Министерство образования Республики Беларусь
Учреждение образования «Витебский государственный
университет имени П.М. Машерова»
Кафедра информатики и информационных технологий

МЕТОДЫ МОНИТОРИНГА КАЧЕСТВА УЧЕБНОГО ПРОЦЕССА

Методические рекомендации

Витебск
ВГУ имени П.М. Машерова
2014

УДК 378.14:004(075.8)
ББК 74.480.2я73
Б90

Печатается по решению научно-методического совета учреждения образования «Витебский государственный университет имени П.М. Машерова». Протокол № 3 от 20.12.2013 г.

Авторы: старший преподаватель кафедры информатики и информационных технологий ВГУ имени П.М. Машерова **Н.В. Булгакова;** доцент кафедры информатики и информационных технологий ВГУ имени П.М. Машерова, кандидат биологических наук **А.А. Чиркина**

Р е ц е н з е н т ы :
декан факультета довузовской подготовки ВГУ имени П.М. Машерова, кандидат педагогических наук, доцент *Л.Л. Ализарчик*; доцент кафедры инженерной физики ВГУ имени П.М. Машерова, кандидат технических наук *В.И. Жидкевич*

Булгакова, Н.В.
Б90 Методы мониторинга качества учебного процесса : методические рекомендации / Н.В. Булгакова, А.А. Чиркина. – Витебск : ВГУ имени П.М. Машерова, 2014. – 35 с.

В методических рекомендациях к выполнению лабораторных работ рассмотрены методы организации мониторинга качества учебного процесса, которые студенты педагогических специальностей изучают в рамках дисциплины «Информационные основы научно-педагогической деятельности». Приведены практические задания по соответствующим темам и вопросы для самоконтроля.

УДК 378.14:004(075.8)
ББК 74.480.2я73

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Важнейшими условиями повышения качества образования являются систематический мониторинг образовательной системы и анализ объективных данных о качестве обучения и уровне подготовленности обучающихся.

Мониторинг – процесс отслеживания состояния объекта (системы или сложного явления) с помощью непрерывного или периодически повторяющегося сбора данных, представляющих собой совокупность определенных ключевых показателей. Мониторинг выполняет одну из следующих функций: установление соответствия осуществляемой деятельности нормативным правилам и нормам; выявление состояния критических точек, в отношении которых будут предприняты определенные действия; обеспечение обратной связи с окружением.

Педагогический мониторинг принято понимать как целенаправленное, специально организованное, непрерывное слежение за функционированием и развитием образовательного процесса и/или его отдельных элементов в целях своевременного принятия адекватных управленческих решений на основе анализа собранной информации и педагогического прогноза. В качестве объектов мониторинга в системе высшего профессионального образования могут выступать: учебный процесс, академическая успеваемость обучаемых, развитие личности обучаемых, профессиональная деятельность преподавателя и т.д.

В мониторинговом исследовании используются разные способы и каналы получения информации для оценивания и диагностики учебного процесса: опрос, наблюдение, анализ статистических данных, анкетирование, тестирование, экспертное оценивание, контент-анализ документации.

Для методов педагогического мониторинга характерен квалиметрический подход, базирующийся на теории педагогических измерений, массовости и независимости процедур тестирования, методах математической статистики и педагогического интерпретационного анализа. Квалиметрический мониторинг, как наиболее информативный и современный способ наблюдения за учебным процессом, чаще всего используется для оценки знаний, умений и навыков, выявления качества учебных достижений обучаемых и соотнесения их уровня с заданным эталоном (стандартом) или статистическими нормами. Принципиальное отличие квалиметрического мониторинга от традиционного заключается в использовании таких технологий контроля, которые ориентированы не на субъективные оценочные механизмы, а на современные объективизированные (дистанцированные от преподавателя) оценочные процедуры, удовлетворяющие требованиям объективности, сопоставимости и репрезентативности оценок. Особенностью квалиметрического мониторинга является использование тестирова-

ния и статистических методов анализа результатов как одного из направлений теории статистических выводов в педагогике.

В педагогическом мониторинге очень важное значение приобретает оценивание – процесс формирования оценки учебных достижений, в котором интегрируются и представляются в определенной шкале данные, полученные при тестировании, приведении экзаменов, выполнении обучаемыми практических работ, рейтинговании их результатов и т.д.

