МЕТОДИКА ОЦЕНКИ КАЧЕСТВА ТЕСТОВЫХ ЗАДАНИЙ

***В.В. Гедранович*,** кандидат педагогических наук, доцент, проректор по научной и международной работе Минского института управления
***А.Б. Гедранович*,** кандидат экономических наук, доцент, заведующий кафедрой информационных технологий и высшей математики Минского института управления

Ключевые слова: тестирование, модель измерений, сложность заданий, уровень компетенций, оценка знаний

Для оценки качества усвоения учебного материала и выявления пробелов в знаниях студентов все чаще применяется тестирование, однако до сих пор ведутся споры об адекватности результатов такой оценки учебных достижений студентов. Несмотря на многочисленные критические отзывы о тестовой системе, следует признать ее одной из самых прогрессивных. Во-первых, грамотно составленные тестовые задания предназначены для объективной оценки знаний, что практически исключает человеческий фактор при выставлении итоговой отметки. Во-вторых, поскольку процедура тестирования хорошо поддается автоматизации, – это самый дешевый способ проведения промежуточного и итогового контроля, что крайне важно в условиях всеобщего перехода к массовому образованию. В-третьих, тестирование может проводиться удаленно, а это позволяет применять его для контроля знаний в дистанционном обучении.

Неоспоримым фактом является то, что в ближайшее время тестирование будет применяться повсеместно: централизованное тестирование для абитуриентов при поступлении в учреждения образования; массовое распространение тестирования в вузах во время текущей аттестации. Можно ожидать, что тестирование будет применяться и при итоговой государственной аттестации в учреждениях образования.

Проблемы использования универсальной системы тестирования в учреждении образования

Можно выделить два типа систем тестирования: предметные и универсальные. Предметными называют системы, которые учитывают специфику дисциплины и используют особые подходы к выявлению учебных достижений обучаемых. Уни-

версальные системы тестирования пригодны для массовой оценки знаний. К ним относятся системы республиканского уровня и системы, применяемые на уровне одного учреждения. Наибольшее распространение, в силу своей экономичности, получили универсальные системы тестирования. Однако для них можно выделить следующие общие проблемы:

– слабая обратная связь (результаты тестирования → тестовые задания);

– уровень сложности тестовых заданий устанавливается преподавателем;

– смещение в оценке уровня предметных компетенций студентов;

– наличие «проблемных» и «бесполезных» тестовых заданий;

– применение единого для всех дисциплин алгоритма выставления итоговой отметки;

– замена качества тестовых заданий на их количество.

В универсальных системах тестирования, как правило, хорошо разработана обратная связь для оценки результатов конкретного студента, но очень слабо – для проведения аудита тестов в целом. Зачастую для анализа тестов преподаватель может получить данные по числу студентов, отвечавших на вопрос теста, и количеству правильных ответов. Такие данные не могут служить основанием для оценки уровня сложности вопроса и его корректности.

Чаще всего уровень сложности тестовых заданий определяет преподаватель или группа преподавателей, основываясь на своем опыте. На самом деле ожидания преподавателя и фактический уровень сложности задания могут значительно отличаться. Вместе с тем в алгоритмах, которые применяются в большинстве систем тестирования, сложность задания играет существенную роль при выставлении итоговой отметки.

Если сложность тестовых заданий определена неверно, то это приводит к смещению в оценке уровня предметных компетенций студентов. В результате отметка студента может быть как завышена, так и занижена – т.е. выставлена неправильно.

К «проблемным» тестовым заданиям относятся задания с некорректными формулировками вопросов, неверными ответами, отсутствием правильных ответов и др. «Бесполезные» тестовые задания не дают информации о предметных компетенциях студентов. К этой категории можно отнести заведомо слишком простые или слишком сложные вопросы. Однако за ответы на них также начисляются баллы, а время, отведенное на тестирование, тратится впустую.

Универсальные системы тестирования должны предоставлять возможность выбора алгоритма выставления итоговой отметки с учетом специфики дисциплин. По ряду дисциплин, например, «Криминалистика» или «Бухгалтерский учет», 40% правильных ответов с учетом случайного выбора правильного ответа (угадывания) не может быть основанием для выставления положительной отметки.

