Теория

Применение заданий в тестовой форме с выбором одного или нескольких правильных ответов

Альмира Искакова,
Национальный центр тестирования
Министерство образования Республики Казахстан

В статье исследуются статистические вопросы оценки качества заданий с выбором нескольких правильных ответов. Проверка качества таких заданий проводилась в рамках мониторингового исследования качества внешней оценки качества учебных достижений.

Ключевые слова: матрица, дисперсия, оценка, корреляция

Внешняя оценка учебных достижений (ВОУД) была введена в 2012 году в связи с внесением изменений в Закон об образовании Республики Казахстан. ВОУД осуществляется в целях оценки качества образовательных услуг и определения уровня освоения обучающимися образовательных учебных программ основного среднего образования.

В мониторинговом исследовании ВОУД на протяжении 4 лет применялись тестовые задания с выбором одного правильного ответа. В 2016 году впервые в рамках ВОУД были использованы задания с выбором одного или нескольких правильных ответов. Особенность заданий с выбором нескольких правильных ответов состоит в том, что надо найти все правильные ответы, при этом количество верных ответов нигде не оговаривается. Следовательно, эти задания труднее, они проверяют знания полнее, глубже и точнее, чем задания с выбором одного правильного ответа.

Несомненным достоинством формы заданий с выбором является её универсальность, однако имеется и существенный недостаток, задания с одним или несколькими правильными ответами не проверяют глубину и полноту знаний для количественных заданий по математике, физике, химии, хотя задания теоретического характера позволяют проверить знания достаточно глубоко.

Рассмотрим примеры заданий по физике:

1. АВТОМОБИЛЬ ДВИГАЛСЯ СО СКОРОСТЬЮ 54 КМ/Ч И ЗА 20 МИН ПРОШЕЛ ПУТЬ

1) 1080 км
2) 27 км
3) 18 км
4) 18000 м
5) 1,8· см
6) 1080 см
7) 1800 см
8) 2,7 км

2. РАБОТА СОВЕРШАЕТСЯ В СЛУЧАЕ

1) искусственный спутник вращается вокруг Земли
2) санки скатываются по абсолютно гладкой ледяной горке
3) груз равномерно поднимают в лифте
4) книга лежит на столе
5) трактор тянет прицеп
6) человек сидит на стуле
7) кран держит груз
8) подводная лодка залегла на дно

Как видно из первого задания, перевод из одной единицы измерения в другую не проверяет полноту и глубину знаний, в отличие от второго задания, для ответа на который требуется понимание физического процесса.

Существенный вопрос при использовании заданий с одним или несколькими правильными ответами — оценка, выставляемая за правильное выполнение. В нашем случает оценивание таких заданий производится следующим образом: 2 балла, если испытуемый не допустил ни одной ошибки, 1 балл в случае одной ошибки и 0 баллов при допущении 2-х и более ошибок. При незнании правильного ответа вероятность допустить хотя бы одну ошибку достаточно высока.

Соответствие требованиям тестовой формы и педагогической корректности содержания — это необходимые, но недостаточные условия для того чтоб задания назвались тестовыми. Превращение заданий в тестовой форме в тестовые задания начинается с момента определения статистических характеристик тестовых свойств заданий. Исходной информацией служит матрица ответов учащихся. Строками матрицы являются результаты испытуемых, упорядоченные в порядке убывания суммарной оценки за тест, а столбцами — тестовые задания. В ячейках матрицы находятся баллы, полученные каждым тестируемым за ответы на соответствующие задания.

Правильно сконструированный тест с нормативно-ориентированной интерпретацией результатов должен обеспечивать близкое к симметричному распределение индивидуальных баллов. В этой связи более удачной мерой вариации (изменчивости) результатов считается дисперсия тестовых баллов. Дисперсия является самым простым способом рассеивания баллов вокруг среднеарифметического значения В математической статистике наиболее употребительная мера рассеивания, отклонения случайных значений от среднего. Для выборочной совокупности дисперсия рассчитывается по следующей формуле:

$$S^2(x) = \frac{\sum_i (x_i - \bar{x})2}{n-1}, \qquad (1)$$

где n — число измерений, x_i — единичное значение, $\bar{x}$ — среднее значение.

Применительно к обработке результатов измерения дисперсия ха-

рактеризует меру вариации результатов.

Так как в ВОУД применялись задания двух форм: с одним правильным ответом (25 заданий) и одним или несколькими вариантами ответов (15 заданий) расчёт дисперсии проводился отдельно и приведён в табл. 1.

Таблица 1

Показатели дисперсии

	25 заданий	15 заданий	40 заданий
Дисперсия	28,8	37,2	106,0

Кроме дисперсии, для характеристики меры изменчивости распределения удобно использовать ещё один показатель вариации, который называется стандартным отклонением. Стандартное отклонение равно корню квадратному из дисперсии:

Таблица 2

Показатели стандартного отклонения

	25 заданий	15 заданий	40 заданий
Отклонение	5,36	6,1	10,3

Сравнивая два теста, в нашем случае, либо тест, состоящий из заданий только с одним правильным ответом (ВОУД 2012–2015), либо комбинированный тест, состоящий как из заданий с одним правильным ответом и одним или несколькими правильными ответами (ВОУД 2016), имеющих одну цель, необходимо выбрать тот инструментарий, который приводит к оценкам испытуемых с большей *надёжностью и валидностью*.

