Е. О. Малыгин, Н. В. Швецов, Е. А. Павлова
Тюменский государственный университет, г. Тюмень
УДК 004.89

РАЗРАБОТКА СЕРВИСА ВАЛИДИЗАЦИИ ТЕСТОВЫХ ЗАДАНИЙ ДЛЯ СИСТЕМЫ MOODLE

Аннотация. В статье представлено выявление качественной шкалы для проверки информативности тестовых заданий и способы улучшить их, если они не соответствуют требованиям. Эти способы основаны на классической и современной теории тестирования. Результаты работы показали, что применение разработанных методов улучшают показатели тестовых заданий, которые используются в других исследованиях на схожие темы [1, 5].

Ключевые слова: оценка знаний, неинформативные вопросы, тестовые задания, теория тестирования.

Введение. В современном образовательном процессе для оценки уровня знаний тестируемых, для одного из наиболее трудозатратных этапов, то есть для контроля полученных знаний [1, 2], широко используются тестовые задания, представляющие собой стандартизированные вопросы с несколькими вариантами ответов на них.

Но тестовые задания должны быть составлены корректно, чтобы они могли точно определять уровень знаний тестируемых, иначе вопрос считается неинформативным.

Неинформативные вопросы — это вопросы, которые не приносят никакой полезной информации и не помогают достичь цели тестирования. Если тест содержит много неинформативных вопросов, то это может привести к ошибочным выводам о знаниях и компетенциях респондентов. Например, если тест содержит много вопросов, которые не имеют отношения к теме тестирования, то результаты могут показать, что респонденты не имеют достаточных знаний на эту тему, хотя на самом деле это не так [10].

Большинство характеристик тестовых заданий относится к классическому анализу заданий в конструировании тестов [3, 4].

Проблема исследования. Итак, целью работы является выявление коэффициентов информативности заданий, чтобы проверить тестовое задание, а также, если оно не соответствует требованиям, улучшить их.

Предметом анализа заданий является изучение полезности отдельных элементов для всего задания. Такой анализ заданий является важным инструментом для разработки тестовых заданий и оценки их надежности (как критерия).

Результатом анализа является определение психометрических критериев различных характеристик тестового задания.

Материалы и методы. В ходе работы было решено разделить неинформативные вопросы на следующие группы:

- Сравнительно легкие/сложные — вопросы, сложность которых значительно различается от сложности остальных вопросов.
- Дублирующие — вопросы, которые проверяют один набор знаний студента.
- Некорректные — вопросы с ошибкой в формулировке.
- Не относящиеся к теме — вопросы, для правильного ответа на которые необходимо обладать набором знаний, не относящимся к теме тестирования.

В работе использовались различные подходы к выявлению таких вопросов. Их можно разделить на Классическую и Современную теории тестирования.

Классическая теория тестирования также использовалась для вычисления характеристик заданий и оценки эффективности разработанных методов.

Современная теория тестирования основана на том, что предполагается существование взаимосвязи между модельной предсказуемостью ответов на задание и общим качеством знания [3].

Обозначим математическую постановку задачи.

Дана матрица баллов тестируемых за задания S_{mn}. Необходимо найти вектор поправочных коэффициентов заданий B_n, состоящий из чисел от нуля до единицы. Где коэффициент, равный нулю, соответствует неинформативному заданию. Где m — количество тестируемых, n — количество заданий.

Для тестирования программы используется скорректированная матрица баллов S'_{mn}:

$$\forall\, i \in m,\ \forall\, j \in n,\ S'_{ij} = S_{ij} \times B_j \qquad (1)$$

Для оценки эффективности вычисленных поправочных коэффициентов, решили вычислять разницу между метриками для исходной матрицы S_{mn} и скорректированной матрицы S'_{mn} в процентах.

Неинформативные вопросы были разделены на четыре категории. Для выявления вопросов каждой категории разработаны различные методы. Результатом каждого такого метода является вектор поправочных коэффициентов B_n.

При помощи моделей IRT можно вычислить матрицу вероятности правильного студента на вопросы. Чтобы получить вероятность для каждого задания можно вычислить средние значения для каждого задания.