Замена качества тестовых заданий на их количество связано, в первую очередь, с организационными проблемами, например, утечкой информации. Предполагается, что студент может легко заучить конкретные ответы на определенные вопросы теста, когда количество тестовых заданий невелико, что приведет к высоким оценкам, не подкрепленным знаниями. Вместе с тем получается, что студент, заучивший ответы на большое количество тестовых заданий, будет обладать высоким уровнем предметных компетенций. Это ошибочный способ решения проблемы, ведь знание ответов на большое количество не очень качественных тестовых заданий не может свидетельствовать о высоком уровне компетентности студента.

Таким образом, прежде чем предлагать тест студентам и надежно полагаться на его результаты, расценивая их как отражение уровня знаний тестируемых, необходимо оценить качество самих тестов.

Краткое описание методики оценки качества тестовых заданий

Методика, позволяющая решить часть обозначенных проблем, может быть реализована в универсальных системах тестирования. Используя математический аппарат, можно по ответам студентов на вопросы тестов выявить фактическую сложность заданий, фактический уровень компетенций студентов, определить примерный список «проблемных» и «бесполезных» вопросов и наметить план мероприятий по улучшению тестовых заданий. Методика базируется на теории измерений латентных переменных,

разработанной Георгом Рашем (*Geor Rasch*) и его последователями [1, 2] и применяемой во многих исследованиях [3–6].

В рамках теории измерений значения всех переменных определяются по единой линейной шкале в безразмерных величинах, называемых логитами (*logits*). Это позволяет, например, сопоставлять сложность тестового вопроса и уровень предметных компетенций студента. Кроме того, к преимуществам теории измерений можно отнести возможность измерения фактической сложности заданий, возможность организации адаптивного тестирования и развитый статистический аппарат оценки результатов тестирования.

Для оценки качества тестовых заданий можно применить дихотомическую модель измерений, оценивающую лишь факт наличия правильного или неправильного ответа студента на вопрос. Формально она может быть представлена следующим образом:

$$P\{X_{ij} = 1\} = P_{ij} = \frac{\exp(\beta_i - \delta_j)}{1 + \exp(\beta_i - \delta_j)},$$

где P_{ij} – вероятность правильного ответа i-ым студентом на j-ое задание;

X_{ij} – индикаторная переменная для ответа i-го студента на j-ое задание, 1 – верный ответ, 0 – неверный ответ;

β_i – уровень предметных компетенций i-го студента, логиты;

δ_j – фактическая сложность j-го задания, логиты.

Оценивание параметров модели (векторы β и δ), как правило, проводится с помощью метода максимального правдоподобия. В нашем исследовании для оценивания параметров была задействована библиотека eRm [7] для статистической платформы R [8].

Экспериментальное исследование

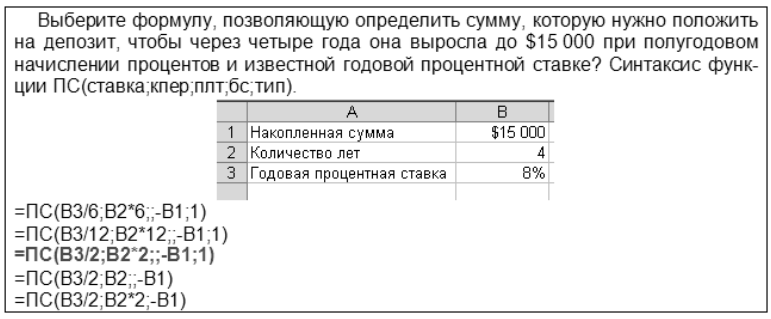
Применение методики оценки качества тестовых заданий рассмотрим на примере анализа практического тестового задания по дисциплине «Финансовая информатика». В эксперименте использованы данные по результатам тестирования 259 студентов из восьми групп разных потоков.

Итоговый контроль по этой дисциплине проводится в форме тестирования с использованием АОС «OpenBook» по двум тестам: теоретическому и практическому. Результаты каждого теста оцениваются по десятибалльной шкале в соответствии с алгоритмом, заданным для каждого теста. Данные для алгоритма выставления отметки по практическому тесту представлены в таблице 1.

