КОНЦЕПТ
научно-методический электронный журнал
Буров А. В. Применение однофакторного дисперсионного анализа к оценке надежности тестовых материалов Единого государственного экзамена // Концепт. – 2014. – № 12 (декабрь). – ART 14341. – 0,3 п. л. – URL: http://e-koncept.ru/2014/14341.htm. – Гос. рег. Эл № ФС 77-49965. – ISSN 2304-120X.
ART 14341 УДК 371.261:519.254

***Буров Андрей Васильевич,***
*магистрант ФГБОУ ВПО «Смоленский государственный университет», г. Смоленск*
burov_andrei@inbox.ru

# Применение однофакторного дисперсионного анализа к оценке надежности тестовых материалов Единого государственного экзамена

***Аннотация.*** *Статья посвящена изучению некоторых возможностей применения однофакторного дисперсионного анализа для оценки надежности тестовых материалов Единого государственного экзамена. Вопрос об объективности тестирования как способа проверки знаний и умений обучающегося актуален как для системы образования, так и для общества в целом. Для оценки надежности тестовых материалов Единого государственного экзамена применяется один из вариантов измерения надежности тестов с помощью аппарата статистической обработки данных, и показывается применение предложенной методики к результатам выполнения учащимися тестовых заданий по русскому языку.*
***Ключевые слова:*** *Единый государственный экзамен, оценка, надежность, тест, статистическая обработка данных.*
***Раздел:*** *(01) педагогика; история педагогики и образования; теория и методика обучения и воспитания (по предметным областям).*

Важнейшим аспектом любой педагогической деятельности являются оценки, которые выставляют преподаватели и разного рода экзаменаторы своим ученикам, абитуриентам, студентам и пр. Последствия таких оценок могут быть самыми различными – от чисто морального эффекта до определения судьбы человека. Тем не менее все прекрасно понимают, что оценки эти субъективны и часто приблизительны. Даже в рамках такой малочувствительной системы оценок, какой является традиционная для России пятибалльная (а по существу, лишь трехбалльная – «3», «4», «5») система, не удается сформулировать конкретные стандарты, определяющие, за что следует ставить «3», а за что можно ставить «4» или «5». Теоретически полезное увеличение чувствительности шкалы вряд ли было бы оправданным при существующем порядке проведения текущего контроля успеваемости и экзаменов, так как на практике привело бы лишь к увеличению влияния субъективизма и его последствий. Проведение контрольных мероприятий в письменной форме требует существенных временных и других затрат, но нисколько не меняет сути дела [1].

Проблема измерения и оценивания результатов обучения является одной из самых важных в педагогической теории и практике. Решение этой проблемы необходимо для оценки эффективности педагогических инноваций и технологий.

Сегодня в качестве *инновационных средств* используют тестирование, рейтинговую и модульную системы оценки качества знаний, учебные портфолио, мониторинг качества.

Тестирование является одной из наиболее технологичных форм проведения автоматизированного контроля с управляемыми параметрами качества. В этом смысле ни одна из известных форм контроля знаний учащихся с тестированием сравниться не может. Тесты контроля уровня знаний применяются на всех этапах дидактического процесса. С их помощью эффективно обеспечивается предварительный, текущий, тематический и итоговый контроль знаний, умений, учет успеваемости, учебных достижений.

КОНЦЕПТ
научно-методический электронный журнал
ART 14341 УДК 371.261:519.254
Буров А. В. Применение однофакторного дисперсионного анализа к оценке надежности тестовых материалов Единого государственного экзамена // Концепт. – 2014. – № 12 (декабрь). – ART 14341. – 0,3 п. л. – URL: http://e-koncept.ru/2014/14341.htm. – Гос. рег. Эл № ФС 77-49965. – ISSN 2304-120X.

