УДК 004.62:519.2

М. С. МАЗОРЧУК, Е. О. СОКОЛОВА, В. С. ДОБРЯК, А. А. СУХОБРУС

Национальный аэрокосмический университет им. Н. Е. Жуковского «ХАИ»

ОБОСНОВАНИЕ ВЫБОРА МЕТОДОВ ИЗМЕРЕНИЯ НАДЕЖНОСТИ ПЕДАГОГИЧЕСКИХ ТЕСТОВ

В данной работе предлагается анализ существующих методов и моделей оценки надежности педагогических тестов и обоснование выбора коэффициентов для оценки надежности, которые максимально отражают качество теста. Рассмотрены три основные метода анализа надежности: метод α-Кронбаха, Спирмена-Брауна и лямбда Гуттмана. При проведении расчетов использовались приложения статистического анализа данных: SPSS и MS Excel. Проведенные расчеты показали, что оценка надежности теста не является единственным методом оценивания качества теста. Также требуется оценивать сложность, валидность, дискриминативность и другие показатели, которые помогут сконструировать качественный тест.

Ключевые слова: *анализ надежности теста, качество педагогического теста, α-Кронбаха, лямбда Гуттмана, коэффициент Спирмена Брауна.*

Введение

В настоящее время методам анализа надежности педагогических тестов уделяется много внимания. Тесты являются одним их эффективных инструментов оценки уровня знаний и подготовки специалистов, особенно при массовом оценивании. Примерами массовой оценки подготовленности испытуемых могут выступать внешнее независимое тестирование в Украине, единый государственный экзамен в России, государственная итоговая аттестация учеников в 4-х и 9-х классах, международные тесты сравнительного мониторингового исследования TIMSS, PIRLS, PISA [1] и другие.

Разработка эффективного теста, который бы позволил дифференцировать учеников, давал возможность определить латентные факторы, влияющие на уровень подготовки, был достаточно сложным и согласованным, измерял различные аспекты подготовки обучаемого – является сложной и актуальной задачей. На создание эффективного теста требуется достаточно много времени; необходимо проведение эксперимента с привлечением педагогов-экспертов в конкретной предметной области; важно провести тщательный анализ психометрических характеристик (оценить надежность теста, сложность вопросов, валидность и т.д.). От качества теста сильно зависит и качество результатов оценивания. Поэтому, выбор методов и моделей оценки педагогических тестов должен быть обоснованным и аргументированным.

Целью данного исследования является анализ существующих методов и моделей оценки надежности педагогических тестов и обоснование выбора коэффициентов для оценки надежности, которые максимально отражают качество теста.

1. Постановка задачи

На первом этапе определимся с понятием «надежности теста». Надежность теста означает устойчивость и согласованность результатов теста по отношению к погрешностям измерения [2, 3]. На тест не должны влиять ни время проведения, ни место, ни люди, которые проходят тестирование.

Устойчивость результатов (test-retest reliability) – это возможность получения одинаковых результатов при повторном (retest) тестировании испытуемых. Для измерения устойчивости применяют методы, основанные на прохождении теста одной и той же группой людей через некоторый промежуток времени или использовании эквивалентных форм теста при повторном тестировании. Эти методы не являются совершенными, поскольку обеспечить одинаковые условия при повторном тестировании не всегда возможно (у испытуемого может быть другое настроение или самочувствие, часть вопросов испытуемые могут запомнить и т.д.), а доказать высокую степень эквивалентности тестов не всегда возможно. Однако данные методы часто применяют на практике особенно при проведении экспериментов на небольших группах.

Согласованность теста означает корреляцию результатов по отдельным вопросам теста с итоговыми общими результатами. Для оценки согласованности применяют метод расщепления (Split-half

reliability), который предполагает разделение теста на части и оценки корреляции между ними. При этом рассчитывают различные коэффициенты, отображающие силу согласованности: коэффициент Спирмена-Брауна, лямбда Гутмана.

