УДК 378.14

КАЧЕСТВЕННАЯ ОЦЕНКА ОБУЧАЮЩИХ МАТЕМАТИЧЕСКИХ ТЕСТОВ

И.Г. Устинова[1], Е.Г. Лазарева[2]

Национальный исследовательский Томский политехнический университет
634050, г. Томск, пр. Ленина, 30
[1] E-mail: igu@tpu.ru
[2] E-mail: lazareva@math.tsu.ru

В настоящее время в учебном процессе широкое распространение получило применение современных компьютерных средств и информационных технологий. Основная цель компьютерного тестирования смещается от оценивания результатов обучения непосредственно к самому процессу обучения. Критерии эффективности обучающих математических тестов еще не изучены должным образом. Целью нашей работы является нахождение и изучение таких критериев. Мы создали 21 обучающий тест по курсу математического анализа и определили следующие критерии эффективности: информативность, валидность, надежность, дискриминативность. Основываясь на статистических данных, полученных в результате тестирований, мы выяснили, что наши тесты информативны, дискриминативны, не всегда имеют высокую надежность, а их валидность требует внешней экспертной оценки.

***Ключевые слова**: критерии эффективности, обучающие математические тесты.*

Развитие современных педагогических технологий привело к повсеместному внедрению в процесс обучения заданий в тестовой форме. Как правило, они рассматриваются как инструмент контроля результатов обучения. Мы в своей работе используем тесты с целью обучения студентов, а не с целью контроля процесса обучения. С помощью заданий в тестовой форме мы знакомим обучающихся с новыми методами, знаниями, идеями и стараемся стимулировать и интенсифицировать процесс познания. В наших тестах присутствуют задания всех типов: выбор одного ответа или нескольких, задания на соответствие, на классификацию, на упорядочивание, задания с открытым ответом. Основными критериями для включения задания в обучающий тест являются его новизна, нестандартность, но при этом доступность решения в рамках изучаемого курса и темы. Поэтому тесты получаются не вполне од-

[1] Ирина Георгиевна Устинова, кандидат технических наук, доцент кафедры «Высшая математика».
[2] Елена Геннадьевна Лазарева, кандидат физико-математических наук, доцент кафедры «Общая математика».

нородными, не охватывают весь изучаемый материал и не могут использоваться с целью контроля знаний.

Цель представленной работы – оценить возможности использования типичных критериев эффективности теста достижений для оценки результатов обучающего тестирования. Объекты исследования: итоги тестирований по курсу математического анализа (21 тест, созданный на платформе «Айрен» [1]).

В.С. Аванесов в [2] установил близость понятий «эффективность» и «оптимальность», причем «последнее трактуется как наилучшее из возможных вариантов с точки зрения удовлетворения нескольким критериям, взятым поочередно или вместе». Введем в рассмотрение критерии эффективности тестов:

1) информативность;

2) надежность;

3) валидность;

4) дискриминативность.

Исследование истоков проблемы. В течение многих лет тесты широко используются в педагогической практике ([2 – 6] и др.) во всем мире. В настоящее время кроме функции контроля обучения тесты все чаще используются с целью обучения [6–8, 10]. Обучающий тест в математике, согласно [6], есть инструмент, позволяющий диагностировать правильность хода решения задачи, вплоть до получения окончательного ответа. Мы расширили понятие обучающего теста, добавив в цели тестирования стимулирование познавательной деятельности обучающихся. Использование обучающих тестов в педагогической практике затруднено из-за отсутствия ряда методических рекомендаций, к которым относится оценка качества обучающих тестов, используемых в учебном процессе. Недостаточное развитие научных и методологических подходов к вопросу выбора параметров качества обучающих тестов появилось из-за различных толкований «качества обучающего теста».

