УДК 373.5:159.9.072.59
МРНТИ 14.25.01

Искакова А.М., Адулова Е.В.
Костанайский региональный университет имени Ахмет Байтұрсынұлы
Центр педагогических измерений
г. Астана, Казахстан

МЕТОДЫ ОПРЕДЕЛЕНИЯ НАДЕЖНОСТИ РЕЗУЛЬТАТОВ ТЕСТИРОВАНИЯ В РАМКАХ КОНКУРСНОГО ОТБОРА ПЕДАГОГОВ

Статья посвящена одному из важных критериев результатов тестирования, определению надежности. Рассмотрены ключевые понятия тестологии, такие как тестовое задание, тест, надежность, уровень трудности заданий, уровень подготовленности испытуемых. Описаны методы определения надежности результатов тестирования. Для проведения расчетов были использованы результаты апробации тестовых заданий конкурсного отбора педагогов. Представлены результаты расчета коэффициента надежности по формулам Спирмена – Брауна, Кьюдера – Ричардсона, Рюлона. В целях обеспечения обьективности результатов тестирования, статистический анализ данных проводился не только с помощью специальных приложений, но и в EXCEL. В статье приведены рекомендуемые значения нормативных показателей коэффициента надежности. Статья представляет интерес для экспертов в области педагогических измерений, которые занимаются вопросами разработки и стандартизации измерительных материалов, используемых для контроля знаний.

Ключевые слова: надежность, тест, тестовое задание, корреляция.

The article discusses the key concepts of testology, such as test items, tests, test reliability, test items` difficulty level, and test takers` qualification level. The methods for determining the reliability of test results are described. The results of reliability coefficient calculation according to Spearman-Brown, Kuder-Richardson, and Rulon formulae are presented. The article presents the recommended values of reliability

coefficient indicators. It is of interest to experts in the field of pedagogical measurements who deal with assessment instruments used to control knowledge development and standardization.

Key words: reliability, test, test item, correlation.

Введение. С 2020 года АОО «Назарбаев Интеллектуальные школы» проводит конкурсный отбор педагогов на занятие вакантных должностей в Назарбаев Интеллектуальные школы. Основная задача первого этапа конкурсного отбора - определение уровня подготовленности педагогов по профильному предмету. Инструментами оценивания в рамках конкурсного отбора педагогов являются тест и эссе. В настоящее время тест является практически единственным технологичеким средством для определения уровня подготовленности претендентов. Несмотря на достаточно широкое применение теста, стоит отметить, что понятийный аппарат науки о тестах до сих пор окончательно не сформирован. В связи с этим многие понятия в области педагогических измерений по–разному трактуются исследователями, предложенные ими определения значительно отличаются друг от друга. В качестве примера рассмотрим одно из ключевых понятий тестологии – тест. Майоров А.Н. определяет тест как специальный инструмент, включающий квалиметрически обоснованную совокупность заданий, стандартизированные условия проведения и заранее разработанные процедуры обработки данных, который используется для диагностики свойств и качеств личности, поддающихся измерению в образовательном процессе [1, с.13]. Аванесов В.С. рассматривает тестирование как систему заданий, расположенных по нарастающей сложности и имеющих определённую форму, что обеспечивает возможность объективного измерения уровня и структуры подготовленности обучающихся [2, с.7]. По мнению Челышковой М.Б., тест представляет собой организованную в соответствии с определённой стратегией совокупность заданий, гарантирующую получение информативных данных об уровне и качестве подготовки тестируемых [3, с.10]. Рубинштейн С.Л. указывает, что тест

следует понимать как испытание, направленное на дифференциацию личности внутри группы, определение её позиционного места и уровня развития [4, с.76]. Анализируя приведенные определения, выбор оставляем за определением, представленным Аванесовым В.С., так оно наиболее полно отражает все характеристики теста.

Тестология разграничивает критерии теста и характеристики тестовых заданий. К ключевым критериям качества теста относятся надежность и валидность. Использование инструментов оценивания без определения надежности и валидности может негативно повлиять на результаты тестирования и, как следствие, отразиться на их объективности. Данная статья посвящена одному из критериев качества теста – надежности.

Цель данной статьи заключается в рассмотрении методов, позволяющих определить надежность результатов тестирования. Представленные методы использовались для оценки качества тестов по предметам математика, физика и информатика.

Задача статьи - определение надежности результатов тестирования в рамках конкурсного отбора педагогов.