Второй метод основан на оценке внутренней согласованности теста (internal consistency), который основан на определении коэффициента α-Кронбаха или коэффициента Кьюдера-Ричардсона (для дихотомических шкал). При использовании данного определяют корреляцию между каждым пунктом теста и общим результатом. Если каждый вопрос измеряет одно и тоже свойство, то в целом при увеличении числа вопросов, увеличивается и показатель α-Кронбаха, т.е. по сути данный показатель отражает гомогенность структуры теста (гомогенность теста свидетельствует о том, что все задания устойчиво измеряют одну и ту же характеристику). Однако, если тест направлен на измерение разных свойств, т.е. имеет гетерогенную структуру, то α-Кронбаха будет низким. Однако, в работе [4] отмечается, что данный показатель отражает внутреннюю согласованность теста, а не одномерность измерения и приводит результаты высокого значения показателя α-Кронбаха для теста, который имеют гетерогенную структуру (значения корреляции между пунктами теста довольно низкие). Высокие значения α-Кронбаха достигаются исключительно за счет увеличения числа вопросов теста, что в свою очередь приводит к ошибочным результатам и получению плохо сконструированного теста. Поэтому, применение этого метода часто необоснованно для измерения надежности педагогических тестов, когда требуется оценить различные умения и навыки испытуемого.

Также использование этого коэффициента требует соблюдения ряда условий [4 - 5], которые не всегда имеют место на практике. Согласно классической теории теста наблюдаемая оценка испытуемых является результатом суммы истинной оценки и ошибки. При использовании α-Кронбаха предполагается, что, во-первых, корреляция между истинными значениями и ошибками равна нулю; во-вторых, среднее значение ошибки по всем испытуемым равно нулю; в-третьих: корреляция между ошибками двух тестов равна нулю [5]. Эти допущения не всегда выполняются, поскольку на истинные результаты теста по каждому пункту может оказывать влияние большое количество скрытых факторов, и утверждение о том, что пункты имеют одинаковую дисперсию в эквивалентных тестах не всегда верно [6]. Также недостатки использования показателя α-Кронбаха обсуждаются в работах [7 - 9].

Основываясь на вышеизложенном, можно сформулировать постановку задачи исследования: необходимо проанализировать существующие подходы и методы к оценке надежности теста, провести численный анализ коэффициентов надежности на основе различных данных (различные объемы выборки, число вопросов, методы анализа и т.д.) и сделать выводы о применимости того или иного метода для оценки надежности педагогического теста.

2. Методы и модели оценки надежности педагогических тестов

Рассмотрим модели оценки показателей надежности тестов при однократном тестировании. Для оценки уровня согласованности теста при использовании метода расщепления (Split-half reliability) используют два основных коэффициента: Спирмена-Брауна и лямбда Гуттмана. Данный метод основан на допущении параллельности двух половин теста и предполагает деление результатов тестирования на две части: данные по нечетным заданиям теста (x) и по четным (y). Корреляция двух половин тестов возрастает по мере роста однородности (гомогенности) теста.

Коэффициент надежности вычисляют как коэффициент корреляции между результатами по двум половинам теста:

$$r_{xy}=\frac{\sum_{i=1}^{n} x_i y_i-\frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\sum_{i=1}^{n} x_i^2-\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}\sqrt{\sum_{i=1}^{n} y_i^2-\frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}}, \qquad (1)$$

где x_i - это суммарные результаты теста i -го испытуемого по нечетным заданиям ($i=\overline{1..n}$);

n - количество испытуемых;

y_i - суммарные результаты теста i -го испытуемого по четным заданиям.

Так как надежность r_{xy} подсчитывают по расщепленному тесту, который в два раза короче, то оценка надежности корректируется по формуле Спирмена–Брауна:

$$r_{sb}=\frac{2r_{xy}}{1+r_{xy}}. \qquad (2)$$

Если расщепление теста производится не на равные части, то используют формулу:

$$r_{sb'}=\frac{kr_{xy}}{1+(k-1)r_{xy}}, \qquad (3)$$

где k - коэффициент расщепления, т.е. общий размер выборки, деленный на размер выборки для каждой формы расщепления (можно заметить, что k =2 при расщеплении теста пополам).