Одним из аспектов качества теста является его эффективность. Перед тем как сделать отбор критериев эффективности обучающих математических тестов, мы провели анализ процесса отбора соответствующих критериев эффективности психологических тестов, так как эта область знания хорошо развита. Например, в [5] в качестве критериев эффективности психологических тестов выбраны: использование шкалы интервалов, надежность, достоверность, дискриминативность, наличие нормативных данных и критериев, установленных экспертами. В педагогической диагностике [11] имеет большое значение качество измерений. Здесь сделать оценку эффективности теста позволяют объективность, надежность, валидность. Вопросы оценки качества в области контролирующего тестирования рассмотрены в работах [12, 13]. Мы предлагаем в качестве критериев эффективности обучающего математи-

ческого теста выбрать информативность, валидность, надежность и дискриминативность. Обучающие математические тесты должны быть информативными, то есть результаты тестирования должны быть связаны со шкалой измерения и, следовательно, легко может быть осуществлен статистический анализ. Обучающие тесты должны быть валидными, так как должны измерять именно то, для чего они созданы. Обучающие тесты должны быть надежными, то есть должна быть возможность получения одинаковых результатов у испытуемых в различных случаях. Обучающие тесты должны быть дискриминативными, то есть обладать способностью разделять испытуемых на отдельные группы в зависимости от уровня выполнения заданий. Мы не рассматриваем объективность, ибо наши тесты не зависят ни от настроения педагога, ни от методов и средств контроля. В дальнейшим мы рассмотрим наличие нормативных данных и критериев, установленных экспертами.

Хорошо известно значительное число критериев надежности теста [5, 14]. Например, в качестве такого критерия можно использовать коэффициент корреляции Пирсона между двумя параллельными тестами на одной и той же выборке студентов. Также в качестве критерия надежности можно использовать коэффициент корреляции результатов экспертных оценок и результатов тестирования. Как правило, в качестве критерия надежности тестового задания используют коэффициент корреляции Спирмена – Брауна [5, 14] и коэффициент надежности Гутмана [15], а также некоторые их модификации. Имеет широкое применение для расчета надежности так называемая формула KR-20 [14], названная так по именам ее основателей Ф. Кудера и М. Ричардсона (формула номер 20 в публикации).

Цели исследования. Целями настоящей работы являются: формулирование критериев эффективности обучающих математических тестов; оценка актуальности критериев для оценки результатов тестирования. Объектом исследования являются результаты тестирований по курсу математического анализа (21 тест, разработанный на платформе «Айрен»). Далее рассмотрим отдельно каждый из выбранных критериев:

– информативность;

– надежность;

– валидность;

– дискриминативность.

Исследование критериев эффективности

Информативность. Обучающие тесты должны быть информативными, то есть такими, которые обеспечивают возможность соотнесения количественной оценки за выполнение теста со шкалой измерений, и, соответственно, быть пригодными для быстрой статистической обработки результатов обследования.

Любое научное исследование начинается с того, что исследователь фиксирует, насколько ярко выражено интересующее его свойство (или свойства) у объекта или объектов исследования, как правило, при помощи чисел (количественная характеристика). Таким образом, следует различать объекты исследования (в нашем случае это тесты, состоящие из тестовых заданий, и результаты тестирований), их свойства (то, что интересует нас как исследователей, составляет предмет изучения – эффективность тестов) и признаки, отражающие в числовой шкале проявление свойств. Поэтому оценку эффективности тестов следует начать с определения шкалы измерений как инструмента статистической обработки результатов тестирования. Пусть A – некоторое множество объектов, а $\{P_i\}_{i=1}^{m}$ – набор отношений на этом множестве. Множество A вместе с заданной на нем системой отношений $\{P_i\}_{i=1}^{m}$ называется системой с отношениями и обозначается $U =< A, \{P_i\}_{i=1}^{m} >$.

Под k-мерной шкалой будем понимать гомоморфизм f эмпирической системы с отношением $U =< A, \{P_i\} >$ в k-мерную числовую систему с отношениями $V =< R^k, \{S_i\} >$ [16]. Таким образом, шкала – это тройка (U, V, f), где f – гомоморфизм из U в V.