Однако не все тесты могут дать желаемый результат. Пользоваться необходимо соответствующими тестовыми измерителями, разработанными и проанализированными в соответствии с правилами и требованиями тестологии, на уровне мировых стандартов. При этом в настоящее время такой тестовой продукции пока слишком мало. В нашей стране только создаются службы сертификации тестовых материалов. Недостаточно квалифицированных специалистов, способных обеспечить высокое качество создаваемых тестов, в связи с чем целесообразно каждому педагогу, школе создавать свой тестовый банк заданий по разделам образовательных программ на основе требований, предъявляемых к данному виду контроля в современной теории конструирования тестов и критериям для проведения внутреннего тестового контроля знаний по всем предметам и направлениям подготовки выпускников.

Вообще,

*тест* (англ. test – проба, испытание, исследование) в психологии и педагогике стандартизированные задания, результат выполнения которых позволяет измерить психофизиологические и личностные характеристики, а также знания, умения и навыки испытуемого;

*педагогический тест* – это инструмент оценивания уровня подготовленности учащихся, состоящий из системы тестовых заданий, стандартизованной процедуры проведения, обработки и анализа результатов;

*педагогическое тестирование* – это форма измерения знаний учащихся, основанная на применении педагогических тестов. Включает в себя подготовку качественных тестов, собственно проведение тестирования и последующую обработку результатов, которая даёт оценку уровня подготовленности тестируемых.

Эксперимент по введению Единого государственного экзамена (ЕГЭ), начатый в 2001 г., открывает новую страницу в развитии отечественной системы образования и имеет инновационный характер не только по замыслу, но и по форме проведения, по масштабам и отсутствию жесткой регламентации со стороны органов власти [2].

Эксперимент имеет две цели: повышение доступности высшего образования и качества среднего школьного образования, реализация которых достигается одновременно за счет совмещения в одной процедуре школьного выпускного экзамена и вступительного экзамена в высшие учебные заведения [3; 4].

К числу основных задач, решаемых с помощью ЕГЭ, можно отнести:

– создание объективной и чувствительной шкалы оценки качества образования;
– повышение доступности профессионального образования;
– снижение психологической нагрузки на выпускников общеобразовательных учреждений;
– совершенствование системы государственного контроля качества общего образования на основе независимой оценки уровня подготовки выпускников.

На ЕГЭ тестовый балл формируется по результатам выполнения заданий контрольно-измерительного материала (КИМ) экзамена. Следовательно, во-первых, оценка должна быть независима от предложенного варианта КИМ (т. е. выставленный балл при выполнении одного варианта заданий должен приближенно совпадать с баллом, полученным при решении другого варианта контрольно-измерительных материалов по этому же предмету на одном и том же уровне знания предмета) и, во-вторых, при выполнении одного и того же варианта тестовых заданий учениками разной степени подготовленности полученные баллы должны различаться, отражая уровень подготовки учащихся.

КОНЦЕПТ
научно-методический электронный журнал
Буров А. В. Применение однофакторного дисперсионного анализа к оценке надежности тестовых материалов Единого государственного экзамена // Концепт. – 2014. – № 12 (декабрь). – ART 14341. – 0,3 п. л. – URL: http://e-koncept.ru/2014/14341.htm. – Гос. рег. Эл № ФС 77-49965. – ISSN 2304-120X.
ART 14341 УДК 371.261:519.254

Под обобщенным термином «уровень подготовленности» понимают уровень обученности испытуемых по указанным разделам, совокупность их умений и соответствующих навыков. Уровень подготовленности участников тестирования является латентным параметром (то есть недоступным для непосредственного измерения), и, чтобы «добраться» до него, необходимо привлечь серьезные научные методы составления тестов и совместной математической обработки результатов тестирования.

Чтобы оценить уровень подготовленности тестируемого в конкретной области знаний, нужно проверить правильность выполнения им достаточно большого количества заданий различной трудности. Это множество заданий можно называть генеральной совокупностью заданий для данной области знания. Понятно, что всякий тест состоит лишь из конечного количества определенных заданий, представляющих собой некоторую выборку из указанной генеральной совокупности.

Таким образом, педагогический тест, в отличие, например, от обычной контрольной работы, можно рассматривать как своеобразный измерительный инструмент определенной разрешающей способности и точности. Нельзя только забывать, конечно, что объект измерения здесь чрезвычайно специфичен, и потому результаты существенно зависят от возможностей разумно формализовать этот объект [5].