На коэффициент надежности Спирмена-Брауна сильно влияет то, каким способом разбивают тест на две половины. Случайное распределение пунктов по двум формам (это обуславливает использование метода деления теста на четные и нечетные задания) повышает вероятность равенства дисперсий между формами расщепления, однако такое равенство не гарантируется и должно проверяться исследователем.

Для оценки надежности теста также широко применяется коэффициент лямбда Гутмана, который не требует равенства дисперсий между двумя расщепленными формами и рассчитывается как:

$$r_{\lambda}=2\left[1-\frac{\sigma_x^2-\sigma_y^2}{\sigma_{\Sigma}^2}\right], \qquad (4)$$

где σ_x^2 - дисперсия результатов по одной форме заданий теста;

σ_y^2 - дисперсия результатов остальных заданий;

σ_{Σ}^2 - дисперсия результатов по всему тесту.

В практике педагогических измерений часто используют способ оценки надежности с помощью формулы Кьюдера–Ричардсона (сокращенно принято обозначать формулу KR–20), которая может применяться только в том случае, если выполнение задания оценивается дихотомически (1 балл – правильно; 0 баллов – неправильно). Данный метод не требует разделения теста на две части, поскольку анализируется корреляция между всеми вопросами теста и суммарным результатом. Данный коэффициент определяется по формуле:

$$r_{KR\text{-}20}=\frac{m}{m-1}\left(1-\frac{\sum p_j q_j}{\sigma_{\Sigma}^2}\right), \qquad (5)$$

где $\sigma_{\Sigma}^2=\dfrac{\sum X_i^2-\dfrac{\left(\sum X_i\right)^2}{n}}{n-1}$ - дисперсия суммарных баллов испытуемых;

m – число заданий в тесте;

p_j - доля правильных ответов на j-е задание теста ($j=\overline{1..m}$);

X_i – индивидуальный балл i – го испытуемого.

Коэффициент (4) является частным случаем показателя надежности теста α-Кронбаха. Коэффициент α-Кронбаха используется для всех типов вопросов и рассчитывается по формуле:

$$r_{\alpha}=\frac{m}{m-1}\left(\frac{\sigma_{\Sigma}^2-\sum\sigma_{X_j}^2}{\sigma_{\Sigma}^2}\right), \qquad (6)$$

где $\sigma_{X_j}^2=\dfrac{\sum_{i=1}^{n}(X_{ij}-\bar{X}_j)^2}{n}$ - дисперсия отдельного пункта теста по всем испытуемым;

X_{ij} – индивидуальный балл i – го испытуемого ($i=\overline{1..n}$) по j-му заданию теста ($j=\overline{1..m}$) ;

$\bar{X}_j$ - среднее значение балла по j-му заданию;

$\sigma_{\Sigma}^2=\dfrac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}$ - дисперсия суммарного результата теста;

$\bar{X}$ - среднее значение суммарного балла теста.

Нижним пределом коэффициента надежности принято считать показатель 0,7. Если значение коэффициента надежности ниже, то надежность теста считается неудовлетворительной, так как возникает большая погрешность измерений. Для профессионально созданных тестов, которые используются в массовом тестировании и по их результатам принимаются ответственные решения, нижний предел коэффициента надежности повышается до 0,8. В таблице 1 представлены величины надежности.

Таблица 1

Распределение величины надёжности

Величина надежности	**Оценка надежности**
0,90…0,99	Отличная
0,80…0,89	Хорошая
0,70…0,79	Удовлетворительная
Менее 0,70	Неудовлетворительная

Рассмотренные методы и модели определения показателей надежности подходят только для оценки нормативно-ориентированных тестов. По мнению многих тестологов, эти методы нежелательно использовать для вычисления надежности критериально-ориентированного теста, так как дисперсия тестовых баллов в критериально-ориентированном тесте небольшая (здесь не нужна большая диффе-

ренциация баллов испытуемых), а соответственно и корреляционная оценка надежности будет низкой.

Следует учесть, что значения надежности, приведенные в таблице 1, имеют место для всех коэффициентов, но значение выше 0,9 не всегда отображает истинную надежность. При этом показатели надежности могут принимать значения меньше -1, если исходная популяция испытуемых очень слабо подготовлена (оценить качество теста в этом случае вообще не предоставляется возможным).