Существует множество шкал наименований, когда например, числа используются как имена объектов исследования. Шкала наименований дает информацию о том, эквивалентны два объекта или нет. Например, результаты тестирования студентов, набравших одинаковое количество баллов, эквивалентны, тогда как сами испытуемые – разные. Шкала наименований используется только для того, чтобы отнести испытуемых к какому-либо классу, например к классу прошедших тестирование. Порядковые шкалы – это шкалы, в которых эмпирическая система есть система с заданным отношением порядка (объекты упорядочены). Например, классификация испытуемых производится по среднему баллу. Недостатком такой шкалы является то, что не учитываются значения разностей между градациями. Можно выделить шкалу интервалов, в которой значения разностей во всех точках данной шкалы равны, начало отсчета произвольно, а единица измерения задана. Значения, полученные по интервальной шкале, инвариантны относительно группы аффинных преобразований. Шкала отношений – это шкала, в которой начало отсчета известно, а единица измерений выбирается по усмотрению исследователя.

Мы используем в своей работе шкалу интервалов (k = 1), так как мы полагаем принципиально важным то, что к экспериментальным данным, обработанным по шкале интервалов, применимо достаточно большое число статистических методов исследований. Испытуемые получают оценку за каждый тест в процентах от полностью правильного ответа на все представленные

вопросы. При этом некоторые тестовые задания предполагают «мягкое оценивание». Например, если дан вопрос с выбором нескольких (k) правильных вариантов из данных ответов, а испытуемый выбрал не все необходимые варианты ($k_1 < k$), то он получит в свою оценку соответствующую часть (k_1/k) от того процента, который мог бы получить ($1/n$, где n – число заданий в тесте), если бы ответил на этот вопрос полностью правильно. Таким образом, в результате тестирования мы имеем данные в процентах, которые необходимо разделять с помощью интервалов.

Надежность. Под надежным тестом будем понимать тест, который дает одинаковые показатели для одного и того же испытуемого при гипотетическом повторном тестировании, то есть тест, в котором результаты тестирования не зависят от всевозможных случайных факторов. Такая надежность называется ретестовой (test-retest reliability) [5, 14]. Для нахождения значения этого показателя вычисляется коэффициент корреляции результатов тестирования одного и того же студента. Для нахождения надежности мы использовали формулу Спирмена – Брауна $r_t' = \frac{2r_t}{1+r_t}$ при расщеплении теста на две части. Здесь r_t' − исправленный коэффициент надежности, а r_t − коэффициент надежности (коэффициент корреляции Пирсона), найденный по половинкам расщепленного теста. Считается, что наименьшим удовлетворительным значением для ретестовой надежности является значение 0,7 [17]. Исследовав 21 тест по курсу математического анализа, мы выяснили, что большинство из них – тесты с достаточной надежностью (см. таблицу). Низкие показатели надежности объясняются, по нашему мнению, двумя основными причинами: небольшое количество заданий в тесте (7–8) и недостаточная ясность формулировок для испытуемых (новизна заданий). Количество заданий в тесте не подлежит изменению, так как с увеличением количества заданий увеличивается продолжительность тестирования, что неудобно по техническим причинам, а также приводит к ослаблению когнитивных функций, которые необходимы для решения обучающих заданий. Неясность формулировок, на которую иногда жаловались наши студенты, являются неотъемлемой частью обучающих тестов в нашем понимании: понять задание, содержащее новую математическую терминологию, – значит уже чему-то научиться. Поэтому значительное увеличение показателя надежности для обучающих тестов вряд ли возможно.