Таблица 1 – Данные для алгоритма выставления отметки по практике

Проценты	Отметка	Проценты	Отметка
[0–20)	1	[55–65)	6
[20–30)	2	[65–75)	7
[30–40)	3	[75–90)	8
[40–50)	4	[90–95)	9
[50–55)	5	[95–100]	10

Практический тест состоит из 27 заданий с выбором единственного правильного ответа (рис. 1). Студент в течение 20 минут должен дать ответы на 10 заданий, сформированных для него по определенному правилу автоматизированной обучающей системой.


Выберите формулу, позволяющую определить сумму, которую нужно положить на депозит, чтобы через четыре года она выросла до $15 000 при полугодовом начислении процентов и известной годовой процентной ставке? Синтаксис функции ПС(ставка;кпер;плт;бс;тип).

	A	B
1	Накопленная сумма	$15 000
2	Количество лет	4
3	Годовая процентная ставка	8%

=ПС(B3/6;B2*6;;-B1;1)
=ПС(B3/12;B2*12;;-B1;1)
=ПС(B3/2;B2*2;;-B1;1)
=ПС(B3/2;B2;;-B1)
=ПС(B3/2;B2*2;-B1)

Рисунок 1 – Пример тестового задания

В результате тестирования получены данные, представленные в таблице 2.

Таблица 2 – Исходные данные

Задание\Студент	Студент №1	Студент №2	...	Студент №259
Задание №1	1	0	...	
Задание №2	0		...	1
Задание №3	1	1	...	
Задание №4	0	0	...	0
...	...	...	...	...
Задание №27		1	...	1

1 – правильный ответ;

0 – неправильный ответ;

пусто – задание не выполнялось.

Предложенная методика позволила получить оценки для параметров, характеризующих фактическую сложность заданий. Измерения в логитах линейно отображены на шкале от 1 до 3 и округлены до целых (таблица 3).

Таблица 3 – Сложность заданий

Задание	**Сложность (δ_j), логиты**	**Сложность (1..3)**
Задание №1	–0,6386	2
Задание №2	–0,0374	2
Задание №3	–1,4499	1
Задание №4	–0,4074	2
Задание №5	1,1894	3
Задание №6	–0,7341	2
...	...	...

Аналогичные операции были проведены и для оценок уровня предметных компетенций студентов (табл. 4), за тем исключением, что в этом случае шкала была от 1 до 10.

Таблица 4 – Уровень компетенций

Студент	**Уровень компетенций (β_i)**		**Оценка за тест**
	логиты	**(1..10)**	
Студент №1	0,6859	7	7
Студент №2	–0,7767	4	4
Студент №3	1,2095	8	8
Студент №4	–0,1532	6	5
Студент №5	–1,0715	4	3
...	...	...	...

Карта, отображающая взаимосвязь сложности тестовых заданий и компетенций студентов, имеет следующую структуру: в верхней части графика приведена частота для студентов с определенным уровнем компетенций; в нижней части графика каждая точка соответствует тестовому заданию, которые расположены в порядке возрастания их сложности (рис. 2).

На карте можно заметить следующие закономерности:

– в исследуемой выборке присутствуют студенты, уровень предметных компетенций которых ниже сложности самого простого вопроса (столбики частоты левее крайней точки), и студенты, уровень предметных компетенций которых выше, чем сложность самого сложного вопроса (столбики частоты правее крайней точки);

– большинство тестовых вопросов позволяют измерить уровень компетенций большинства студентов (шкалы пересекаются на диапазоне от -1,45 до 1,19).

На карте заданий (рис. 3) по шкале OY указана сложность заданий в логитах, а по шкале OX – наблюдаемое значение *t*-статистики, тестирующей нулевую гипотезу об однородности тестовых заданий.

