифмическом масштабе:

$$\ln a = \theta, \ln d = \delta \Leftrightarrow a = e^{\theta}, d = e^{\delta}. \tag{3}$$

Тогда функция успеха примет вид:

$$p = \frac{e^{\theta}}{e^{\theta}+e^{\delta}} = \frac{1}{1+e^{-(\theta-\delta)}}. \tag{4}$$

Формула (4) называется основной логистической моделью Раша. График функции (4) приведен на рис. 2.

Аргументы уровня подготовленности и трудности $\theta, \delta \in (-\infty, \infty)$ измеряются на одной шкале с единицей измерения **1 логит**. Очевидно, что

$$\frac{1}{1+e^{-1}} = 0,731$$

то есть разница в 1 логит повышает вероятность успеха в 0,731 раза.

2.2. Оценка латентных параметров на основе «сырых» баллов

Рассмотрим схему тестирования набором из k заданий.

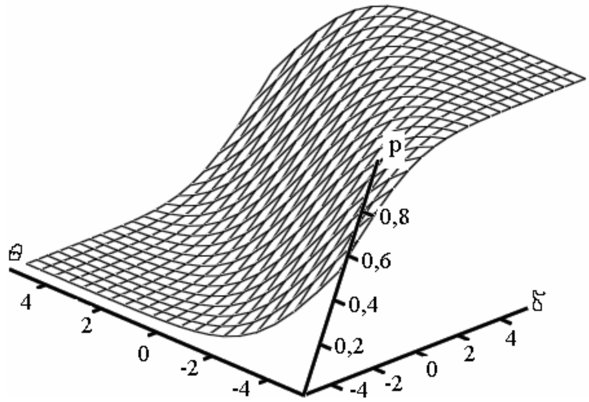

Рис. 2. Логистическая функция Раша

Пусть в тестировании принимают участие n человек. Результат выполнения каждого задания оценивается по дихотомному принципу.

Обозначим $R = (r_{ij})$ матрицу ответов ($i = 1,\dots,n; j = 1,\dots,k$).

Элементы r_{ij} являются случайными величинами: они принимают значения 1 с вероятностью $p_{ij} = p(\theta_i, \delta_j)$.

Вычислим «сырые» баллы участников и заданий:

$$b_i = \sum_{j=1}^{k} r_{ij}, i = 1,\dots,n; \tag{5}$$

$$c_j = \sum_{i=1}^{n} r_{ij}, j = 1,\dots,k. \tag{6}$$

Очевидно, что множество значений $b_i \in \{0,1,\dots,k\}$, то есть все количество участников можно разделить строго на $k+1$ группу по значению «сырого» балла.

Перепишем снова выражения (2) и (4)

$$\lambda = \frac{p}{1-p} = \frac{p}{q},\ p = \frac{e^{\theta}}{e^{\theta}+e^{\delta}}, q = \frac{e^{\delta}}{e^{\theta}+e^{\delta}}.$$

Тогда

$$\lambda = e^{\theta-\delta}. \tag{7}$$

Напомним, что символ λ означает отношение латентных параметров подготовленности и трудности в линейной шкале. Логарифм величины (7) дает непосредственную разницу латентных параметров уровня подготовленности и трудности θ, δ в шкале логитов. Обозначим эту величину так

$$l = \ln \lambda = \theta - \delta. \tag{8}$$

Очевидно, что для каждого участника и каждого задания можно определить величину

$$l_{ij} = \theta_i - \delta_j,\ i = 1,\dots,n; j = 1,\dots,k \tag{9}$$

если соответствующие вероятности не равны 0 либо 1.

Причем величину l_{ij} можно измерять на основе «сырых» баллов. Неизвестными величинами в системе уравнений (9) являются параметры уровня подготовленности и трудности θ_i, δ_j.

Но, поскольку участники, набравшие один и тот же «сырой» балл, имеют одно и тоже значение уровня подготовленности (в соответствии с моделью Раша!), количество уравнений сокращается с $n \times k$ до $(k+1) \times k$, что при $n \gg k$ дает существенную редукцию системы уравнений.