3. Численный эксперимент оценки надежности тестов

Рассмотрим на примере результатов тестирования знаний по математике, украинскому языку и литературе и английскому языку показатели надежности (2) – (6). Если количество вопросов в тесте было нечетным, то использовалась формула (3), а если четным, то (2). Если в тест входили вопросы только формата MSQ (Multiple Choice Questions), т.е. только один из вариантов ответа являлся правильным и за вопрос давался 1 балл, то использовался коэффициент KR-20 (5), а если результат по вопросу был больше 1, то рассчитывался коэффициент α-Кронбаха. В расчетах приводится только значение α-Кронбаха. На основе популяции 45000 учеников были сформированы случайные выборки по 50, 100, 200, 500, 1000, 1500, 2000 и более 3000 человек. Оценивались тесты из 5, 10, 15, 20, 25, 30, 35 и более вопросов. Расчеты проводились с помощью программных инструментариев SPSS [10] и MS Excel.

На рисунке 1 изображены графики изменения коэффициентов надежности в зависимости от длины теста и объема выборки для теста по математике. Как видно из графиков, коэффициент α-Кронбаха и Спирмена-Брауна существенно растет с увеличением длины теста, тогда как лямбда Гутмана, начиная с длины теста в 20 вопросов, практически не изменяется. При малом количестве вопросов все коэффициенты надежности не превышают 0,7 даже при больших объемах выборки, тогда как при длине теста более 15 вопросов значения показателей отличаются. Однако, для данного теста уже при длине в 15 вопросов мы можем видеть хорошую надёжность (коэффициент α-Кронбаха больше 0.8 при любом объеме выборки, а остальные коэффициенты при выборке более 200 человек также имеют высокие значения).

Также следует отметить, что на рис.1 приведены графики только для выборки 1000 испытуемых. Это обусловлено тем, что показатели надежности практически не изменяются для популяции выше 1000 испытуемых, поэтому они и не приводились.

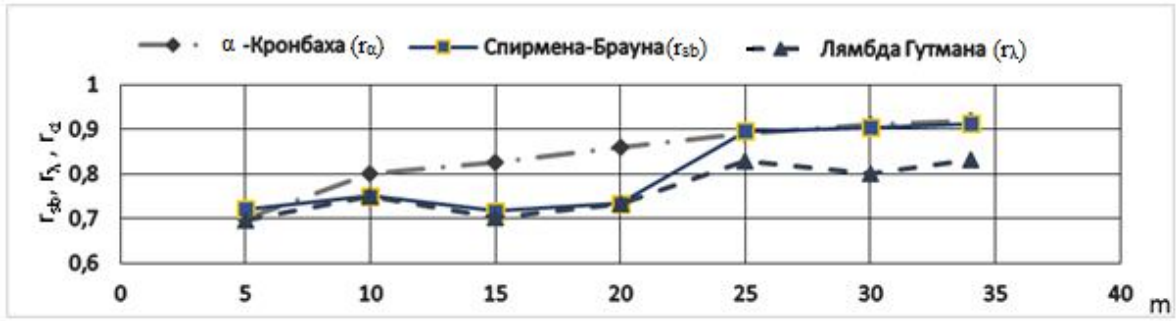
Рассмотрим как изменяются коэффициенты надежности для тестов по различным предметам длиной в 15 вопросов. На рис. 2 представлены зависимости показателей надежности теста от вида для выборки в 1000 человек. Как видно из графика, для английского языка по всем трем коэффициентам можно сделать выводы о том, что даже тест длинной в 15 вопросов является достаточно надежным, тогда как по украинскому языку и литературе, требуется более длинный тест.

Рассмотрим изменение коэффициента α-Кронбаха для различных групп участников тестирования. Результаты теста были переведены в шкалу 100-200 по методу эквипроцентильной нормализации [11]. На рис. 3 приведено распределение для теста по математике, где на горизонтальной оси указаны группы участников, которые набрали не более i баллов (i = 110, 120, …, 190). А на рис. 4 – показаны изменения коэффициента α-Кронбаха для групп участников, которые набрали более i баллов (i = 110, 120, …, 190).