Материалы и методы. Надёжность теста обычно определяется как стабильность и воспроизводимость получаемых результатов при повторном измерении у одной и той же выборки испытуемых в разное время, при применении альтернативных, но эквивалентных вариантов заданий, а также при изменении ряда условий проведения обследования [5, с.103]. Этот показатель относится к важнейшим характеристикам качества теста, поскольку он отражает точность измерений и степень устойчивости результатов к влиянию случайных внешних факторов [6, с.176].

В справочнике-словаре Бурлачука Л.Ф. и Морозова С.М. понятие «надёжность» предлагается трактовать в двух аспектах - широком и узком [7, с.82]. В широком понимании надёжность рассматривается как степень соответствия различий в тестовых результатах фактическим различиям в измеряемых качествах испытуемых. В узком же значении авторы связывают этот

термин с уровнем устойчивости и согласованности показателей, полученных при повторных измерениях тем же тестом.

В психометрической литературе надёжность преимущественно интерпретируется через показатель внутренней согласованности теста. Для её оценки традиционно применяется коэффициент α Кронбаха, характеризующий, насколько задания в тесте взаимосвязаны между собой. Значение коэффициента изменяется в диапазоне от 0 до 1, и чем оно выше, тем более согласованно функционируют элементы теста. Надежность результатов тестирования определяется различными методами, наиболее применимы на практике следующие методы:

– ретестовый метод;

– метод параллельных форм;

– метод расщепления теста.

Применение ретестового метода и метода параллельных форм в случае конкурсного отбора является неприемлемым, так как данные методы предполагают проведение повторного тестирования той же группы испытуемых. Метод расщепления теста является наиболее оптимальным, так как он предполагает однократное тестирование. Метод расщепления основан на допущении параллельности двух половин теста и предусматривает деление результатов тестирования на две части: данные по нечетным заданиям теста (x) и по четным (y).

Рассмотрим альтернативный метод для оценивания надежности по двум половинам теста на формуле Рюлона, без применения формулы коррекции. Данный метод включает использование разности оценок по половинам тестов:

$$r = 1 - \frac{S_d^2}{S_y^2}, \quad (1)$$

где S_y^2- дисперсия суммарных баллов результата, S_d^2- дисперсия разностей между результатами каждого кандидата по обеим половинам теста. Дисперсия вычисляется по формуле:

$$S_d^2 = \frac{\sum_{i=1}^{n}\left((x_i - y_i) - \frac{\sum_{i=1}^{n}(x_i - y_i)}{n}\right)^2}{n-1}, \quad (2)$$

Активаци
Чтобы актив

где x_i – индивидуальный балл i – го кандидата при ответе на тест по нечетным заданиям, y_i – индивидуальный балл i – го кандидата при ответе на тест по четным заданиям, n – число кандидатов, (x_i- y_i) – разность сумм баллов в строках с номером i субматриц с нечетными и четными заданиями.

Дисперсия суммарных баллов результата определяется по формуле (3):

$$S_x^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}, \quad (3)$$

где n – число кандидатов, x_i – индивидуальный балл i–го кандидата. Метод Рюлона является простым в использовании, однако он не приводит к единственной оценке коэффициента надежности теста, так как существует множество способов деления теста пополам, данные способы могут привести к различным оценкам надежности.

Рассмотрим следующий метод расчета коэффициента надежности полного теста с использованием коэффициента корреляции *r,* скорректированного по формуле Спирмена-Брауна. В связи с тем, что подсчет надежности проводится по расщепленному тесту, который в два раза короче, то оценка надежности теста корректируется по формуле (4):

$$r = \frac{2r_{\text{расщ}}}{1 + r_{\text{расщ}}}. \quad (4)$$

При применении формулы Спирмена–Брауна необходимо учитывать, что её использование опирается на допущение о полной параллельности двух получаемых половин теста. Соответственно, чем сильнее нарушается это условие, тем меньше будет точность рассчитанного показателя надёжности. В случае однократного проведения теста задания можно разделить на две эквивалентные части, а далее полученные результаты интерпретировать двумя

основными способами: либо определяя корреляцию между половинами с последующей коррекцией по формуле Спирмена–Брауна, либо рассчитывая надёжность с использованием формулы Рюлона.