Но использовать трехпараметрическую и четырехпараметрическую модель в рамках поставленной задачи не получится, потому что они используют параметры, которые невозможно вычислить на основе результатов теста.

$$P_{ij} = \frac{e^{r_j(b_i - b_j)}}{1 + e^{r_j(b_i - b_j)}} \quad (2),$$

где $b_i - b_j$ — совместная аддитивность (разность параметров тестируемого и задания). В качестве этого значение используется индекс дискриминативности:

$$\forall\, j \in n,\ r_j = \frac{\overline{S_j} - \overline{S}}{S_x} \sqrt{\frac{N_j}{N - N_j}} \quad (3)$$

где S_x — среднеквадратическое отклонение баллов, $\overline{S_j}$ — средний балл для тестируемых, верно решивших задание j, а N_j — количество таких тестируемых.

Вычислив среднее арифметическое матрицы Pmn по оси Y, получим вектор средней вероятности правильного ответа на задания P_j.

Но если использовать такой вектор в качестве коэффициентов, то задания с меньшим шансом правильного ответа, которые является сложными, будут давать меньше баллов, чем задания, на которые шанс правильного ответа высокий.

Полученный вектор необходимо преобразовать. Для достижения этой цели мы решили инвертировать значения вектора. Тогда

задания с высоким шансом ответа будут иметь низкий коэффициент.

Рассмотрим различные методы преобразования вероятности в коэффициент информативности задания.

Разница с идеалом, где C — желаемая вероятность:

$$B_j = 1 - \left|P_j - C\right| \quad (4)$$

Инвертирование:

$$B_j = 1 - P_j \quad (5)$$

Инвертирование с порогом:

$$B_j = 1 - P_j\,, \quad D \leq B_j \leq 1 \quad (6)$$

где D — порог для коэффициента неинформативности.

Таким образом, мы получим вектор поправочных коэффициентов.

Связь вопросов друг с другом можно определить при помощи вычисления корреляции вопросов друг с другом. Однако, методы вычисления корреляции основаны на изменение значений некоторых параметров.

Корреляция будет высокой при изменении правильности ответа на один вопрос, а правильность ответа на другом тоже изменится, причем в том же направлении, что и первый вопрос.

Этот метод может не выявить дубликаты вопросов, если ответы на хотя бы один из вопросов одинаковы у всех студентов. Для решения этой проблемы вычислим еще один коэффициент. Он будет учитывать возможность этой ситуации.

Необходимо подсчитать процент совпадения ответов на вопрос с каждым другим вопросом. Если ответы студентов на два вопроса совпадают у всех студентов, то для этих вопросов значение данного метода будет равно 100%. Соответственно, применив инвертирование, можем получить и коэффициент.

Таким образом, мы имеем два коэффициента для вычисления дубликатов. Чтобы свести их к одному, можно вычислить среднее между ними. Вычислим матрицу корреляции заданий Cnn и матрицу совпадение ответов Dnn.

$$\forall i \in n,\ \forall j \in n,\ r_{ij} = \frac{\sum_{t=0}^{m}\left(S_{ti} - \overline{S_i}\right)\left(S_{tj} - \overline{S_j}\right)}{\sqrt{\sum_{t=0}^{m}\left(S_{ti} - \overline{S_i}\right)^2 \cdot \sum_{t=0}^{m}\left(S_{tj} - \overline{S_j}\right)^2}} \quad (7)$$

$$\forall i \in n,\ \forall j \in n,\ D_{ij} = \frac{\sum_{t=0}^{m}\left(1 - \left|S_{ti} - S_{tj}\right|\right)}{n} \quad (8)$$

После чего вычисляем среднее арифметическое по оси Y для каждой из полученной матрицы. Получим два вектора коэффициентов зависимости r_n и Dn. Используем min-max нормализацию для вектора Dn. Далее вычислим среднее арифметическое между двумя полученными векторами, получив вектор E_n.

Теперь используем инвертирование для получения вектора B_n.

Некорректные задания и задания, которые не связанные с темой, имеют одну общую черту. Процент правильных ответов на такие задания примерно равен среди студентов с разной успеваемостью.