Обозначим через n_b количество участников тестирования, набравших одинаковый «сырой» балл $b = 0,1,\ldots,k$. Пусть $\Theta(b)$ обозначает уровень подготовленности соответствующей группы. Тогда систему уравнений (9) можно переписать так:

$$l_j(b) = \Theta(b) - \delta_j, \qquad (10)$$

$j = 1,...,k;\ b = 1,\ldots,k-1.$

Случаи $b = 0, b = k$ означают, что либо испытуемый не выполнил ни одного задания, либо выполнил все k. Для этих групп расчет значений $l_j(b)$ должен выполняться по особой процедуре.

Для всех же остальных величин $l_j(b)$ их можно определить так:

$$l_j(b) = \ln\frac{p_j(b)}{q_j(b)}, \qquad (11)$$

где $p_j(b)$ – относительная частота верного решения j-го задания участниками, набравшими один и тот же «сырой» балл.

С учетом, например, $l_j(0) = -5,\ l_j(k) = 5$, а также возможных аналогичных случаев для других групп b, в которых возможны значения $p_j(b) = 0$, $p_j(b) = 1$, мы имеем несовместную систему $k \times (k+1)$ уравнений вида (10), содержащую $2k+1$ неизвестных. При этом матрица коэффициентов вырождена, и ее ранг равен $2k$, то есть количество независимых уравнений на 1 меньше, чем количество неизвестных параметров. Поэтому одно из значений необходимо полагать свободным, выразив через него остальные неизвестные. Это значение задает начало отсчета на шкале. Удобным является совместить начало отсчета со средним значением Θ параметра θ в логитах.

Систему уравнений (10) можно решить несколькими способами – на основе системы нормальных уравнений, методом моментов, либо методом максимального правдоподобия. Данные методы подробно описаны в [3] и применяются в пакете WinSteps для оценки латентных параметров.

Как было отмечено ранее, что не смотря на заявленную в названии теории индивидуальную связь сложности задания с уровнем подготовленности участника, на практике мы имеем такое же количество подгрупп спудеев, набравших одинаковый балл в шкале логитов, что и в системе исходных «сырых» баллов. То есть, каноническая теория IRT не улучшает качество оценивания как функцию разнообразия исходных баллов, хотя остальные ее преимущества очевидны.

Таким образом, актуальным представляется совершенствование теории в направлении повышения разнообразия количества участников тестирования с разными баллами. Решение этой задачи можно получить лишь в том случае, когда участники, набравшие одинаковый балл (пусть даже и в логитах), затем перераспределяются внутри подгруппы в соответствии со сложностями заданий, например, в лексикографическом порядке, либо на основе пересчета уровней подготовленности при зафиксированных значениях сложностей задания.

Именно эти модификации канонической теории IRT предлагаются в настоящей работе.

3. Улучшение качества оценивания в современной теории тестирования

Пусть в результате решения системы уравнений (10) получены значения уровней подготовленности в группах, $b = 0,1,\ldots,k$ и уровни сложности заданий δ_j, $j = 1,\ldots,k$.

Рассмотрим группу с одинаковым уровнем подготовленности

$$\Theta(b^*) = \theta_{n_1} = \ldots = \theta_{n_{k^{\delta^*}}},$$

где $n_1,\ldots,n_{\kappa^{\delta^*}}$ номера участников в данной группе.

Для этих участников количество заданий, выполненных правильно, одинаково (их «сырые» баллы одинаковы, то есть для дихотомических тестов одинаково количество правильных ответов, то есть заданий). Но сами задания могут быть различными.

Выполним сортировку группы $\left\{n_1,\ldots,n_{k^{\delta^*}}\right\}$ в лексикографическом порядке, то есть первое место в группе займет участник, выполнивший самое трудное задание с максимальным уровнем сложности в данной группе.