Тест – это инструмент, состоящий из квалиметрически выверенной системы тестовых заданий, стандартизированной процедуры проведения и заранее спроектированной технологии и анализа результатов для измерения качеств и свойств личности, учебных достижений, изменение которых возможно в процессе систематического обучения [7]. Тестирование как процесс применения тестов является частью процесса оценивания. В контексте педагогического мониторинга тестирование предполагает предъявление и использование научно обоснованных тестов, обладающих необходимыми статистическими характеристиками и обеспечивающих высокое качество измерений.

Педагогическим тестом называется система заданий специфической формы, определенного содержания, равномерно возрастающей трудности – система, создаваемая с целью объективно оценить структуру и измерить уровень подготовленности учащихся (студентов) [1].

Для того чтобы тестом можно было пользоваться для объективной оценки уровня знаний обучаемых, необходимо оценить его качество, то есть проверить на соответствие ряду требований. Определение качества тестов состоит из трех этапов. *Первый этап* – оценка качества содержания, правильности форм заданий, установление соответствия учебного материала целям измерения и выборке тестируемых. *Второй этап* – определение квалиметрических характеристик тестовых заданий по результатам статистического анализа данных апробационного тестирования. Для расчетов основных показателей использовалась классическая (традиционная) теория тестов. В результате этого этапа выявляются качество тестовых заданий, пригодность для данной группы обучающихся. Таким образом формируется база тестовых заданий, пригодных для конструирования тестов. *Третий этап* – оценка качества теста как измерительного инструмента на основе вычислений коэффициентов надежности тестовых результатов, определения валидности и оптимального времени тестирования. Определение валидности и надежности тестовых результатов характеризует эффективность теста как инструмента измерения учебных достижений обучаемых. Также на третьем этапе подлежат оцениванию методы определения баллов, используемых при интерпретации результатов тестирования.

Помимо тестовых методов и технологий в мониторинговых исследованиях широко используются *экспертные методы*. *Экспертиза* представляет собой исследование, разрешение при помощи компетентных в данной

области людей какого-либо вопроса, требующего специальных знаний. *Эксперт* – сведущее лицо, приглашаемое в спорных или трудных случаях для экспертизы. *Экспертный метод* – это комплекс логических и математических процедур, направленных на получение информации, ее анализ и обобщение с целью подготовки и принятия компетентного управленческого решения. Суть метода состоит в проведении экспертами анализа проблемы с качественной и количественной оценкой суждений, а также формальной обработкой результатов индивидуальных экспертных оценок. Данный метод обеспечивает оценку альтернативных решений и выбор предпочтительных вариантов решения.

В педагогике область применения экспертных методов весьма обширна. Это компетентная оценка образовательных программ, педагогических технологий и методик воспитания и обучения, новых учебников и учебных пособий. Таким образом, метод экспертных оценок позволяет анализировать сложные педагогические процессы, явления или ситуации, которые характеризуются качественными, неформализуемыми признаками, что затрудняет их количественный анализ и оценку.

В данных методических рекомендациях рассмотрены экспертный и тестовый методы организации мониторинга качества учебного процесса, раскрыты научные подходы к составлению и анализу качества контрольно-измерительных материалов, которые студенты педагогических специальностей изучают в рамках дисциплины «Информационные основы научно-педагогической деятельности». В каждой теме приводятся практические задания, в которых подробно представлены последовательность выполняемых студентами работ и материалы для самоконтроля.

Тема 1. Экспертный метод первичного отбора заданий

1.1. Теоретические сведения и пример выполнения задания

В настоящее время независимое тестирование становится важнейшей составляющей контрольно-оценочной системы. Чем качественнее тест, тем точнее количественная оценка уровня учебных достижений и ее приближение к латентной характеристике испытуемого – подготовленности.

Говоря об экспертизе качества контрольно-измерительных материалов, необходимо упомянуть, что существуют два направления осуществления экспертизы: содержательное и тестологическое.

Содержательная экспертная оценка очень важна для первичного отбора тестовых заданий при создании тестов. Она проводится специалистами соответствующей предметной области. На этапе подготовки теста группа экспертов выполняет комплексную оценку качества тестовых заданий. Содержательная экспертиза включает в себя анализ структуры контрольно-измерительных материалов на соответствие образовательному стандарту и анализ содержания отдельных заданий контрольно-измерительных материалов. Тестологическая экспертиза контрольно-измерительных материалов проводится экспертами тестовых материалов, прошедшими необходимую подготовку [9].