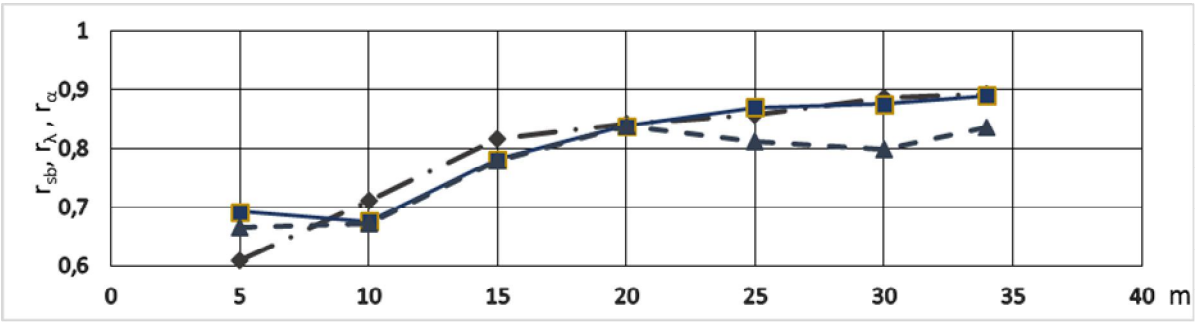
Как видно из диаграмм (рис. 3 и рис.4) коэффициент α-Кронбаха может принимать и отрицательные значения и быть по модулю значительно больше 1. Аналогичные результаты были получены и по остальным коэффициентам – лямбда Гутмана и Спирмена-Брауна. Также видно, что, для слабо подготовленных участников тестирования и для хорошо подготовленных, надежность теста не удовлетворяет заданным требованиям, что дает основание для поиска других форм оценки знаний учащихся, например, разработки двухуровневых тестов.

Заключение

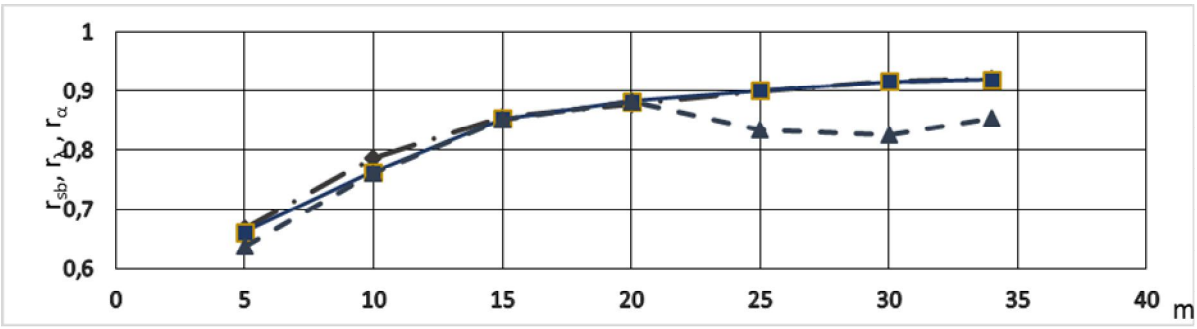
На основании проведенного анализа и приведенных расчетов можно сделать выводы, что коэффициент надежности α-Кронбаха не является наиболее рациональным решением для оценки внутренней согласованности и одномерности теста. Данный коэффициент возрастает пропорционально длине теста, что дает основание для разработчиков тестов чаще использовать именно данный показатель, чтобы доказать надежность инструмента оценивания. Однако коэффициент лямбда Гутмана зависит от длины теста, однако не растет прямо пропорционально увеличению числа вопросов. Поэтому, использование данного коэффициента более рационально. Это дает возможность не разрабатывать длинные тесты, а ограничиваться длиной в 15-20 вопросов. Также важно оценивать надежность для групп испытуемых с различным уровнем подготовки.