Если таких участников несколько, для них применяется та же процедура, только для второго по сложности задания, и так далее. Таким образом, все участники внутри группы будут проранжированы с учетом сложности заданий.

Такое решение приемлемо для ранжирования в рамках одного теста. Если же необходимо изменить количественные значения показателей уровня подготовленности спудеев, предлагается итерационная процедура, основанная на уже полученных параметрах модели Раша для данного теста и на изменении показателей уровня подготовленности при фиксированных значениях сложности задания.

Наряду с однопараметрической моделью Раша (4), в практике тестирования большее распространение получили двухпараметрические модели, имеющие вид:

$$p(\theta) == \frac{1}{1+e^{-a(\theta-\delta)}} \qquad (12)$$

в которых вероятность успеха спудея с уровнем подготовки θ определяется как трудностью задания δ, так и дискриминационным свойством задания a – коэффициентом дискриминации.

Предлагается для всех участников группы с одинаковым уровнем подготовленности $\Theta\left(b^*\right)$ провести итерационную процедуру уточнения подготовленности по формуле, которую обычно используют в методе моментов совместно с формулой уточнения сложности заданий:

$$\hat{\theta}_{j,s+1} = \hat{\theta}_{j,s} + \frac{\sum_{i=1}^{k} a_i \left(r_{j,i} - p\left(\hat{\theta}_{j,s}\right)\right)}{\sum_{i=1}^{k} a_i^2 \left(p\left(\hat{\theta}_{j,s}\right)\left(1-p\left(\hat{\theta}_{j,s}\right)\right)\right)}, \qquad (13)$$

$j = n_1, \ldots, n_{k_{\delta^*}}$, $s = 0,1,2,\ldots$.

Данная итерационная процедура выполняется для всех элементов множества

$$\Theta\left(b^*\right) = \left\{\theta_{n_1}, \ldots, \theta_{n_{k_{\delta^*}}}\right\},$$

элементы которого вначале итерационной процедуры одинаковы, то есть

$$\theta_{n_1,0} = \ldots = \theta_{n_{k_{\delta^*}},0} = \Theta\left(b^*\right).$$

4. Пример

Рассмотрим пример анализа результатов тестирования группы из 13 человек ($n = 13$) с помощью теста, состоящего из трех заданий ($k = 3$). Матрица ответов $R = \left(r_{ij}\right)$, $i = 1,\ldots,n;\ j = 1,\ldots,k$, приведена в табл. 1. Данный тест разделил группы на две категории – с одним и двумя значениями суммарного «сырого» балла -1 и 2.

Рассмотрим применение классического метода ЭПН.

Частотный анализ результатов представлен в табл. 2. Также в таблице приведен перевод в шкалу 100-200 методом ЭПН.

Результат построения двухпараметрической модели Раша (12) и анализа теста с помощью IRT представлен в табл. 3.

Таблица 1

Матрица ответов

№ участника	Задание 1	Задание 2	Задание 3	Суммарный балл
1	0	1	1	**2**
2	1	0	1	**2**
3	1	1	0	**2**
4	1	0	0	**1**
5	1	0	0	**1**
6	1	0	0	**1**
7	1	0	0	**1**
8	1	0	0	**1**
9	0	1	0	**1**
10	0	1	0	**1**
11	0	1	0	**1**
12	0	0	1	**1**
13	0	0	1	**1**

Таблица 2

Частотный анализ результатов

Балл	Частота	Кумул. частота	Процентиль	ЭПН
0	0	0	0%	100
1	10	76,9%	38%	145
2	3	100%	88%	167
3	0	100%	100%	200

Если приравнять 3 логита 200 баллам, линейное преобразованием шкалы логитов в шкалу 100-200 результаты в логитах можно реализовать то по формуле

$$150 + 16,667 \cdot \log it \qquad (14)$$

получим результирующие значения – 162 и 138 баллов соответственно.

Однако, как и в случае ЭПН применение IRT не изменяет количества участников тестирования с одинаковым баллом. Из рассматриваемого примера видно, что уровень сложности заданий различный, и очевидно, что этот фактор следует участь при ранжировании участников с одинаковыми баллами.