Для оценки вопросов с точки зрения дискриминативности используется кластерный анализ [8]. Мы разбили матрицу баллов S_{mn} на k кластеров с помощью метода KMeans по уровню знаний тестируемых. За уровень знаний тестируемых принимается суммарный балл по матрице S_{mn}.

Получим k матриц баллов S_{kx_kn} и вычислим процент выполнений заданий:

$$\forall z \in k,\ \forall j \in n,\ P_{zj} = \frac{\sum_{i=0}^{x_z} S_{zij}}{x_z} \quad (9)$$

Бисериальный коэффициент поможет определить насколько хорошо каждое задание служит поставленной цели измерения. Но для этого нужно, чтобы одна переменная измерялась в дихотомической шкале, а другая в интервальной. В нашем случае баллы за каждое задание расположены в диапазоне от 0 до 1, поэтому был применен порог 0.5 для их размещения на дихотомической шкале. Итак, результаты выполнения тестовых заданий размещаются на дихотомической шкале, а индивидуальный балл тестируемых на интервальной.

При вычислении значения коэффициента для каждого тестового задания в каждом тесте нужно учитывать, что мы будем считать нулем коэффициенты у тестовых заданий, на которые ответили верно или ошиблись все тестируемые.

$$\forall j \in n,\ B_j = \frac{\overline{S_{1j}} - \overline{S}_{0j}}{s_x} \sqrt{\frac{m_{1j} \cdot m_0 j}{m_j \cdot (m_j - 1)}} \qquad (10)$$

где m_{1j} — число испытуемых, выполнивших задание j, m_{0j} — число испытуемых, не выполнивших его, S_{1j} — средний индивидуальный балл, справившихся с заданием j, S_{0j} — средний индивидуальный балл, не справившихся с заданием, s_x — стандартное отклонение для индивидуальных баллов всех студентов.

Вектор валидности каждого задания мы и будем считать вектором поправочных коэффициентов.

$$\forall j \in n,\ B_j = r_{pb}^{j} \qquad (11)$$

Но нужно учитывать, что коэффициент корреляции r_{pb} должен быть достаточно высоким. Так, Челышкова М.Б [9] предоставляет следующую рекомендацию для оценки бисериального коэффициента: $r_{pb} \geq 0{,}5$.

Таким образом, при использовании такого вектора поправочных коэффициентов задания с меньшей валидностью будут давать меньше баллов, чем те задания, которые оказались более валидными.

Для доступа к функционалу разработанной программы разработан REST API веб-сервис на Django.

Сервис принимает на вход результаты тестирования, осуществляет предобработку входных данных и направляет их в программу анализа результатов тестирования. На выход сервис отправляет пользователю информацию о неинформативных вопросах в тестовом задании, которое отправил пользователь.

Результаты тестирования представляют из себя строку формата csv, которая представляет матрицу баллов студентов за задания.

Для взаимодействия с REST API сервисом пользователь может использовать разработанное веб-приложение или плагин Moodle.

Для взаимодействия с API разработано собственное веб-приложение на Django. Данное приложение и API сервис является одним Django приложением, чтобы избавиться от лишних HTTP запросов.

Клиентская часть разработана на Django Templates, чтобы данные, которые используются только на HTML странице, передавались сразу внутри этой HTML страницы.

На главной странице пользователю предлагается загрузить данные о результатах тестирования в формате CSV в форму.

После загрузки результатов пользователю необходимо удалить столбцы и строки таблицы, которые не содержат данных о баллах за задания. Интерфейс представлен на рис. 1.