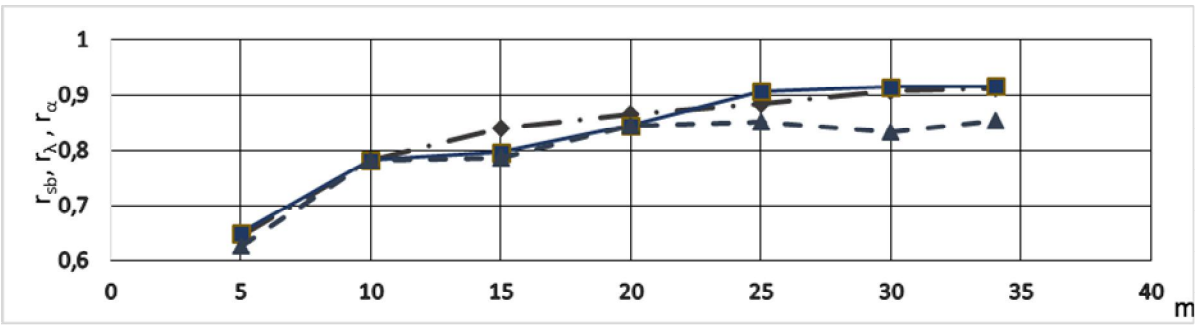
а



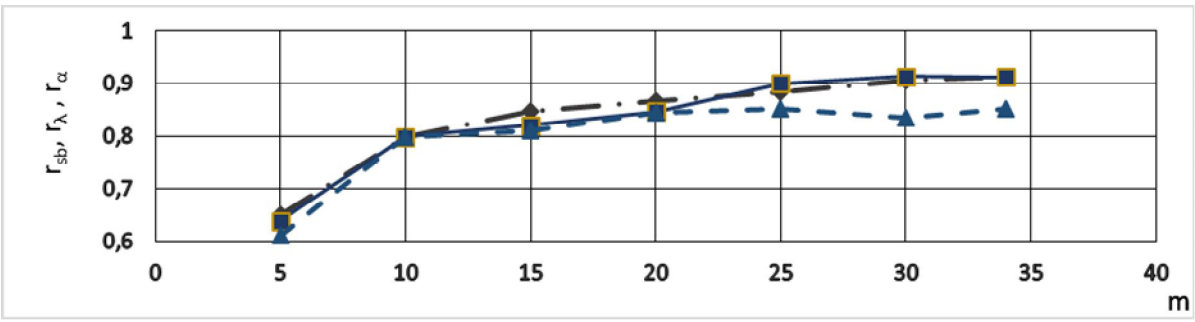
б



в



г



д

Рис. 1. Значение коэффициентов Спирмена-Брауна, лямбда Гутмана и α-Кронбаха для тестов различной длины и размеров выборки: а – 50, б – 100, в – 200, г – 500, д - 1000