Валидность. Тест называется валидным, если он измеряет то, для чего предназначен [17]. Существует несколько видов валидности: при очевидной (внешней) валидности у испытуемых складывается впечатление, что тест измеряет именно то, для чего он создан; конкурентная валидность оценивается

по корреляции результатов тестирования с результатами других тестов, предназначенных для решения аналогичных задач; прогностическая валидность определяется при помощи корреляции между показателями теста и некоторым критерием, характеризующим то же самое свойство у испытуемых, но в более позднее время: например, корреляция между показателями тестирования в первом семестре и успеваемостью данного студента во втором семестре. К тестам достижений в основном применяется содержательная валидность. Однако валидность наших обучающих тестов не очевидна, так как они не охватывают весь изучаемый материал и часто фокусируются на деталях. Мы старались подготовить тесты так, чтобы основные понятия и факты теории, методы практических занятий использовались при решении тестовых заданий, но полностью охватить весь материал невозможно. Кроме того, многие задания по математике предполагают выполнение большой последовательности действий и поэтому не подходят для обучающего теста. Поэтому валидность наших тестов может быть оценена только с качественной точки зрения, путем независимой профессиональной экспертизы (см. [12]).

Дискриминативность. Дискриминативность означает различительную способность теста (способность отделять испытуемых с высоким баллом по тесту от тех, которые набрали низкий балл) [18]. Одной из целей разработчика тестов является достижение хорошего распределения показателей. Произвести оценку дискриминативности теста можно при помощи коэффициента дискриминации, коэффициента корреляции Гилфорда, коэффициента дельта δ Фергюсона [14]. Именно последний из перечисленных коэффициентов мы использовали при исследовании наших тестов. Дискриминативность, измеряемая показателем дельта Фергюсона, принимает максимальное значение $\delta = 1$ при равномерном распределении [6]. Дельта Фергюсона находится по формуле $\delta = \dfrac{(n+1)\left(N^2 - \sum_{i=1}^{k} w_i^2\right)}{nN^2}$, где N – количество испытуемых, n – количество вопросов теста, $w_i, i = \overline{1,k}$ – количество итоговых баллов, попавших в каждый из k интервалов шкалы. Если $\delta = 0$, то все испытуемые получили одинаковое количество баллов, то есть тест не является дискриминативным. При создании обучающих тестов равенство $\delta = 0$ означает, что по сути тест не является обучающим, ибо все испытуемые одинаково правильно ответили на задания теста, то есть материал, представленный в тесте, уже усвоен.

В таблице приведены статистические характеристики результатов тестирований по разработанным нами тестам. При исследовании типа распределения обучающих тестов использовался статистический критерий Пирсона [19]. В частности, из этой таблицы следует, что наши тесты являются дискриминативными.

Статистические характеристики обучающих тестирований

Тест	Число тестовых заданий	Количество тестируемых	Выборочное среднее	Выборочное среднее квадратическое отклонение	Тип распределения итоговых баллов	Надежность r_t'	Дельта Фергюсона δ
1. «Множества»	7	67	46,19	22,71	Нормальное	0,690	0,987
2. «Вещественные числа»	7	56	46,25	26,8	Равномерное	0,717	1
3. «Числовые функции	7	51	52,94	28,63	Равномерное	0,578	1
4. «Предел последовательности – 1»	7	52	41,92	29,06	Равномерное	0,752	1
5. «Предел последовательности – 2»	8	43	58,37	24,72	Нормальное	0,747	0,926
6. «Предел последовательности – 3»	7	48	45	25,2	Нормальное	0,676	0,909
7. «Предел функции – 1»	7	49	51,84	29,45	Равномерное	0,734	1
8. «Предел функции – 2»	8	39	48,59	24,18	Нормальное	0,837	0,926
9. «Непрерывность функции»	8	39	54,19	27,62	Равномерное	0,714	1
10. «Производная – 1»	8	31	72,22	23	Не определяется	0,833	0,812
11. «Производная – 2»	8	31	56,76	25,86	Равномерное	0,587	1
12. «Производная –3»	7	36	61,53	22,11	Нормальное		0,884
13. «Комплексные числа»	8	45	62,11	25,92	Равномерное	0,827	1
14. «Неопределенный интеграл – 1»	8	52	54,81	26,46	Равномерное	0,720	1
15. «Неопределенный интеграл – 2»	7	43	52,09	27,83	Равномерное	0,747	1
16. «Неопределенный интеграл – 3»	8	41	63,17	26,93	Равномерное	0,690	1
17. «Неопределенный интеграл – 4»	7	44	53,75	27,24	Равномерное	0,751	1
18. «Определенный интеграл – 1»	7	38	54,08	25,41	Нормальное	0,392	0,844
19. «Определенный интеграл – 2»	7	33	49,09	27,95	Равномерное	0,647	1
20. «Определенный интеграл – 3»	7	29	59,83	26,57	Равномерное	0,720	1
21. «Определенный интеграл – 4»	7	30	50,33	26,17	Нормальное	0,761	0,907