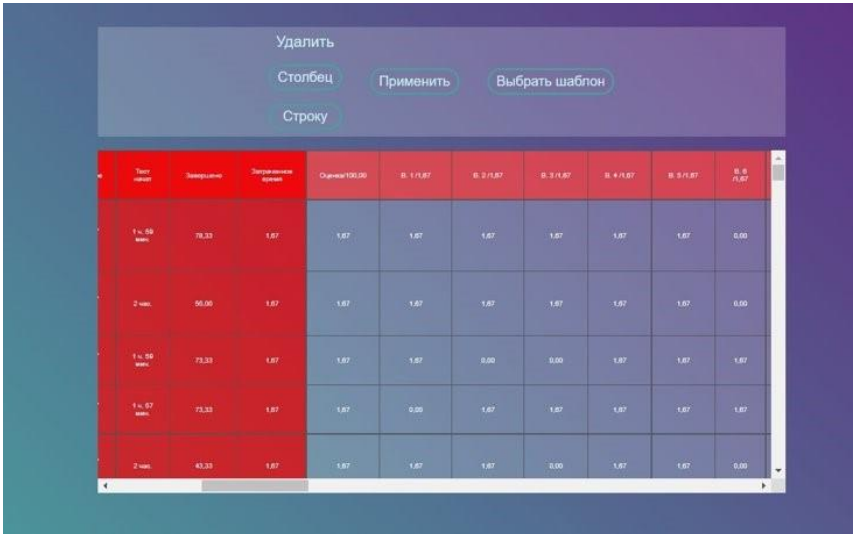

Рис. 1. Интерфейс удаления лишних столбцов и строк
при загрузке данных о результатах тестирования пользователем

Также, пользователь может выбрать шаблон входных данных с различных сервисов. Например, формат результатов тестирования сайта Moodle.

Матрица отправляется в программу анализа результатов, и пользователь получает информацию о неинформативных вопросах (рис. 2).

УП Итоговый тест 2022	
Коэффициенты информативности	
Легкие/сложные вопросы	44%
Дублирующие вопросы	25%
Некорректные вопросы	17%
Валидные вопросы	21%
Средний коэффициент	27%
Количество неинформативных вопросов	
Легкие/сложные вопросы	17 28%
Дублирующие вопросы	6 10%
Некорректные вопросы	7 11%
Валидные вопросы	9 15%
Всего	21 35%

Вопрос 1	Легкий	5
Вопрос 4	Легкий	8
	Дубликат	5
Вопрос 6	Некорректный	6
Вопрос 7	Легкий	9
Вопрос 9	Легкий	6
	Дубликат	7
Вопрос 11	Сложный	8
Вопрос 14	Легкий	6
Вопрос 18	Некорректный	7
Вопрос 19	Легкий	5
Вопрос 21	Сложный	8
Вопрос 25	Сложный	8

Рис. 2. Вывод информации о неинформативных вопросах

Для каждой категории неинформативных вопросов вычисляется оценка от нуля до десяти, которая определяет насколько вопрос неинформативен.

Пользователю предлагается список вопросов, которых есть оценки, значение которых равно пяти или более.

Результаты. Проведено тестирование разработанных методов на результатах тестирований по дисциплинам «Управление проектами», «Структуры и алгоритмы компьютерной обработки данных», «Языки программирования» и «Объектно-ориентированное программирование» в 2022 г.

Для тестирования используется скорректированная матрица баллов S'_{mn}

$$\forall\, i \in m,\ \forall\, j \in n,\ S'_{ij} = S_{ij} \cdot B_j \qquad (12)$$

Чтобы вычислить эффективность вычисленных поправочных коэффициентов, вычислялась разница между метриками для исходной матрицы S_{mn} и скорректированной матрицы S'_{mn} в процентах.

Использованы метрики:

- KR-20 — надежность и согласованность задания [6];
- Дельта Фергюсона — дискриминативность [5];
- Item Total Correlation — согласованность на противоречии заданий [7].

Вычисленная средняя разница между исходной и скорректированной матрицей по разным метрикам показала положительный рост (табл. 1).

Таблица 1

Эффективность вычисленных поправочных коэффициентов

Метрика	*Средняя разница между исходной и конечной матрицей (%)*
KR-20	22.50
Дельта Фергюсона	1.36
Item Total Correlation	1.02

Заключение. Проведенное тестирование выявило, что разработанные методы эффективней всего позволяют повысить уровень надежности и согласованности тестового задания по метрике Кудера-Ричардсона. Не удивительно, ведь неинформативные задания напрямую влияют на надежность и согласованность.