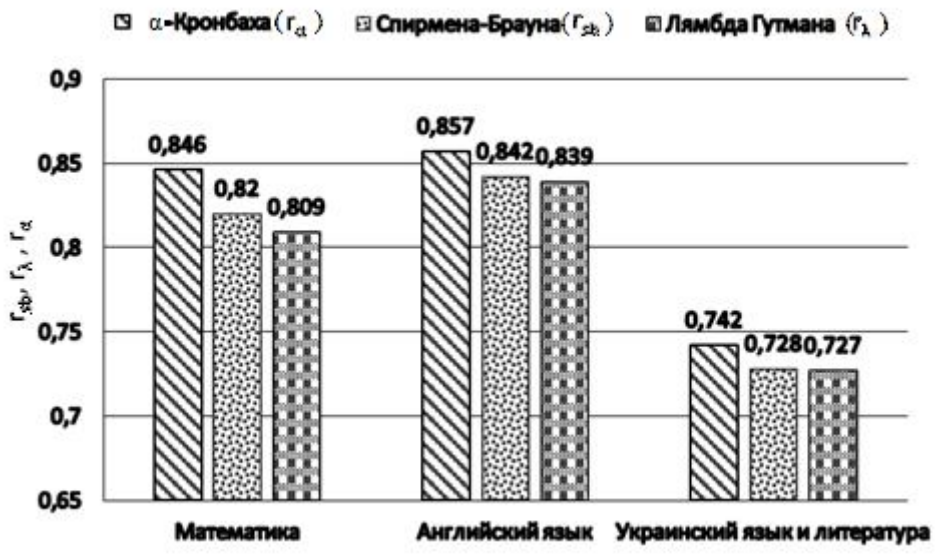


Рис. 2. Значения коэффициентов надежности для тестов по различным предметам

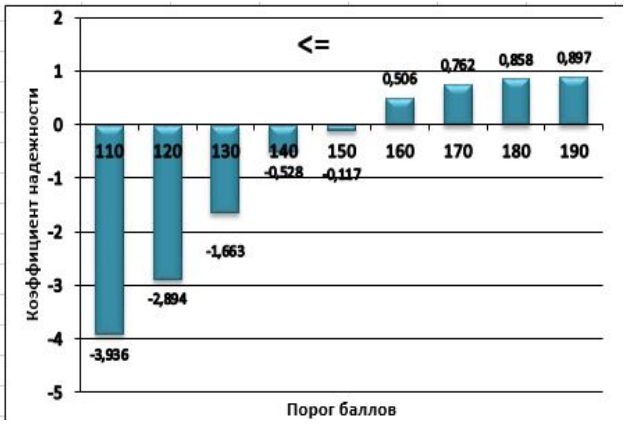


Рис. 3. Распределение α-Кронбаха для групп і ≤

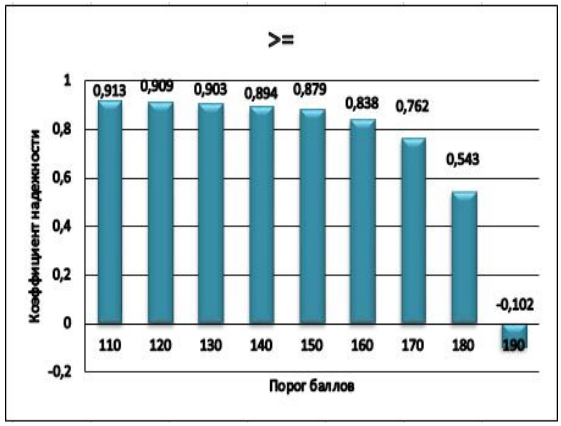


Рис. 4. Распределение α-Кронбаха для групп і ≥

Оценка надежности теста не является единственным методом оценивания качества теста. Требуется также оценивать сложность, валидность, дискриминативность и другие показатели, которые помогут сконструировать качественный тест, что, в свою очередь, приведет к получению достоверных и объективных результатов об уровне знаний студентов.

Литература

1. Nohara, D. A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA) [Електрон. Ресурс] / David Nohara // National Center for Education Statistics. – 2001. – № 07. – Режим доступа: http://nces.ed.gov/pubs2001/200107.pdf. – 5.09.2014.

2. Handout on reliability [Електрон. ресурс] / Laurier's Athletes of the Week. – 2013. – Режим доступа: http://web.wlu.ca/bgebotys/book/reliability.pdf. – 5.09.2014.

3. Челышкова, М. Б. Теория и практика конструирования педагогических тестов [Текст] : учеб. пособие / М. Б. Челышкова. – М. : Логос, 2002. – 432 с.

4. Panayides, P. Coefficient Alpha: Interpret with caution [Text] / P. Panayides // Europe`s Journal of Psichology. – 2013. – Vol. 9(4). – P. 687-696.

5. Ким, В. С. Тестирование учебных достижений [Текст] : монография / В. С. Ким. – Уссурийск : Изд-во УГПИ, 2007. – 214 с.

6. Крокер, Л. Введение в классическую и современную теорию тестов [Текст] / Л. Крокер, Дж. Алгина. – М. : Логос, 2010. – 668 с.

7. Starkweather, Dr. J. Step out of the past: Stop using coefficient alpha; there are better ways to calculate reliability [Text] / Dr. J. Starkweather // Research and Statistical Support. – 2012. – Vol. 6. – P. 6 –12.

8. Benton, T. An empirical assessment of Guttman`s Lambda 4 reliability coefficient [Text] / T. Benton // Cambridge assessment. – 2013. – Vol. 4. – P. 1-9.