Обсуждение результатов. В таблице приведены статистические характеристики результатов обучающих тестирований. Статистическая обработка данных возможна вследствие использования шкалы интервалов. В первом столбце таблицы содержится название (тема) теста, во втором – количество тестовых заданий. Следует отметить, что число испытуемых (третий столбец) меняется от теста к тесту. Сначала это значение уменьшается из-за снижения интереса студентов к тестовой форме обучения, а затем в силу усиления мотивации со стороны преподавателя (уменьшение количества заданий на контрольной работе, освобождение от контрольной работы, уменьшение количества вопросов на экзамене и т. д.) увеличивается. Выборочное среднее (сумма баллов, полученная студентами за тест, разделенная на количество обследованных) – довольно простая характеристика теста. Так, если средний балл близок к 100, то тест ничего не стоит, ничему не учит. А если средний балл значительно меньше 50, то тест труден для этой группы студентов. В этой ситуации возможны следующие шаги: либо адаптировать тест для данных студентов, либо обсудить все неясные вопросы и провести повторное тестирование. Пятый столбец («Выборочное среднее квадратическое отклонение») содержит отклонение балла от выборочного среднего. Шестой столбец содержит тип распределения тестовых баллов. В 62 % случаев это равномерное распределение и в 38 % − нормальное. То, что в большинстве случаев распределение тестовых баллов получилось равномерным, говорит о том, что наши тесты являются дискриминативными. Действительно, как можно заметить из таблицы, соответствующее равномерному закону распределения значение дельта Фергюсона (последний столбец) равно единице. Это является еще одним доказательством дискриминативности тестов.

Надежность тестов (седьмой столбец таблицы) рассчитывается по формуле Спирмена – Брауна. Для этого мы нашли коэффициент корреляции между двумя частями теста (средний балл по четным и средний балл по нечетным вопросам для каждого испытуемого), а затем рассчитали исправленный коэффициент надежности.

Заключение. В результате проведенного исследования итогов тестирований мы установили, что рассматриваемые нами тесты являются информативными, их валидность следует оценивать качественно, а дискриминативность не вызывает сомнений. Надежность тестов не всегда достаточно высока. Обучение с помощью решения заданий теста, согласно [8], – это процесс, в начале которого студент знает и умеет меньше, чем в конце. Поэтому мы считаем, что стремиться к повышению повторяемости результата (надежности теста) и придавать этому фактору большое значение при оценке эффективности обучающих тестов не стоит.