Составление качественных тестов требует использования научных методов отбора содержания, теории педагогических измерений, применяемых для проверки соответствия тестов обоснованным критериям качества. Одним из таких критериев является *надежность* тестовых материалов. Под надежностью понимается устойчивость (или согласованность) результатов теста, получаемых при повторном его применении к тем же испытуемым в различные моменты времени, при использовании разных наборов эквивалентных заданий или же при изменении условий обследования [6]. Такое понимание надежности лежит в основе вычисления ошибки измерения отдельного показателя, благодаря чему мы можем предсказывать диапазон случайных колебаний тестового балла у конкретного человека, возникающих, вероятно, под действием посторонних или неизвестных факторов.

Другими словами, мы должны быть уверены, что тест адекватно отражает генеральную совокупность заданий и дает устойчивые результаты при повторном использовании его вариантов. Надежность теста должна показать, в какой мере результаты теста можно считать реальными, а в какой – приписанными влиянию случайных факторов. В качестве количественной меры надежности будем рассматривать коэффициент (надежности) $r \in [0; 1]$, определяющий долю дисперсии «истинного» балла в общей дисперсии.

В данной работе поставлена задача оценить уровень надежности заданий базового уровня контрольно-измерительных материалов по русскому языку Единого государственного экзамена разных лет.

Для определения коэффициента надежности воспользуемся идеей однофакторного дисперсионного анализа.

Контрольно-измерительные материалы по русскому языку состоят из трех частей. *Часть А* – задания с выбором ответа. К каждому из них даны 4 варианта ответов, из которых только один правильный. *Часть В* – задания с ответом в краткой форме, ответ нужно сформулировать самостоятельно. *Часть С* состоит из одного задания и представляет собой небольшую письменную работу по приведенному тексту (сочинение).

Результаты выполнения заданий части А теста по русскому языку занесены в таблицу – матрицу ответов. Результат выполнения каждого задания оценивается по

КОНЦЕПТ
научно-методический электронный журнал
Буров А. В. Применение однофакторного дисперсионного анализа к оценке надежности тестовых материалов Единого государственного экзамена // Концепт. – 2014. – № 12 (декабрь). – ART 14341. – 0,3 п. л. – URL: http://e-koncept.ru/2014/14341.htm. – Гос. рег. Эл № ФС 77-49965. – ISSN 2304-120X.
ART 14341 УДК 371.261:519.254

дихотомическому принципу – ставится 1, если задание выполнено верно, и 0 в противном случае. Ответы всех участников на все задания вариантов образуют прямоугольную таблицу – матрицу размера $n \times k$ ($n$ – количество тестируемых, $k$ – количество заданий). Обозначим матрицу ответов через $A = (a_{ij})$. Рассматривая имеющиеся баллы как реализации случайной величины, выполним дисперсионный анализ таблицы. Основная идея состоит в выделении «факторной дисперсии», порождаемой в данном случае влиянием тем и участников тестирования, и «остаточной дисперсии», обусловленной случайными причинами.

Рассмотрим частичные суммы элементов матрицы *A* по строкам и столбцам. Назовем *первичным баллом i-го участника* сумму элементов *i*-й строки матрицы $A$ (т. е. количество верно выполненных заданий варианта участником с номером *i*):

$$b_i = \sum_{j=1}^{k} a_{ij}\ , i = \overline{1;n}$$

и, аналогично, *первичным баллом j-го задания* назовем сумму элементов столбца с номером *j*, т. е.

$$c_j = \sum_{i=1}^{n} a_{ij}\ , j = \overline{1;k}\ .$$

Разность $n - c_j$ отражает меру трудности *j*-го задания при выполнении его группой из $n$ участников тестирования.

Понятно, что чем выше первичный балл задания, тем оно легче.