Дискриминативность заданий практически не изменяется. Однако, это связано с тем, что разработанные методы не улучшают задания, а лишь уменьшают влияние некорректных заданий. Поэтому разделение тестируемых по измеряемому признаку, по сути, не изменяется.

Item total correlation также практически не изменяется, что является показателем того, что уровень противоречия заданий не изменяется. Это является хорошим результатом, потому что разработанные методы не делают задания противоречивыми, то есть, учитывают согласованность.

Более того, согласованность улучшается (по метрике KR-20). Методы выявляют вопросы, которые препятствуют оценке знаний сами по себе. И не учитывается связь заданий друг с другом. Кроме метода нахождения дублирующих вопросов, который выявляет задания, которые слишком похожи по смыслу.

В результате работы были разработаны методы для выявления неинформативных вопросов, а также вычисления поправочных коэффициентов для тестовых заданий с учетом их значимости. Для доступа к функционалу разработанной программы был разработан веб-сервис, а для взаимодействия с API разработано веб-приложение. Также была проведена оценка созданных подходов для корректировки баллов за задания. Выявлены наиболее эффективные подходы. В веб-приложении визуализированы коэффициенты информативности вопросов, приведена статистика о неинформативных вопросах по всему тестовому заданию.

СПИСОК ЛИТЕРАТУРЫ

1. Чередниченко О. Ю. Модели тестирования знаний и методы оценки надежности полученных результатов / О. Ю. Чередниченко, С. И. Ершова, О. В. Янголенко, Т. Н. Запорожец — Текст: электронный // Восточно — Европейский журнал передовых технологий — 2011. — № 4(54). — URL: https://cyberleninka.ru/article/n/modeli-testirovaniya-znaniy-i-metody-otsenki-nadezhnosti-poluchennyh-rezultatov (дата обращения: 12.01.2023).
2. Ким В. С. Тестирование учебных достижений: монография / В. С. Ким. — Уссурийск : Издательство УГПИ, 2007. — 214 с. — URL: http://uss.dvfu.ru/static/kim_testing_monograph (дата обращения: 12.01.2023). — Текст: электронный.
3. Crocker L.M. Introduction to classical and modern test theory / L.M. Crocker, A. James. — New York: Holt, Rinehart, and Winston, 1987. — 482 p. — URL: https://archive.org/details/introductiontocl00croc/page/n7/mode/2up. (date of the application 12.01.2023). — Text: electronic.
4. Наследов А. Д. Профессиональный статистический анализ данных / А. Д. Наследов. — Санкт-Петербург: Питер, 2011. — 416 с. — Текст: непосредственный.
5. Hankins M. How discriminating are discriminative instruments? / M. Hankins. Health and Quality of Life Outcomes — 2008. — P. 1-5. —

URL: https://rdcu.be/c3cWE (date of the application: 12.01.2023) — Text: electronic.
6. Kuder G. The theory of the estimation of test reliability / G. Kuder, M. Richardson. — Psychometrika. — 1937. — Vol. 2(3). — P. 151-160. — URL: https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1342881 (date of the application: 12.01.2023) — Text: electronic.
7. Henrysson S. Correction of item-total correlations in item analysis / S. Henrysson. — Psychometrika. — 1963. — P. 211-218. — URL: https://link.springer.com/article/10.1007/BF02289618 (date of the application: 12.01.2023). — Text: electronic.
8. Карасева А. Е. Кластерный анализ готовности студентов к исследовательской деятельности / А. Е. Карасева, К. В. Хорошун, Р. В. Терюха. — Текст: непосредственный // Научные труды КубГТУ. — 2015. — № 5. — С. 1-15.
9. Челышкова М. Б. Теория и практика конструирования педагогических тестов: учебное пособие / М. Б. Челышкова. — Москва : Логос, 2002. — 432 с. — Текст: непосредственный.
10. Коржик И. А. Тестовая система Moodle и качество тестовых заданий / И. А. Коржик, И. В. Протасова, А. П. Толстобров. — Текст: электронный // Современные информационные технологии и ИТ-образование — 2012. — URL: https://cyberleninka.ru/article/n/testovaya-sistema-moodle-i-kachestvo-testovyh-zadaniy/viewer (дата обращения 27.02.2023).