УДК 004.942:378.147

И.Н. Зянчурина, В.И. Некрасов

Организация тестирования с учетом степени сложности тестов

Национальный аэрокосмический университет им. Н.Е. Жуковского «Харьковский авиационный институт»

Подведение итогов результатов обучения является важной составной частью учебного процесса. Достижение высокого качества обучения возможно только при наличии объективных методов диагностики. Помимо объективности тесты обладают высокой степенью достоверности, также возможна проверка тестов на надежность и валидность.
В данной работе рассмотрены вопросы организации прохождения и вычисления норм тестов (нормирование индивидуальных результатов, нормативно-ориентированный подход, нормальное распределение индивидуальных баллов и т.д.).
Разработанная информационная система позволяет рассчитать индивидуальный балл учащегося, провести статистическую обработку полученных результатов и шкалирование результатов измерений, рассчитать характеристики заданий и показатели качества тестов.
Ключевые слова: качество тестов, шкалирование результатов, ошибка измерения, дистрактор, дискриминативность, коэффициент надежности.

Образование является важнейшей сферой социальной жизни. В обеспечении качественного образования заинтересован каждый субъект образовательного процесса (педагог, учащиеся, администрация и пр.). В практике любого преподавателя есть конфликтные случаи недовольства учащегося (студента) экзаменационной оценкой, в то же время подобные конфликты практически исключены при тестировании. В связи с этим процесс тестирования учебных достижений все шире применяется в образовательной деятельности.

В данной статье рассмотрены вопросы организации прохождения и вычисления норм тестов. Весь спектр проделанной самостоятельной работы учащихся начинается с этапа создания вопроса и завершается результатами ответов на этот вопрос. Проведение учета ответов на одни и те же вопросы выносится на этап анализа.

Разработанная информационная система позволяет рассчитать индивидуальный балл учащегося, провести статистическую обработку полученных результатов тестирования, разобрать шкалирование результатов тестовых измерений, провести расчет характеристик тестовых заданий, а также рассчитать показатели качества тестов. Система состоит из нескольких подпрограмм: модуль подбора тестов (которые будут проверяться на нормы), модуль реализации тестирования и модуль обработки результатов тестирования.

1. Расчет индивидуального балла учащегося. Шкалирование с помощью Z-шкалы (шкалы отклонений)

Для подсчета индивидуального балла каждого испытуемого X_i сравнивается ответ на каждый вопрос с правильным ответом, записанным в базе данных, и выводится на экран сумма правильных ответов.

Этот метод основан на подсчете отклонения «сырого» балла X_i от среднего значения индивидуальных баллов X по группе тестируемых. Значение Z_i — шкалированный результат каждого испытуемого находят по формуле

$$Z_i = \frac{X_i - \bar{X}}{S_x}, \quad (1)$$

где X_i — сырой балл *i*-го испытуемого;

$\bar{X}$ — среднее значение индивидуальных баллов N испытуемых группы (/=1, 2,..., N),

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}; \quad (2)$$

S_x – стандартное отклонение по множеству сырых баллов, подсчитанное по формуле

$$S_x = \sqrt{S_x^2}, S_x^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N-1}. \quad (3)$$

С помощью этой формулы мы вычисляем значения Z_i, составляем таблицу соответствия значений «сырого» балла X_i разности $X_i - \bar{X}$ и значения Z_i. Положительные значения Z_i свидетельствуют о хороших результатах, отрицательные — о плохих. Стандартное отклонение по множеству значений равно единице. С ее помощью можно привести баллы учеников, полученные по различным тестам, к одному удобному для сравнения виду путем нормирования индивидуальных результатов.

2. Статистическая обработка полученных результатов тестирования. Построение гистограммы распределения частот правильных ответов. Полигон частот

На основании таблицы частот можно построить полигон частот. По оси абсцисс отложены тестовые баллы, а по оси ординат - соответствующие частоты.

Обычно считают, что хороший нормативно-ориентированный тест обеспечивает нормальное распределение индивидуальных баллов репрезентативной выборки учеников, когда среднее значение баллов находится в центре, а остальные значения концентрируются вокруг среднего по нормальному закону, т.е. примерно 70% значений находятся в центре, а остальные сходят на нет к краям распределения (рис. 1).