В случае конкурсного отбора, когда выполнение заданий оценивается дихотомически (правильный ответ – 1 балл, неправильный ответ – 0 баллов), для расчета коэффициента надежности лучше использовать формулу Кьюдера – Ричардсона. Данная формула была получена Кьюдером и Ричардсоном в связи с тем, что известные методы расщепления теста пополам не в состоянии предоставить единую оценку надежности теста. Коэффициент надежности определяется по следующей формуле:

$$r = \frac{m}{m-1}\left(1 - \frac{\sum_{j=1}^{m} p_j q_j}{S_x^2}\right), \quad (5)$$

где, m – число заданий в тесте, p_j – доля правильных ответов на j–е задание теста, q_j – доля неправильных ответов на j–е задание теста, S_x^2- дисперсия индивидуальных баллов кандидатов. Доля правильных ответов при дихотомическом оценивании традиционно называется мерой трудности задания и определяется по формуле:

$$p_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}, \quad (6)$$

Следует отметить, что чем больше величина коэффициента p_j, тем большая часть испытуемых успешно справляется с заданием j. Так что на самом деле коэффициенты p_j должны интерпретироваться как показатели легкости заданий, а не трудности (6).

Результаты. В рамках данного исследования были использованы результаты конкурсного отбора 2022 и 2023 гг.

При проведении статистического анализа результатов тестирования для вычисления показателей вариации используются методы матричной алгебры. Матрица результатов тестирования является формой представления данных,

доступных для их обработки и интерпретации. В случае дихотомического оценивания матрица будет представлена в следующем виде:

Таблица 1. Матрица результатов тестирования

	Номера заданий				
	1	2	3	…	25
Person_001	1	0	0	1	0
Person_002	1	1	0	1	1
Person_003	0	0	1	0	0
Person_004	0	0	1	1	0
Person_005	0	0	1	0	0

Исходная тестовая матрица обычно имеет вид прямоугольной матрицы. В целях обеспечения достоверности статистических выводов число претендентов (строк) всегда должно быть больше числа заданий (столбцов); чем больше претендентов, тем выводы будут обоснованнее. В нашем случае 25 тестовых заданий по каждому предмету, число претендентов – более 100.

Для проведения расчетов коэффициента надежности были использованы формулы Рюлона, Спирмена – Брауна и Кьюдера – Ричардсона. Математико-статистическая обработка результатов тестирования осуществлялась с помощью пакета MS Excel и программы Winsteps. В качестве образца приведены расчеты коэффициента надежности по предмету математика. Для расчета надежности результатов тестирования были выполнены следующие действия:

1) подсчитаны индивидуальные баллы кандидатов (сумма баллов по строкам) X_i;
2) подсчитано количество правильных ответов на каждое задание (сумма баллов по столбцам) R_j.
3) подсчитано количество неправильных ответов на каждое задание W_j.
4) определена доля правильных ответов p_j и доля неправильных ответов на каждое задание q_j.

Далее надежность результатов тестирования расчитывалась по формулам Рюлона и Спирмена – Брауна. Матрица результатов тестирования была разбита на две части: первая часть содержала ответы по нечетным заданиям, вторая часть - ответы по четным заданиям. Были подсчитаны индивидуальные баллы претендентов по нечетным заданиям x_i и по четным заданиям y_i, а также их разность в соответствующих строках. Первый расчет коэффициента надежности проводился по формуле Рюлона (1), вычислена дисперсия разностей между результатами каждого кандидата по обеим половинам теста и дисперсия показателей по полному тесту (2,3). После выполнения расчетов коэффициент надежности принял значение – 0,81. Альтернативный метод расчета надежности основан на корреляции эквивалентных половин тест. Подсчитан коэффициент корреляции Пирсона для двух половин теста, который составил 0,69 в нашем случае. Полученное значение коэффициента надежности теста скорректировано по формуле Спирмена – Брауна (4), в результате вычислений скорректированный коэффициент надежности составил 0,82. И, наконец, надежность теста была определена по формуле Кьюдера – Ричардсона (5). После определения значения дисперсии и суммы произведений доли правильных ответов на долю неверных ответов коэффициент надежности составил 0,83.

Итоговые показатели коэффициента надежности за 2022 и 2023 годы в разрезе предметов, рассчитанные различными методами, представлены в таблицах 2 и 3.

Таблица 2. Коэффициент надежности результатов тестирования (2022 год)

Используемая формула	Математика	Физика	Информатика
Формула Кьюдера – Ричардсона	0,8313	0,6255	0,7885
Формула Спирмена – Брауна	0,8213	0,6462	0,7940
Формула Рюлона	0,8103	0,6381	0,7913

Таблица 3. Коэффициент надежности результатов тестирования (2023 год)

Используемая формула	Математика	Физика	Информатика
Формула Кьюдера – Ричардсона	0,8681	0,8161	0,7672
Формула Спирмена – Брауна	0,8662	0,8310	0,7928
Формула Рюлона	0,8565	0,8212	0,7928

Данные в таблицах 2 и 3 свидетельствуют о том, что по информатике наблюдается стабильность показателя надежности в 2022 и 2023 гг. По предмету математика прослеживается незначительная положительная динамика. Наибольшие изменения претерпел показатель надежности по физике: если в 2022 значение показателя надежности находилось практически у нижнего предела рекомендуемого диапазона, то в 2023 году надежность теста значительно повысилась, достигнув показателя 0,83. К одной из причин, повлиявших на рост коэффициента надежности, можно отнести внесение изменений в спецификацию теста и технологическую матрицу.