Таким образом, мы установили применимость таких критериев эффективности, как информативность, надежность, дискриминативность и валидность, к изучению эффективности обучающих математических тестов. Мы выяснили причины невысокой надежности обучающих тестов – неясность формулировок, небольшое число заданий. Также мы объяснили, почему исследование содержательной валидности тестов невозможно без внешней экспертной оценки.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Программа тестирования знаний «Айрен». URL: http://irenproject.ru/
2. *Аванесов В.С.* Композиция тестовых заданий. – М.: Центр тестирования, 2002. – 240 с.
3. *Ким В.С.* Тестирование учебных достижений. – Уссурийск: Изд-во УГПИ, 2007. – 214 с.
4. *James E. Carlson, Matthias von Davier.* Item Response Theory. ETS Research Report Series. Volume 2013, Issue 2, pages i–69, December 2013.
5. *Клайн П.* Справочное руководство по конструированию тестов. – Киев: ПАН Лтд., 1994. – 288 с.
6. *Майоров А.Н.* Теория и практика создания тестов для системы образования. – М.: Интеллект-Центр, 2001. – 296 с.
7. *Углев В.А.* Обучающее компьютерное тестирование // Теоретические и прикладные вопросы современных информационных технологий: Мат-лы VIII Всероссийск. науч.-техн. конф. – Улан-Удэ: ВСГТУ, 2007. – С. 312-316.
8. *Лазарева Е.Г., Устинова И.Г., Подстригич А.Г.* Использование тестирующих программ в процессе обучения высшей математике // Вестник Томского государственного педагогического университета (Tomsk State Pedagogical University Bulletin). – 2012. – Вып. 7 (122). – С. 217–222.
9. *Сеногноева Н.А.* Обучающие тесты: Инновационная педагогическая технология. – Нижний Тагил: Нижнетагильская государственная социально-педагогическая академия, 2005. – 155 с.
10. *Кадневский В.М.* Из истории создания и применения тестов для системы образования // Педагогическая диагностика. – 2003. – № 3. – С. 39-50.
11. *Ингекамп К.* Педагогическая диагностика. – М.: Педагогика, 1991. – 240 с.
12. *Челышкова М.Б.* Теория и практика конструирования педагогических тестов. – М.: Логос, 2002. – 432 с.
13. *Черепанов В.С.* Экспертные оценки в педагогических исследованиях. – М.: Педагогика, 1989. − 152 с.
14. *Kuder G.F., Richardson M.W.* The theory of the estimation of test reliability // Psychometrika. – 1937. − V. 2. − № 3. − P. 151-160.
15. *Guttman L.* A basis for analyzing test-retest reliability // Psychometrika. – 1945. − V. 10. − P. 255-282.
16. *Пфанцагль И.* Теория измерений. – М.: Мир, 1976. – 165 с.

17. *William M.K.* Trochim. Measurement Validity Types. Cornell University. The Research Methods Knowledge Base. URL: www.socialresearchmethods.net/kb/measval.php
18. *Белоус В.В., Домников А.С., Карпенко А.П.* Тестовый метод контроля качества обучения и критерии качества образовательных тестов. Обзор // Наука и образование. – 2011. − № 4. – С. 1-28.
19. *Уилкс С.* Математическая статистика. – М.: Наука, 1967. – 632 с.

Поступила в редакцию 17.03.16;
в окончательном варианте 21.03.16

UDC 378.14

QUALITY EVALUATION OF LEARNING MATH TESTS

I.G. Ustinova[1], E.G. Lazareva[2]

National Research Tomsk Polytechnic University
30, Lenin avenue, Tomsk, 634050
[1] E-mail: igu@tpu.ru
[2] E-mail: lazareva@math.tsu.ru

In present days modern computer tools and information technologies implementation in educational process are widespread. The main purpose of computer-based testing shifted from the assessment instrument of learning outcomes to the purpose of teaching. Performance criteria of learning mathematical tests are not yet studied properly. The objectives of the present work are: to formulate the performance criteria of learning math tests; to evaluate the relevance of achievement test performance criteria for the evaluation of learning test results. Learning math tests should be informative, valid, reliable, discriminative. The research subjects are test results in mathematical analysis 21 tests. Several methods of statistics are used (sample mean, sample standard deviation, Pearson's correlation coefficient, etc.) to study these criteria. We found out our tests informative, discriminative criterion, valid, nevertheless, we describe the tests at hand are not always have a high reliability.

Keywords*: performance criteria, learning math tests.*

REFERENCES

1. Programma testirovaniya znaniy "Ayren" [Testing program knowledge "Iren"]. http://irenproject.ru/ (accessed February 1, 2016).
2. *Avanesov V.S.* Kompoziciya testovykh zadaniy [The composition of the test tasks]. Moskow, Centr testirovaniya Publ., 2002. 240 p.
3. *Kim V.S.* Testirovanie uchebnykh dostizheniy [Testing of educational achievements]. Ussuriysk, UGPI Publ., 2007. 214 p.
4. *James E.* Carlson, Matthias von Davier. Item Response Theory. ETS Research Report Series, 2013, vol. 2013, Issue 2. 1-69 pp.