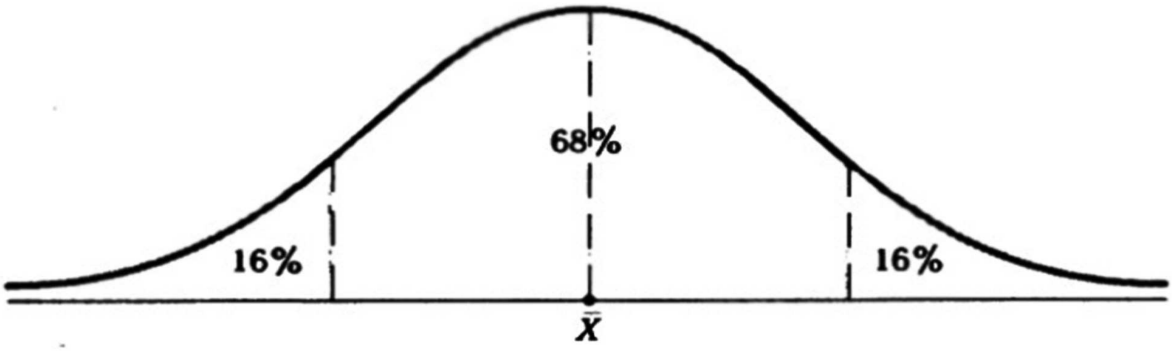

Рис. 1. Нормальная кривая распределения индивидуальных баллов

Гистограммы для распределения индивидуальных баллов по слишком легкой или слишком трудной подборке тестов могут выглядеть таким образом (рис. 2, 3).

Рис. 2. Гистограмма распределения баллов по трудному тесту

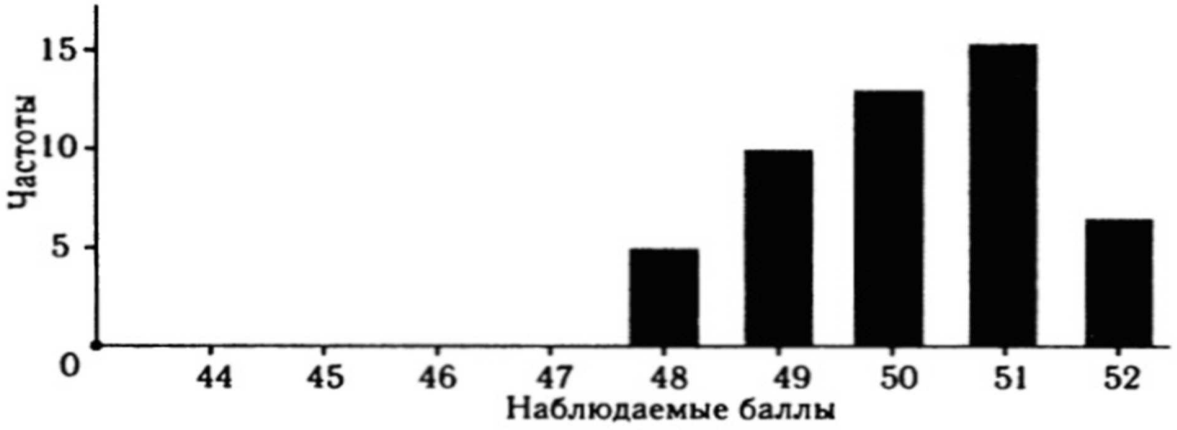

Рис. 3. Гистограмма распределения баллов по легкому тесту

3. Шкалирование результатов тестовых измерений

Для объективной оценки уровня подготовленности опрашиваемых по сравнению с другими участниками, которые проходят тестирование, часто применяется технология шкалирования результатов. В процессе шкалирования первичные результаты переводятся в сформированную по определенным правилам числовую систему, в которой отношение между свойствами объекта выражено в соответствующих числах, каждому первичному баллу ставится в соответствие определенный тестовый балл. Обычно применяются такие методы шкалирования.

Шкалирование с помощью процентильных рангов. Процентиль является производным показателем, указывающим на долю тех, кто правильно выполнил задание теста, от общего количества тестируемых в группе.

4. Расчет характеристик тестовых заданий

По результатам опробационного тестирования определяются характеристики тестовых заданий - трудность и дискриминативность.