Пусть, по определению,

$$SS_{\text{общ}} = \sum_{i=1}^{n}\sum_{j=1}^{k}\left(b_{ij} - \bar{b}\right)^2 -$$

общая сумма квадратов отклонений наблюдаемых значений от их общего среднего;

$$SS_{\text{тем}} = n\sum_{j=1}^{k}\left(\frac{1}{n}b_j - \bar{b}\right)^2 -$$

факторная сумма квадратов отклонений средних значений по столбцам от общего среднего, характеризует рассеяние между темами;

$$SS_{\text{исп}} = k\sum_{i=1}^{n}\left(\frac{1}{k}b_i - \bar{b}\right)^2 -$$

факторная сумма квадратов отклонений средних значений по строкам от общего среднего, характеризует рассеяние между участниками тестирования;

$$S_{\text{ост}} = S_{\text{общ}} - S_{\text{тем}} - S_{\text{исп}} -$$

остаточная сумма квадратов, характеризует внутреннее рассеяние.

По этим результатам легко оценить соответствующие несмещенные оценки дисперсии. Их получают делением сумм квадратов отклонений на соответствующее число степеней свободы:

КОНЦЕПТ
научно-методический электронный журнал
ART 14341 УДК 371.261:519.254
Буров А. В. Применение однофакторного дисперсионного анализа к оценке надежности тестовых материалов Единого государственного экзамена // Концепт. – 2014. – № 12 (декабрь). – ART 14341. – 0,3 п. л. – URL: http://e-koncept.ru/2014/14341.htm. – Гос. рег. Эл № ФС 77-49965. – ISSN 2304-120X.

$$D_{\text{тем}} = \frac{S_{\text{тем}}}{k-1};\ D_{\text{исп}} = \frac{S_{\text{исп}}}{n-1};\ D_{\text{ост}} = \frac{S_{\text{ост}}}{(n-1)(k-1)}.$$

Извлекая квадратный корень из последней оценки, получаем среднеквадратичную ошибку измерений $e$.

Для вычисления коэффициента надежности теста примем во внимание, что задания всех вариантов одинаковы по темам и трудности. Следовательно, спецификация заданий дополнительного возмущения в результаты не вносит. Поэтому дисперсия среди испытуемых состоит из дисперсий реально существующего рассеяния баллов тестируемых и случайной ошибки измерений.

Тогда коэффициент надежности находится по формуле [7]:

$$r = \frac{D_{\text{исп}} - D_{\text{ост}}}{D_{\text{исп}}}.$$

Коэффициенты надежности и оценки соответствующих дисперсий, вычисленные для контрольно-измерительных материалов по русскому языку, приведены в таблице. Анализируя таблицу, можно сделать вывод о том, что коэффициент надежности заданий теста базового уровня сложности достаточно высок и постоянен, уровень дисперсии, обусловленной влиянием случайных факторов, со временем уменьшается.

Для данных ЕГЭ по русскому языку (блок заданий, оцениваемых по дихотомической шкале) имеем:

**Оценки дисперсий и коэффициенты надежности контрольно-измерительных материалов Единого государственного экзамена по русскому языку**

| | **2005** | **2006** | **2007** | **2008** | **2009** |
|---|---|---|---|---|---|
| $D_{\text{исп}}$ | 0.956 | 1,069 | 1.061 | 1,068 | 1,059 |
| $D_{\text{ост}}$ | 0.187 | 0,176 | 0.172 | 0,170 | 0,175 |
| коэффициент надежности ***r*** | **0.803** | **0,835** | **0.838** | **0,84** | **0,846** |
| ошибка измерений $e$ | 0.433 | 0,419 | 0.414 | 0,413 | 0,418 |

| | **2010** | **2011** | **2012** | **2013** | **2014** |
|---|---|---|---|---|---|
| $D_{\text{исп}}$ | 1.049 | 0,971 | 0,934 | 0.947 | 0,911 |
| $D_{\text{ост}}$ | 0.160 | 0,166 | 0,160 | 0.157 | 0,152 |
| коэффициент надежности ***r*** | **0.847** | **0,829** | **0,828** | **0.834** | **0.830** |
| ошибка измерений $e$ | 0.400 | 0,408 | 0,400 | 0.396 | 0,389 |

Вообще, количество открытых проблем в теории педагогического тестирования в настоящее время, по-видимому, гораздо больше, чем тех, которые уже получили в той или иной мере удовлетворительное решение. Тем не менее использование тестирования в реальной педагогической деятельности и сейчас позволяет заметно повысить детальность и точность оценивания результатов этой деятельности со всеми вытекающими отсюда последствиями и потому привлекает все большее количество сторонников. В основе недооценки тестирования лежит, как правило, только недостаточная информированность. Последний фактор имеет, к сожалению, реальную почву, поскольку почти вся содержательная литература по тестированию опубликована в основном на иностранных языках.