Цель экспертной оценки – оценить задание по заданным критериям и исключить из теста «слабые» задания. В процессе экспертизы эксперту необходимо:

- выполнить каждое задание теста (указывается правильный ответ, приводится решение задания там, где необходимо);
- проанализировать формулировки задания (проверяется предметная корректность формулировок);
- оценить содержание заданий на их тематическую принадлежность и уровень сложности (базовый, повышенный или высокий);
- в конструктивной форме сформулировать замечания для разработчиков к каждому из заданий с предложениями по внесению изменений;
- сформулировать заключение о пригодности теста для использования.

В процессе оценивания эксперту предлагается заполнить карту, в которую вносится информация о каждом задании по следующим характеристикам: номер правильного ответа; требования (одного или нескольких) к уровню подготовки испытуемых; уровень сложности; значимость содержания задания; ожидаемый процент выполнения испытуемыми; ожидаемое время выполнения задания; неудачные задания. Эксперты оценивают тестовые задания по следующим показателям качества: компактность формулировки задания и вариантов ответов; логичность (формулировка тестового задания в виде суждения); корректность (отсутствие лишних слов); достаточность (необходимое количество вариантов ответов); содер-

жательность (соответствие тестового задания учебной программе); значимость (уровень значимости содержания тестового задания); время вывода заключения (ответа); однозначность; ясность смысла тестовой ситуации.

Такого рода формально-содержательная экспертиза является сложной и требующей больших временных затрат деятельностью.

Другой вариант применения экспертных методов – экспертное ранжирование с целью отбора наиболее подходящих для практических целей тестовых заданий. Группе экспертов предлагается проранжировать по качеству и важности ряд тестовых заданий. Ранговая оценка сводится к обозначению степени важности каждого тестового задания. Наиболее важный показатель обозначают рангом R = 1, а наименее значимый – рангом R = n, где n – число оцениваемых заданий. Если эксперт считает несколько тестовых заданий равноценными по значимости, то им присваиваются равные ранги, но сумма их должна быть равна сумме мест при их последовательном расположении. Если эксперты оценивают важность единичных показателей с помощью баллов, то необходимо преобразование балльной оценки в ранговую. Очевидно, что сумма рангов для каждого тестового задания покажет его место в системе тестовых заданий. Задания, набравшие наибольшуюсумму рангов (то есть находящиеся на последнем месте по мнению экспертов)считаются наиболее «слабыми» подлежат исключению из базы тестовых заданий. Однако, прежде чем оценивать тестовые задания, следует оценить группу экспертов и согласованность их мнений.

Рассмотрим на примере метод экспертного ранжирования.

Пусть группа состоит из $m = 4$ экспертов, по результатам опроса которых нужно ранжировать $n = 5$ тестовых заданий: *A1, A2, A3, A4, A5*. Для обработки результатов опроса нужно ввести в таблицу данные, как показано на рис. 1:

	A	B	C	D	E	F
1		***Номер тестового задания***				
2	***Эксперт***	***A1***	***A2***	***A3***	***A4***	***A5***
3	***1***	6	2	3	3	2
4	***2***	5	2	3	4	1
5	***3***	7	3	3	3	2
6	***4***	4	3	2	4	3

Рис. 1. **Матрица с результатами работы 4 экспертов по пяти тестовым заданиям**

На первом этапе необходимо выполнить оценку качества самих экспертов. Оценка качества экспертов (K_{oc}) по отклонению от средней оценки экспертной группы производится по формулам:

$$\rho_j = \frac{\sum_{i=1}^{n} \left| \overline{R}i - R_{ij} \right|}{2 \sum_{i=1}^{n} \overline{R}i} \qquad (1)$$

$$K_{oc} = 10 \cdot (1 - \rho_j) \qquad (2)$$

где R_i – среднее значение ранга i-того оцениваемого тестового задания;

R_{ij} – значение i-того ранга, назначенное j-тым экспертом

Расчеты в таблице выполняются следующим образом:

- для вычисления среднего ранга нужно ввести в ячейку B7 формулу =СРЗНАЧ(B3:B6) и скопировать ее на диапазон ячеек C7:F7;
- для вычисления отклонения ранга в ячейку B9 вводится формула =ABS(B$7-B3) и копируется на соответствующий диапазон ячеек;
- в столбце G рассчитывается сумма отклонений для каждого эксперта, в ячейке G7 вычисляется итоговая сумма;
- в столбцах H и I выполняются вычисления по формулам (1) и (2).

Результирующая таблица приведена на рис. 3.