В работе авторов Siti Shahirah Saidi, Nyet Moi Siew представлена таблица с рекомендуемыми значениями показателей надежности [8,540]. Аналогичные значения показателей были представлены также в статье автора Ayu Faradillah [9,4]. Исходя из представленных данных, нижним пределом значения коэффициента надежности можно принять значение 0,5 (качество теста низкое). В случае, если показатель надежности 0,6 и выше, то качество теста можно считать удовлетворительным. Согласно данным в таблицах 2 и 3, качество теста, используемого в рамках конкурсного отбора, следует считать хорошим.

Таблица 4. Значения нормативных показателей надежности

	Значения	Качество теста
Альфа – Кронбаха (Cronbach Alpha)	<0,5	низкое
	0,5 - 0,6	удовлетворительное
	0,6 - 0,7	хорошее

	0,7 - 0,8	высокое
	>0,8	очень высокое

В целях получения полного представления об используемом инструменте проводился анализ тестовых заданий и критериев теста с применением программы Winsteps. Суммарная статистика по тестовым заданиям и участникам тестирования представлена в таблицах 5 и 6. Значения коэффициента надежности (Cronbach Alpha), полученные ранее в программе Excel, подтвердились, более того, были определены такие параметры, как среднее значение уровня подготовленности кандидатов, ошибка измерения, стандартное отклонение, статистика согласия.

Среднее значение уровня подготовленности кандидатов (-0,62) является отрицательным числом, что свидетельствует о том, что уровень их подготовленности ниже уровня трудности тестовых заданий [10, 8]. Средняя трудность заданий - на уровне 0,00 логитов. Согласно Rasch Measurement Theory, у кандидата есть шанс 50:50 выполнить задание с уровнем трудности 0,00 логит [11, 321]. Чем сложнее задание, тем ниже вероятность его выполнения. Поэтому в случае с данным тестом уровень трудности заданий высокий и, как следствие, шансы на успех у кандидатов низкие.

Таблица 5. Суммарная статистика по претендентам (предмет математика)

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	9,5	25.0	-.62	.50	1.00	.0	1.02	.1
S.D.	5,1	0,0	1.12	.08	.14	.6	.31	.7
MAX	22,0	25.0	2.22	.77	1.38	1.7	2.99	2.2

MIN	2,0	25.0	-2.75	.43	.77	-1.6	.53	-1.5
CRONBACH ALPHA "TEST" RELIABILITY = .83								

Таблица 6. Суммарная статистика по заданиям (предмет математика)

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	35.7	94.0	.00	.25	1.00	-.1	1.02	.0
S.D.	13.8	0	.84	.02	.18	1.6	.28	1.5
MAX	77.0	94.0	1.53	.32	1.47	3.2	1.65	4.0
MIN	14.0	94.0	-2.47	.23	.69	-3.5	.62	-3.1
CRONBACH ALPHA "TEST" RELIABILITY = .83								

Заключение

Надежность инструмента оценивания в рамках конкурсного отбора педагогов определялась несколькими способами. Коэффициенты надежности, полученные в результате использования различных методов, имеют значения, незначительно отличающиеся друг от друга. Анализ результатов конкурсного отбора свидетельствует о том, что надежность тестов по предметам математика, физика и информатика является допустимой. Тест конкурсного отбора позволяет провести объективный отбор кандидатов, однако следует рассмотреть вопросы содержательной валидности теста для выяснения причин низкой подготовленности кандидатов по отдельным темам. Детальный анализ

валидности результатов тестирования, который позволит оценить качество теста с другой позиции, будет представлен в следующей статье.