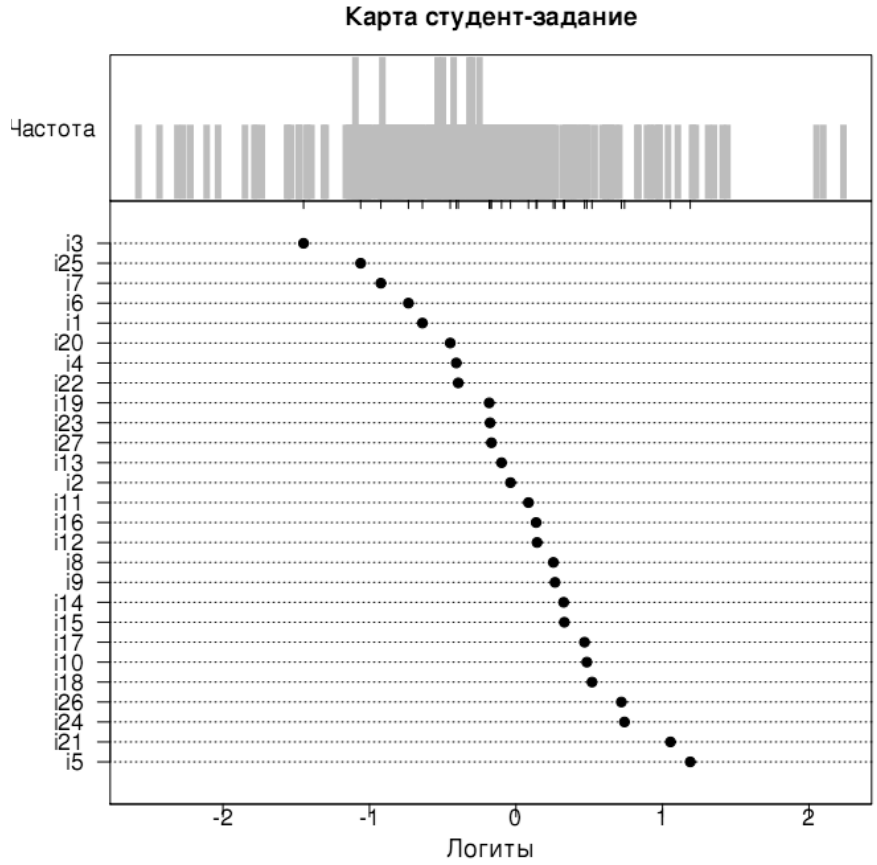


Рисунок 2 – Карта взаимосвязи сложности тестовых заданий и уровня компетенций студентов