9. Callender, J. C. An Empirical Comparison of Coefficient Alpha, Guttman's Lambda - 2, and MSPLIT Maximized Split-Half Reliability Estimates [Text] / J. C. Callender // Journal of Educational Measurement. – 1979. – Vol. 16(2). – P. 89-99.

10. Анализ надежности в SPSS [Електрон. ресурс] / Центр дистанционной поддержки обучения РГПУ им. А. И. Герцена. – Режим доступа: http://moodle.herzen.spb.ru. – 5.09.2014.

11. Нейман, Ю. М. Введение в теорию моделирования и параметризации педагогических тестов [Текст] / Ю. М. Нейман, В. А. Хлебников. – М. : Прометей, 2000. – 169 с.

12. Tavakol, M. Making sense of Cronbach's alpha [Text] / M. Tavakol, R. Dennick // International Journal of Medical Education. – 2011. – № 2. – P. 53-55.

Поступила в редакцию 5.09.2014, рассмотрена на редколлегии 18.11.2014

Рецензент: д-р техн. наук, проф., зав. каф. охраны труда, стандартизации и сертификации Р. М. Трищ, Украинская инженерно-педагогическая академия, Харьков.

ОБГРУНТУВАННЯ ВИБОРУ МЕТОДІВ ВИМІРЮВАННЯ НАДІЙНОСТІ ПЕДАГОГІЧНИХ ТЕСТІВ

М. С. Мазорчук, О. О. Соколова, В. С. Добряк, А. А. Сухобрус

В даній роботі пропонується аналіз існуючих методів і моделей оцінки надійності педагогічних тестів та обгрунтування вибору коефіцієнтів для оцінки надійності, які максимально відображають якість тесту. Розглянуто три основні методи аналізу надійності: метод α-Кронбаха, Спірмена-Брауна і лямбда Гуттмана. При проведенні розрахунків використовувалися дода тки статистичного аналізу даних: SPSS і MS Excel. Проведені розрахунки показали, що оцінка надійності тесту не є єдиним методом оцінювання якості тесту. Також вимагається оцінювати складність, валідність, дискримінативність та інші показники, які допоможуть сконструювати якісний тест.

Ключові слова: аналіз надійності тесту, якість педагогічного тесту, α-Кронбаха, лямбда Гутмана, коефіцієнт Спірмена Брауна.

RATIONALE CHOICE RELIABILITY PEDAGOGICAL METHODS OF MEASUREMENT OF TEST

M. S. Mazorchuk, O. O. Sokolova, V. S. Dobriak, A. A. Suhobrus

This paper proposes an analysis of existing methods and models for assessing reliability of pedagogical tests and a rationale for the selection of coefficients that are used to assess test reliability and best reflect the quality of the test. Three basic methods of reliability analysis (α-Cronbach's, Spearman-Brown and Guttman lambda methods) have been considered in the research. The applications of statistical data analysis: SPSS and MS Excel have been used in calculations. The calculations have shown that the assessment of test reliability is not the only method of evaluating the quality of the test. It is also required to assess complexity, validity, discriminative ability and other indicators that help to construct a qualitative test.

Keywords: analysis of the reliability of the test, the quality of teaching the test, α-Cronbach's, Guttman lambda, coefficient Spearman Brown.

Мазорчук Мария Сергеевна – канд. техн. наук, доц., доц. каф. информатики, Национальный аэрокосмический университет им. Н. Е. Жуковского «ХАИ», Харьков, Украина, e-mail: mazorchuk_mary@inbox.ru.

Соколова Елена Олеговна – аспирант кафедры информатики, Национальный аэрокосмический университет им. Н. Е. Жуковского «ХАИ», Харьков, Украина, e-mail: mango26.88@mail.ru.

Добряк Виктория Сергеевна - ассистент кафедры информатики, Национальный аэрокосмический университет им. Н. Е. Жуковского «ХАИ», Харьков, Украина, e-mail: viktoriya--13@mail.ru.

Сухобрус Анатолий Андреевич – канд. техн. наук, доцент, профессор кафедры авиационных приборов и измерений, Национальный аэрокосмический университет им. Н. Е. Жуковского «ХАИ», Харьков, Украина, e-mail: anik328@rambler.ru.