Эти два критических свойства тестовых оценок до некоторой степени зависят от дисперсии оценок, однако для нормативно-ориентированных тестов дисперсия является необходимым, но не достаточным условием гарантии полноценности теста

Надёжность результатов теста

Коэффициент надёжности результатов рассчитывается как коэффициент корреляции экспериментальных данных при выполнении двух половин одного и того же теста (метод расщепления) или одного и того же теста, но в разное время (ретестовый метод), или результатов тестирования параллельными вариантами. Рассмотрим различные методы вычисления надёжности теста.

1) Ретестовый метод — основан на повторном применении одного и того же теста на одной и той же группе испытуемых (рекомендуется не ранее, чем через 2 недели, и не позже, чем через 3 недели). Коэффициент надёжности в этом случае рассчитывается как коэффициент корреляции между оценками испытуемых по двум тестированиям.

Коэффициент надёжности, вычисленный ретестовым методом, может дать завышенное значение, особенно если повторное тестирование проводится слишком близко по времени. Учащиеся могут запомнить ответы к некоторым заданиям, что негативно скажется при оценке надёжности теста.

2) Метод параллельных форм.

Для исследования надёжности теста этим методом используется корреляция между результатами выполнения одной группой испытуемых двух параллельных форм теста. На практике этот метод используется крайне редко ввиду невоз-

можности разработки полностью параллельных вариантов. Однако, если проверена гипотеза о параллельности вариантов теста, этот метод можно применять.

Описанные два метода на практике используются редко, т.к. предполагают двукратное тестирование.

3) Метод расщепления — позволяет оценить надёжность теста при одном предъявлении теста группе испытуемых. Результаты тестирования делятся на две группы, например, в одну группу берутся все нечётные задания, в другую — все чётные задания. В качестве коэффициента надёжности берется коэффициент корреляции между оценками испытуемых по двум группам заданий.

В результате расщепления количество заданий (длина теста) уменьшается в два раза, поэтому значение коэффициента надёжности теста будет заниженным. Для его коррекции используют формулу Спирмена–Брауна:

$$r_{скор} = \frac{2 \cdot r_н}{1 + r_н}.$$

Метод расщепления основан на предположении о параллельности двух половин теста, что не всегда оказывается верным. Корреляция двух половин теста возрастает по мере роста гомогенности тестовых результатов. В этой связи коэффициент надёжности, вычисленный таким способом, иногда называют коэффициентом внутренней согласованности теста.

Отметим, что итоговые тесты таким способом лучше не расщеплять, т.к. необходимо при расщеплении учитывать содержание теста.

4) Формула Кьюдера–Ричардсон (KR-20).

Представляет собой упрощенный вариант коэффициента Кронбаха альфа, для случая дихотомических заданий. Формула Кьюдера-Ричардсон (KR-20) очень удобна:

$$r_{KR-20} = \frac{m}{m-1} \cdot \left(1 - \frac{\sum_{j=1}^{m} p_j q_j}{D_x} \right),$$

где m — число заданий в тесте; p_j — трудность j-го задания теста; $q_j = 1 - p_j$; D_x – дисперсия баллов испытуемых по всему тесту.

Вычислим по этой формуле надёжность теста, результаты выполнения которого приведены в табл. 3.

Таблица 3

Показатели надёжности

	25 заданий	15 заданий	40 заданий
Отклонение	0,83	0,85	0,89

Как видим, показатели надёжности достаточно приемлемые, и у 15 заданий с одним или несколькими вариантами ответов коэффициент надёжности на 0,2 превышает надёжность заданий с выбором одного правильного ответа.

Рекомендуется для большей точности для оценки коэффициента надёжности использовать различные методы.

В качестве нижнего предела допустимых значений надёжности обычно выбирают значение 0,7. При более низких значениях использование теста нецелесообразно ввиду большой погрешности измерения. К профессионально разработанным тестам предъявляются более жёсткие требования: тесты с надёжностью менее 0,8 считаются непригодными.

Положение с выводами о качестве теста осложняется тем, что коэффициент надёжности зависит от свойств выборки испытуемых, по результатам которых оценивается надёжность теста. Поэтому при

каждом использовании теста необходимо оценивать его надёжность и только после этого говорить о достоверной интерпретации выполнения теста.

К числу источников неудовлетворительной надёжности теста можно отнести:

1) субъективизм при оценке результатов выполнения заданий теста;

2) угадывание (как показывают исследования, угадывание существенно снижает надёжность теста, особенно в тех случаях, когда слабые ученики прибегают к догадке при выполнении наиболее трудных заданий теста);

3) отсутствие логической корректности формулировок заданий (как правило, некорректные задания искажают истинную картину, что в целом негативно отражается на надёжности теста);

4) неоправданный выбор весовых коэффициентов;

5) количество заданий (длина) теста;

6) отсутствие стандартной инструкции к тесту;

7) условия тестирования (шум, плохое освещение и т.д.);

8) плохое самочувствие испытуемого и пр.

Таким образом, применение в мониторинговом исследовании тестовых заданий с одним или несколькими вариантами ответов позволит проверить знания учащихся более полнее.