Применим предложенные методы улучшения качества оценивания. При лексикографическом упорядочивании участников, набравших по 2 «сырых» балла, нетрудно видеть, что правильным будет следующий порядок: $1 \succ 2 \succ 3$, поскольку участник с №1 решил самое трудное задание в 0,47 логитов и следующее за ним по трудности в 0,1 логита, участник с № 2 – более простую комбинацию.

Таблица 3

Результат построения двухпараметрической модели Раша

№ участника	Задание 1	Задание 2	Задание 3	Суммарный балл	Уровень подготовленности (логит)
1	0	1	1	2	**0,72**
2	1	0	1	2	**0,72**
3	1	1	0	2	**0,72**
4	1	0	0	1	**-0,72**
5	1	0	0	1	**-0,72**
6	1	0	0	1	**-0,72**
7	1	0	0	1	**-0,72**
8	1	0	0	1	**-0,72**
9	0	1	0	1	**-0,72**
10	0	1	0	1	**-0,72**
11	0	1	0	1	**-0,72**
12	0	0	1	1	**-0,72**
13	0	0	1	1	**-0,72**
	-0,57	**0,1**	**0,47**	Сложность задания (логит)	

Что касается участников, набравших «сырой» балл 1, то порядок будет такой:

$$13 \equiv 12 \succ 11 \equiv 10 \equiv 9 \succ 8 \equiv 7 \equiv 6 \equiv 5 \equiv 4.$$

Данный подход ранжирует участников внутри групп, однако не добавляет количественной информации в результирующие баллы.

Рассмотрим применение итерационной процедуры (13). В результате построения модели Раша (12) в пакете WInSteps были получены следующие значения коэффициентов дискриминации заданий

$$a_1 = 0,43; \; a_2 = 1,04; \; a_3 = 1,18.$$

Начальные значения уровней подготовленности – это последний столбец табл. 3.

Результаты итерационной процедуры (13), начиная со 2 шага, приведены в табл. 4.

Таблица 4

Итерационная процедура уточнения подготовленности

№ уч.	Итерация 2	Итерация 3	Итерация 4	Итерация 5
1	1,6575	1,8891	1,9096	1,9097
2	0,6856	0,6857	0,6857	0,6857
3	0,4625	0,4683	0,4683	0,4683
4	-1,1946	-1,5062	-1,5442	-1,5448
5	-1,1946	-1,5062	-1,5442	-1,5448
6	-1,1946	-1,5062	-1,5442	-1,5448
7	-1,1946	-1,5062	-1,5442	-1,5448
8	-1,1946	-1,5062	-1,5442	-1,5448
9	-0,2226	-0,1996	-0,1997	-0,1997
10	-0,2226	-0,1996	-0,1997	-0,1997
11	-0,2226	-0,1996	-0,1997	-0,1997
12	0,0004	0,0240	0,0240	0,0240
13	0,0004	0,0240	0,0240	0,0240

Листинг программы расчетов в пакете Matab для одного тестируемого приведен на рис. 3.

Воспользовавшись формулой (14), представим все полученные результаты в единой табл. 5.

```
Persons=1;Items=3;
R=[1 1 0];% Change this raw for each
D1new=[0,72];                   %case
D1=D1new;
D=[];
for k=1:10 %Number of iterations
    D1=D1new;
D=[D,D1];

a=[0.43 1.04 1.18];
b=[-0.57 0.1 0.47];
for j=1:Persons
nn=0;
for i=1:Items
    p=1/(1+exp(-a(i)*(D1(j)-b(i))));
    nn=nn+a(i)*(R(j,i)-p);
end;
dn=0;
for i=1:Items
    p=1/(1+exp(-a(i)*(D1(j)-b(i))));
    dn=dn+a(i)*a(i)*p*(1-p);
end;
dDj=nn/dn;
D1new(j)=D1(j)+dDj;
end;
end;
```