Трудность тестовых заданий. Трудность задания вычисляется по формуле (мера легкости тестового задания)

$$p_j = \frac{R_j}{N}, \tag{4}$$

где P_j - доля правильных ответов на *j*-е задание*;*

R_j - количество испытуемых, выполнивших *j*-е задание верно;

N - количество испытуемых в группе;

j - номер задания*.*

Трудность тестового задания также можно вычислить в процентах P_j:

$$P_j = \frac{R_j}{N} \cdot 100\%. \tag{5}$$

Из формулы видно, что чем выше показатель трудности, тем задание легче, и, соответственно, чем меньше показатель трудности задания, тем задание сложнее. Например, если P = 30 %, то это значит, что только 30% испытуемых справились с этим заданием, а если P = 70 %, то 70 % справились с заданием. Получается, что первое задание сложнее, чем второе.

В рамках нормативно-ориентированного подхода наиболее удачными считаются задания средней трудности p=q=0,5, которые обеспечивают максимальную дисперсию теста σ=p*q. Это произведение достигает максимального значения (0,5 х 0,5 = 0,25) при p=0,5.

5. Анализ качества дистракторов в заданиях закрытой формы

Одно из важнейших требований, которое предъявляется к заданиям закрытой формы, - это правдоподобность дистракторов (равноценная вероятность выбора дистрактора при неправильном ответе). Дистракторный анализ предполагает подсчет долей испытуемых, выбравших каждый дистрактор. В идеальном варианте каждый дистрактор должен выбираться в равной доле от всех неправильных ответов. В табл. 1 приведено идеальное распределение долей.

Таблица 1

Распределение доли дистракторов

Номер задания	1-й ответ	2-й ответ	3-й ответ	4-й ответ
J	0,1	0,7	0,1	0,1

В табл. 1 показано, что правильно выполнили задание (выбрали 2-й ответ) 70 % испытуемых. Остальные 30 %, которые дали неправильные ответы, равномерно выбрали 1-й, 3-й, 4-й ответы, т.е. в задании были даны равновероятные дистракторы. Но такая идеальная картина распределения выбора неправильных ответов в реальной практике встречается редко.

6. Дискриминативность тестового задания

Дискриминативность (дифференцирующая способность, различающая способность) задания - это способность задания дифференцировать испытуемых по уровню достижений на сильных и слабых.

Один из способов вычисления дискриминативности - вычисление с применением метода крайних групп, где для расчета берутся показатели самых слабых и самых сильных испытуемых. Чаще всего это 27 (30) % худших и 27 (30)% лучших по результатам выполнения тестового задания.

Индекс дискриминативности определяется как разность долей правильных ответов сильной и слабой групп:

$$(r_{dis})_i = p_i^1 - p_i^0, \quad (6)$$

где r_{dis} - индекс дискриминативности;

p^1 - доля правильных ответов в сильной подгруппе (27 % от всего количества);

p^0 - доля правильных ответов в слабой подгруппе (27 %).

Если трудность задана в процентах, то

$$(r_{dis})_i = \frac{P_i^1 - P_i^0}{100\%}. \quad (7)$$

Значение индекса дискриминативности располагается в интервале [-1; 1].

В табл. 2 даны результаты расчета индекса дискриминативности.

Таблица 2

Результаты расчета индекса дискриминативности

Номер задания	P_j для слабой подгруппы	P_j для сильной подгруппы	Индекс r_{dis}
Question1	100	100	0
Question2	100	100	0
Question3	81,8	100	0,182
Question4	72,7	100	0,273
Question5	100	100	0
Question6	100	100	0
Question7	100	92,9	*-0,071*
Question8	90,9	100	*0,091*
Question9	90,9	100	*0,091*
Question10	9,1	57,1	0,48

7. Расчет показателей качества тестов

Статистическая обработка результатов тестирования позволяет, с одной стороны, объективно определить результаты испытуемых, с другой – оценить качество самого теста, тестовых заданий, в частности оценить его надежность.

Проблеме надежности уделено много внимания в классической теории тестов. Эта теория не потеряла своей актуальности и в настоящее время. Несмотря на появление более современных теорий классическая теория продолжает сохранять свои позиции.

Классическая теория тестов основывается на таких пяти основных положениях:

1. Эмпирически полученный результат измерения X представляет собой

(8)

сумму истинного результата измерения T и ошибки измерения E:

$$X = T + E.$$

Величины T и E обычно неизвестны.

2. Если мы рассматриваем наблюдаемую тестовую оценку как случайную переменную X, то истинный результат измерения можно выразить как математическое ожидание T=M(X).

3. Корреляция между истинной оценкой и ее ошибочным компонентом для генеральной совокупности испытуемых равна нулю $\rho(T,E)=0$.