[1] Irina G. Ustinova, Cand. of Tech. Sci., Associate Professor of "Higher Mathematics".
[2] Elena G. Lazareva, Cand. of Phys. and Math. Sci., Associate Professor of "General Mathematics".

5. *Klain P.* Spravochnoe rukovodstvo po konstruirovaniyu testov [Reference design test]. Kiev, PAN Ltd. Publ., 1994. 288 p.
6. *Mayorov A.N.* Teoriya I praktika sozdaniya testov dlya sistemy obrazovaniya [Theory and practice of creating tests for the education system]. Moskow, Intellekt-centr Publ., 2001. 296 p.
7. *Uglev V.A.* Obuchayushchee komp'yuternoe testirovanie [Training computer testing]. Trudy konferentsii 2007 "Teoreticheskie I prikladnye voprosy sovremennykh informacionnykh tekhnologiy" [Materials of the Conference of 2007 "Theoretical and applied issues of modern information technologies"]. Ulan-Ude, VSGTU Publ., 2007. 312-316 pp.
8. *Lazareva E.G., Ustinova I.G., Podstrigich A.G.* Ispol'zovanie testiruyushchikh program v processe obucheniyavysshey matematike [Using testing programs in learning higher mathematics]. Vestnik Tomskogo gosudarstvennogo pedagogicheskogo universiteta. – Tomsk State Pedagogical University Bulletin, 2012, vol. 7(122). 217-222 pp.
9. *Senognoeva N.A.* Obuchayushchie testy: Innovacionnaya pedagogicheskaya tekhnologiya [Educational tests: innovative educational technology]. Nizhniy Tagil, Nizhnetagil'skaya gosudarstvennaya social'no-pedagogicheskaya akademiya Publ., 2005. 155 p.
10. *Kadnevskiy V.M.* Iz istorii sozdaniya I primeneniya testov dlya sistemy obrazovaniya [From the history of creation and application of tests for the education system]. Pedagogicheskaya diagnostika, 2003, № 3. 39-50 pp.
11. *Ingekamp K.* Pedagogicheskaya diagnostika [Pedagogical diagnostics]. Moskow, Pedagogika Publ.,1991. 240 p.
12. *Chelyshkova M.B.* Teoriya I praktika konstruirovaniya pedagogicheskikh testov [Theory and practice of designing of pedagogical tests]. Moskow, Logos Publ., 2002. 432 p.
13. *Cherepanov V.S.* Ekspertnye ocenki v pedagogicheskikh issledovaniyakh [Expert assessments in educational research]. Moskow, Pedagogika Publ.,1989. 152 p.
14. *Kuder G.F., Richardson M.W.* The theory of the estimation of test reliability. Psychometrika, 1937, vol. 2, no 3. 151-160 pp.
15. *Guttman L.* A basis for analyzing test-retest reliability. Psychometrika,1945, vol. 10. 255-282 pp.
16. *Pfancagl' I.* Teoriya izmereniy [Theory of measurement]. Moskow, Mir Publ., 1976. 165 p.
17. *William M.K.* Trochim. Measurement Validity Types. Cornell University. The Research Methods Knowledge Base. http://www. socialresearchmethods.net/kb/measval.php (accessed February 1, 2016).
18. *Belous V.V., Domnikov A.S., Karpenko A.P.* Testovyy metod kontrolya kachestva obucheniya I kriterii kachestva obrazovatel'nykh testov [the testing method of the control of quality of education and criteria of quality of educational tests]. Nauka I obrazovanie, 2011, no 4. 1-28 pp.
19. *Uilks S.* Matematicheskaya statistika [Mathematical statistics]. Moskow, Nauka Publ., 1967. 632 p.

Original article submitted 17.03.16;
revision submitted 21.03.16