Рис. 3. Листинг программы итерационной процедуры (13)

Таблица 5

Перевод в шкалу 100-200

№ участника	Суммарный балл	ЭПН	IRT	Модификация IRT
1	**2**	167	162	182
2	**2**	167	162	161
3	**2**	167	162	158
4	**1**	145	138	124
5	**1**	145	138	124
6	**1**	145	138	124
7	**1**	145	138	124
8	**1**	145	138	124
9	**1**	145	138	146
10	**1**	145	138	146
11	**1**	145	138	146
12	**1**	145	138	150
13	**1**	145	138	150

Следует отметить, что после итерационной модификации баллов участники группы с более низким начальным баллом не достигли уровня баллов более высокой группы, то есть разделение внутри одной группы не пересекается с другими группами, что не противоречит понятию справедливости оценивания.

Заключение

В результате применения классической теории тестирования и современной теории IRT можно улучшить показатели сопоставимости результатов тестирования.

Однако наилучшего результата можно достичь путем модификации современной теории тестирования за счет ранжирования внутри групп с одинаковым баллом на основе учета сложностей заданий.

Литература

1. Самылкина Н.Н. Современные средства оценивания результатов обучения / Н.Н. Самылкина. – М.: Бином, Лаборатория знаний, 2007. – 172 с.

2. Челышкова М.Б. Теория и практика конструирования педагогических тестов: Учебное пособие / М.Б. Челышкова. – М.: Логос, 2002. – 432 с.

3. Нейман Ю.М. Введение в теорию моделирования и параметризации педагогических тестов / Ю.М. Нейман, В.А. Хлебников. – М.: Москва, 2000. – 168 с.

4. Frank B. Baker. The basics of item response theory / Baker Frank B. – Univ. of Wisconsin, 2001. – 185 p.

Поступила в редакцию 14.01.2011

Рецензент: д-р техн. наук, проф., зав. каф. авиационных приборов и измерений Н.Д. Кошевой, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков, Украина.

МЕТОДИ ОЦІНЮВАННЯ РЕЗУЛЬТАТІВ ТЕСТУВАННЯ В АВТОМАТИЗОВАНИХ СИСТЕМАХ НАВЧАННЯ

О.Ю. Соколов, О.Г. Молчанова

Метою роботи є дослідження різних методів оцінювання, які використовуються в сучасних системах тестування. В роботі розглянуто методи, що ґрунтуються на класичній теорії тестування та теорії IRT (Item Response Theory). Було показано, що в результаті застосування класичної теорії тестування та сучасної теорії IRT можливо покращити показники зрівняння результатів тестування. Запропоновано підходи к покращенню якості оцінювання рівнів підготовки студентів як функції різноманітності балів. У якості програмних засобів використано пакети статистичної обробки SPSS, аналізу тестів WinSteps, пакет MatLab.

Ключові слова: класична теорія тестів, складність завдання, уміння, латентні параметри, логіт.

MODEL OF EDUCATIONAL ASSESSMENT IN AUTOMATIZED LEARNING SYSTEMS

O.U. Sokolov, O.G. Molchanova

The purpose of work is to investigate of different methods of assessment which are using in modern systems of testing. In the article methods which based on classical test theory and modern theory of IRT are discussing. It was shown that the application of classical test theory and modern theory of IRT can improve the performance of comparability of test results. New approaches are proposed for increasing of quality of assessment the levels of efficiency of students like function of quality is variety of marks. As program instruments we used packages like SPSS, WinSteps and MatLab.

Key words: classical test theory, difficulty of item, ability, latent parameters, logit.

Соколов Александр Юрьевич – д-р техн. наук, проф., заведующий кафедрой информатики Национального аэрокосмического университета им, Н.Е. Жуковского «ХАИ», Харьков, Украина.

Молчанова Ольга Георгиевна – аспирантка кафедры информатики Национального аэрокосмического университета им, Н.Е. Жуковского «ХАИ», Харьков, Украина.