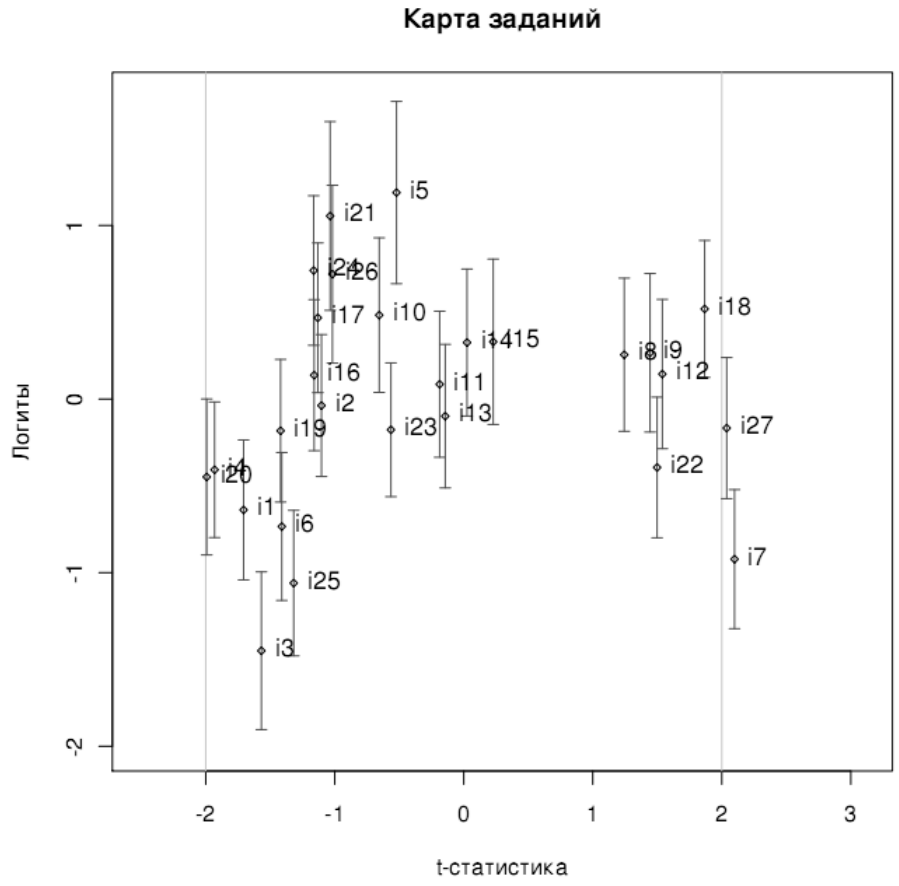


Рисунок 3 – Карта заданий

Если *t*-статистика для некоторого задания оказалась меньше -2 (левее нижней границы диапазона) – это указывает на то, что, скорее всего, это задание «проблемное», т.е. такое, что студенты с уровнем компетенций ниже, чем сложность задания, систематически показывают результаты лучшие, чем у более сильных студентов.

Те задания, которые оказались «справа», могут быть «бесполезными», т.е. не вносящими дополнительную ясность в оценку предметных компетенций студентов. На практике время, отведенное на такие вопросы, тратится впустую.

Аналогичные рассуждения имеют место и для карты студентов (рис. 4).

Слева – «везунчики», систематически отвечающие на вопросы с уровнем сложности выше своих компетенций, но не отвечающие на вопросы с уровнем сложности ниже своих оцененных компетенций. Возможно, что это является результатом угадывания или списывания.

Справа – «трудяги», которые подтверждают свой уровень ответом практически на каждый вопрос: если он заметно сложнее уровня их компетенций, то систематически дается неверный ответ, если ниже – верный.