Список использованной литературы:

1. Майоров А.Н. Материалы курса «Основы теории и практики разработки тестов для оценки знаний школьников»: лекции 1–4. – М: Педагогический университет «Первое сентября», 2010. – С.13 – учебно-методическое пособие
2. Аванесов В.С. Теория и методика педагогических измерений: ЦТ и МКО УГТУ-УПИ, 2005. – С.7 – учебное пособие
3. Челышкова М.Б. Теория и практика конструирования педагогических тестов: Учебное пособие. – М.: Логос, 2002. – С. 10
4. Рубинштейн С.Л. К критике метода тестов / Против педологических извращений в педагогике. – Л., 1938. – С.76 -книга
5. Анастази А. Психологическое тестирование. М:Питер, 2007. – С.103 - книга
6. Мороз Л. С., Методы определения надежности и валидности тестов для контроля знаний // Труды БГТУ, Серия VI физико-математические науки и информатика - 2010. – С.176 – 179 – статья из журнала
7. Бурлачук Л.Ф., Морозов С.М. Словарь-справочник по психологической диагностике, - Киев: Наука, 1989. – С. 82 -словарь
8. Siti Shahirah Saidi, Nyet Moi Siew. Reliability and Validity Analysis of Statistical Reasoning Test Survey Instrument using the Rasch Measurement Model // International Electronic Journal of Mathematics Education – 2019. №3, 535-546
9. Ayu Faradillah, Hella Jusra, Trisna Roy Pradipta. Validity and Reliability Analysis of Algebraic Reasoning Test Instrument // Virtual Conference on Engineering, Science and Technology (ViCEST) – 2020. 4–6
10. Abdul Aziz, A., Jusoh, M.S., Omar, A.R., Amlus, M.H., & Awang Salleh, T.S. (2014). Construct Validity: A Rasch Measurement Model Approaches. J. Appl. Sci. & Agric., 9(12), 7-12

11. Rasch, G. On general laws and the meaning of measurement in psychology. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, IV, pp. 1961. 321. Berkeley, California: University of California Press

References

1. *Mayorov A.N. Marerialy kursa Osnovy teorii i praktiki razrabotki testov dlya otsenki znanui shkolnikov: lektsii 1-4.-M: Pedogogitchesky universiset Pervoye sentyabrya, 2010.- S.13 - uchebno-metodicheskoye posobiye*
2. *Avanesov V.S. Teoria I metodika pedagogicheskikh izmereniy:TzT I MKO UGTU-UPI, 2005.-S.7.- uchebnoye posobiye*
3. *Tchelyshkova M.B. Teoria I praktika construirovaniya pedagogicheskikh testov: Uchebnoye posobiye. - M.: Logos, 2002.-S. 10*
4. *Rubinshteyn S.L. K kritike metoda testov / Protiv pedologicheskikh izvrashcheniy v pedagogike. – L., 1938. – S. 76*
5. *Anastazi A. Psykhologicheskoye testirovaniye. M.: Piter, 2007.-S.103-kniga*
6. *Moroz L. S., Metody opredeleniya nadezhnosti i validnosti testov dlya kontrolya znaniy // Trudy BGTU, Seriya VI fiziko-matematicheskiye nauki i informatika - 2010. – S.176 – 179 – stat'ya iz zhurnala*
7. *Burlachuk L.F., Morozov S.M. Slovar-spravochnik po psikhologicheskoy diagnostike, - Kiev: Nauka, 1989. – S. 82*
8. *Siti Shahirah Saidi, Nyet Moi Siew. Reliability and Validity Analysis of Statistical Reasoning Test Survey Instrument using the Rasch Measurement Model // International Electronic Journal of Mathematics Education – 2019. №3, 535-546*
9. *Ayu Faradillah, Hella Jusra, Trisna Roy Pradipta. Validity and Reliability Analysis of Algebraic Reasoning Test Instrument // Virtual Conference on Engineering, Science and Technology (ViCEST) – 2020. 4–6*

10. *Abdul Aziz, A., Jusoh, M.S., Omar, A.R., Amlus, M.H., & Awang Salleh, T.S. (2014). Construct Validity: A Rasch Measurement Model Approaches. J. Appl. Sci. & Agric., 9(12), 7-12*
11. *Rasch, G. On general laws and the meaning of measurement in psychology. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, IV, pp. 1961. 321-334. Berkeley, California: University of California Press*

Наши авторы:

Искакова Альмира Мухтаровна – магистр технических наук, Костанайский региональный университет имени Ахмет Байтұрсынұлы, n.a.almira.24@gmail.com

Iskakova Almira Muchtarovna – master of technical sciences, A. Baitursynov Kostanay Regional University, n.a.almira.24@gmail.com

Адулова Елена Владимировна – магистр, «Центр педагогических измерений» АОО Назарбаев Интеллектуальные школы, adulova_ye@cpi.nis.edu.kz

Adulova Yelena – master of science in educational leadership, “Center for Pedagogical Measurements” under AEO “Nazarbayev Intellectual Schools", adulova_ye@cpi.nis.edu.kz