КОНЦЕПТ
научно-методический электронный журнал
Буров А. В. Применение однофакторного дисперсионного анализа к оценке надежности тестовых материалов Единого государственного экзамена // Концепт. – 2014. – № 12 (декабрь). – ART 14341. – 0,3 п. л. – URL: http://e-koncept.ru/2014/14341.htm. – Гос. рег. Эл № ФС 77-49965. – ISSN 2304-120X.
ART 14341 УДК 371.261:519.254

## Ссылки на источники

1. Нейман Ю. А., Хлебников В. А. Введение в теорию моделирования и параметризации педагогических тестов. – М., 2000. – 168 с.
2. Звонников В. И., Челышкова М. Б. Современные средства оценивания результатов обучения. – М.: Изд. центр «Академия», 2007. – 224 с.
3. Болотов В. А. Единый государственный экзамен: на пути к созданию системы независимой оценки качества образования // Единый государственный экзамен: сб. ст. – М., 2004.
4. Болотов В. А. Основные подходы к созданию общероссийской системы оценки качества образования // Единый государственный экзамен: сб. ст. – М., 2005.
5. Нейман Ю. А., Хлебников В. А. Указ. соч.
6. Анастази А., Урбина С. Психологическое тестирование. – СПб.: Питер, 2007. – 688 с.
7. Нейман Ю. А., Хлебников В. А. Указ. соч.

***Andrey Burov,***
*Master student, Smolensk State University, Smolensk*
burov_andrei@inbox.ru

ISSN 2304-120X 12
9 772304 120142

**Application of data statistical processing methods for validity estimation of tests for the Universal state exam**

**Abstract.** The paper is dedicated to the study of data statistical processing methods (one-way ANOVA test) for validity estimation of tests for the Universal state exam. The question is topical both for society and education. For validity estimation of testing items we apply some methods of statistical processing of data and propose the mechanism for data processing and control of testing items at the Universal state exam.

**Key words:** Universal state exam, estimation, reliability, test, statistical processing of data.

**References**

1. Nejman, Ju. A. & Hlebnikov, V. A. (2000) *Vvedenie v teoriju modelirovanija i parametrizacii pedagogicheskih testov*, Moscow, 168 p. (in Russian).
2. Zvonnikov, V. I. & Chelyshkova, M. B. (2007) *Sovremennye sredstva ocenivanija rezul'tatov obuchenija*, Izd. centr “Akademija”, Moscow, 224 p. (in Russian).
3. Bolotov, V. A. (2004) “Edinyj gosudarstvennyj jekzamen: na puti k sozdaniju sistemy nezavisimoj ocenki kachestva obrazovanija”, *Edinyj gosudarstvennyj jekzamen: sb. st*., Moscow (in Russian).
4. Bolotov, V. A. (2005) “Osnovnye podhody k sozdaniju obshherossijskoj sistemy ocenki kachestva obrazovanija”, *Edinyj gosudarstvennyj jekzamen: sb. st.,* Moscow (in Russian).
5. Nejman, Ju. A. & Hlebnikov, V. A. (2000) Op. cit.
6. Anastazi, A. & Urbina, S. (2007) *Psihologicheskoe testirovanie*, Piter, St. Peterburg, 688 p. (in Russian).
7. Nejman, Ju. A. & Hlebnikov, V. A. (2000) Op. cit.

***Рекомендовано к публикации:***
*Горевым П. М., кандидатом педагогических наук, главным редактором журнала «Концепт»*
*Утёмовым В. В., кандидатом педагогических наук*