4. Когда испытуемые выполняют два отдельных теста и оценки каждого испытуемого по двум тестам (или по двум тестированиям с помощью одной и той же формы теста) предполагаются случайно выбранными из двух независимых распределений возможных наблюдаемых оценок, корреляция между ошибочными компонентами оценок по этим двум тестированиям равна нулю:

$$\rho(E1, E2) = 0. \quad (9)$$

5. Ошибочные компоненты одного теста не коррелируют с истинными компонентами любого другого теста:

$$\rho(E1, T2) = 0. \quad (10)$$

Кроме этого, основу классической теории тестов составляют два определения – параллельные и эквивалентные тесты.

Ошибка измерения - статистическая величина, отражающая степень отклонения наблюдаемого балла от истинного балла испытуемого. Дисперсия наблюдаемых тестовых баллов будет равна сумме дисперсий истинных и ошибочных составляющих:

$$S_x^2 = S_T^2 + S_E^2. \quad (11)$$

Значит, чем ближе показатель дисперсии наблюдаемых баллов к дисперсии баллов истинных, тем выше корреляция между множеством наблюдаемых баллов X и множеством истинных баллов T, т.е. тест надежнее. Поэтому надежность теста (коэффициент надежности теста – $r_н$) определяется через отношение дисперсии истинного балла к дисперсии наблюдаемого тестового балла:

$$r_н = \frac{S_t^2}{S_x^2} = 1 - \frac{S_e^2}{S_x^2}. \quad (12)$$

Стандартная ошибка измерения находится как корень квадратный из дисперсии ошибочной компоненты:

$$S_e = \sqrt{S_e^2}. \quad (13)$$

Достижение высокого качества обучения возможно только при наличии объективных методов диагностики. К сожалению, традиционная форма оценивания уровня знаний в форме опроса, экзамена, проводимого человеком, весьма субъективна.

При использовании пятибалльной шкалы преподаватель выставляет оценки с разбросом плюс, минус один балл, то есть с точностью 20%. Из этого следует, что за одни и те же знания испытуемый может быть оценен разными экзаменаторами на «2», «3» и «4» балла. Более того, один и тот же экзаменатор в разные моменты времени, например с интервалом в один семестр, зачастую, по-разному оценивает один и тот же ответ.

Разработанная программа позволяет осуществлять разработку тестов, автоматизировать процедуру тестирования, а также реализовать обработку и интерпретацию результатов.

Список литературы

Ким В.С. Коррекция тестовых баллов на угадывание //Педагогические измерения, 2006. – № 4. – С. 47 – 55.

Майоров А.Н. Теория и практика создания тестов для системы образования. – М.: «Интеллект-центр», 2001. – 296 с.

Орлов А.И. Теория измерений и педагогическая диагностика //Педагогическая информатика, 2004. – № 1. – С. 22 – 31.

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

Рецензент: доктор техн. наук, проф., заведующий кафедрой информационных управляющих систем и технологий Федорович О.Е., Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков

Поступила в редакцию 12.05.11

Організація тестування з урахуванням ступеня складності тестів

Підведення підсумків результатів навчання є важливою складовою частиною навчального процесу. Досягнення високої якості освіти можливе тільки за наявності об'єктивних методів діагностики. Крім об'єктивності тести мають високий ступінь достовірності, також можлива перевірка тестів на надійність і валідність.

У цій роботі розглянуто питання організації проходження й обчислення норм тестів (нормування індивідуальних результатів, нормативно-орієнтований підхід, нормальний розподіл індивідуальних балів тощо).

Розроблена інформаційна система дозволяє розрахувати індивідуальний бал учня, провести статистичну обробку отриманих результатів і шкалювання результатів вимірювань, розрахувати характеристики завдань і показники якості тестів.

Ключові слова: якість тестів, шкалювання результатів, помилка вимірювання, дистрактор, дискримінативність, коефіцієнт надійності.

The organization of testing taking into account degree of complexity of tests

Summarizing of results of training is an important component of educational process. Quality achievement of formation is possible only in the presence of objective methods of diagnostics. Besides objectivity, tests possess high level of reliability, check of tests for reliability and validity also is possible.

In the given work questions of the organization of passage and calculation of test norms of tests (rationing of the individual results, the is standard-focused approach, normal distribution of individual points etc.) are considered.

Examinees The developed information system allows to calculate individual point of the pupil, to carry out statistical processing of the received results and scaling of results of measurements, to calculate characteristics of tasks and indicators of test quality

Keywords: quality of tests, scaling results, a measurement error, distractor, discriminative, reliability factor.