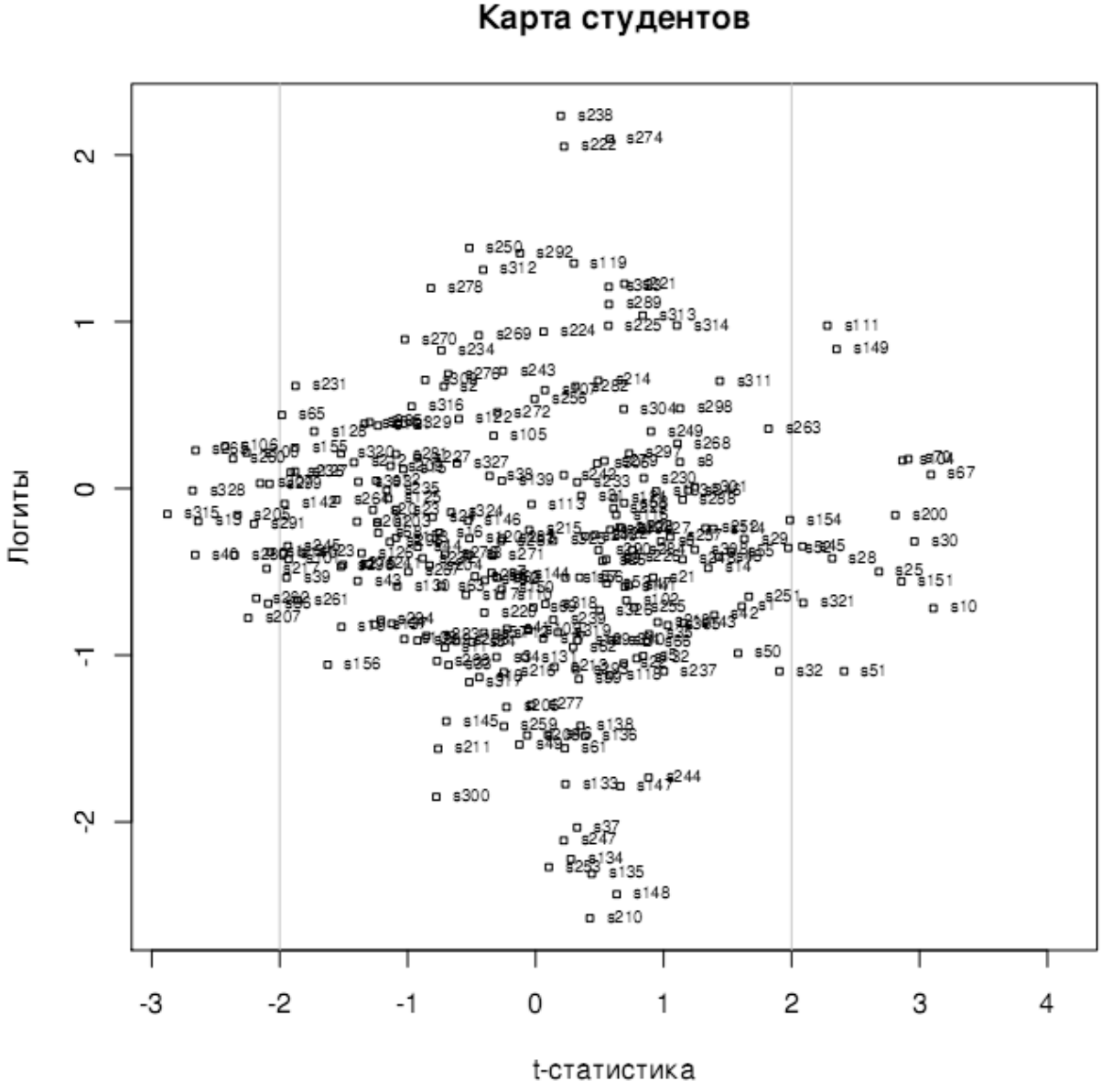


Рисунок 4 – Карта студентов

Таким образом, основываясь на результатах эксперимента, можно сделать вывод о том, что использование теории измерений для оценки и корректировки сложности тестовых заданий, использование теории измерений для отслеживания «проблемных» и «бесполезных» заданий облегчает проведение постоянного аудита качества тестовых заданий, используемых для оценки учебных достижений студентов. Применение теории измерений возможно только при предоставлении преподавателям полного отчета по результатам тестирования в централизованных системах тестирования.

ЛИТЕРАТУРА

1. Rasch, G. Probabilistic models for some intelligence and attainment tests / G. Rasch. – University of Chicago Press**,** Chicago, 1980.

2. Wright, B.D. Measurement essentials. 2nd edition / B.D. Wright, M. Stone. – Wilmington, Delaware, 1999. – 221 p.

3. Маслак, А.А. Измерение уровня развития инфраструктуры сферы образования в субъектах РФ / А.А. Маслак, С.А. Поздняков, А.А. Данилов // Высшее образование в России. – 2008. – №2. – С. 102–108.

4. Маслак, А.А. Измерение латентных переменных в образовании / А.А. Маслак, Т.С. Анисимова // Экономика и образование сегодня. – 2007. – №13. – С. 85–88.

5. Гедранович, А.Б. Измерение качества образовательных услуг вузов с помощью латентных переменных / А.Б. Гедранович // Управление в социальных и экономических системах: материалы XIX-й междунар. науч.-практ. конф., Минск, 18 мая 2010 г. – Минск: Издво МИУ. – 2010. – С. 271–272.

6. Гедранович, В.В. Квалиметрический инструментарий в управлении учебно-познавательной деятельностью студентов / В.В. Гедранович // Инновационные образовательные технологии. – 2005. – № 1. – С. 58–65.

7. Mair, P. Extended Rasch modeling / P. Mair, R. Hatzinger // The eRm package for the application of IRT models in R. Journal of Statistical Software. – 2007. – 20(9). – P. 1–20.

8. R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

РЕЗЮМЕ

Рассмотрены проблемы использования универсальных систем тестирования в учреждениях образования. Предложена методика оценки качества тестовых заданий на основе дихотомической модели измерений, базирующаяся на теории измерений латентных переменных. Основываясь на результатах эксперимента, обоснована целесообразность применения предложенной методики.

SUMMARY

The paper discusses the problems of using universal testing systems in educational institutions. The authors proposes the method of assessing the quality of tests on the basis of dichotomous measurement, which is well-known in the measurement theory of latent variables. The results of the experiment proves the feasibility of the proposed method.