	A	B	C	D	E	F	G	H	I
1		***Номер тестового задания***					***Сумма***	ρ_j	K_{oc}
2	***Эксперт***	***A1***	***A2***	***A3***	***A4***	***A5***	***отклонений***		
3	***1***	6	2	3	3	2	***1,75***	0,05	**9,5**
4	***2***	5	2	3	4	1	***2,75***	0,08	**9,2**
5	***3***	7	3	3	3	2	***2,75***	0,08	**9,2**
6	***4***	4	3	2	4	3	***4,25***	0,13	**8,7**
7	Средний ранг	5,50	2,50	2,75	3,50	2,00	***16,25***		
8									
9	Отклонение 1 эксперта	0,50	0,50	0,25	0,50	0,00			
10	Отклонение 2 эксперта	0,50	0,50	0,25	0,50	1,00			
11	Отклонение 3 эксперта	1,50	0,50	0,25	0,50	0,00			
12	Отклонение 4 эксперта	1,50	0,50	0,75	0,50	1,00			

Рис. 2. **Результирующая таблица**

Решающее правило: в экспертную группу принимаются те эксперты, для которых $K_{oc} \geq 8{,}5$. По величине K_{oc} также можно сделать дополнительные выводы о группе специалистов, мнению которых можно доверять в наибольшей степени, и о том, какие эксперты наименее надежны. В данном случае все эксперты удовлетворяют указанному требованию.

Далее для оценки общей согласованности мнений экспертов вычисляется коэффициент конкордации ω Кендалла или по-другому – коэффициент множественной ранговой корреляции.

Для расчетов нужно перейти на другой лист и снова ввести начальные данные. Сначала определяют сумму рангов по заданиям. Для этого в

диапазон ячеек B7:F7 нужно ввести формулы, суммирующие значения рангов по столбцам. Например, в ячейку B7 вводится формула = СУММ(B3:B6), остальные ячейки диапазона заполняются аналогично.

Полученные значения позволяют построить среднюю априорную диаграмму рангов, но предварительно необходимо оценить степень согласованности мнений всех экспертов с помощью коэффициента конкордации ω. Для вычисления коэффициента конкордации нужно вычислить ряд промежуточных величин. Сначала находится среднее значение суммы рангов (ячейка D13 =СРЗНАЧ(B7:F7)). Далее нужно вычислить разность между суммой каждого тестового задания и средней суммой рангов (отклонение от среднего значения вычисляется в строке 8), затем найти квадрат отклонения для каждого задания (строка 9), и, наконец, сумму квадратов отклонений (ячейка D14 =СУММ(B9:F9)). На рис. 3 представлен вид расчетной таблицы.

	A	B	C	D	E	F
1		***Номер тестового задания***				
2	***Эксперт***	***A1***	***A2***	***A3***	***A4***	***A5***
3	***1***	6	2	3	3	2
4	***2***	5	2	3	4	1
5	***3***	7	3	3	3	2
6	***4***	4	3	2	4	3
7	Сумма рангов	=СУММ(B3:B6)	=СУММ(C3:C6)	=СУММ(D3:D6)	=СУММ(E3:E6)	=СУММ(F3:F6)
8	Отклонение	=B7-D13	=C7-D13	=D7-D13	=E7-D13	=F7-D13
9	Квадрат отклонения	=B8^2	=C8^2	=D8^2	=E8^2	=F8^2
10						
11	Количество экспертов m=			4		
12	Количество тестовых заданий k=			5		
13	Среднее значение суммы рангов			=СРЗНАЧ(B7:F7)		
14	Сумма квадратов отклонения			=СУММ(B9:F9)		

Рис. 3. **Расчетная таблица для вычисления коэффициента конкордации**

Далее нужно вычислить еще ряд величин: $T_i = \sum(t_i^3 - t_i)$; где t_i – число одинаковых рангов в i-м ранжировании. Покажем на примере: первый эксперт присвоил тестовым заданиям *A2* и *A5* одинаковые ранги 2, а заданиям *A3* и *A4* одинаковые ранги 3. Само значение ранга нас не интересует, только количество повторяющихся рангов. Ранг 2 повторяется два раза и ранг 3 повторяется два раза, следовательно, $T_1 = (2^3–2) +(2^3–2) = (8-2) + (8-2) = 12$. $T_2 = 0$, так как второй эксперт присвоил заданиям различные ранги; третий эксперт присвоил одинаковый ранг трем заданиям, значит $T_3 = 3^3-3 = 27-3 = 24$; T_4 вычисляется аналогично T_2.

Обозначив через S сумму квадратов отклонений, а через $\sum_{i=1}^{m} T_i$ сумму величин T_i по всем экспертам, можно вычислить коэффициент конкордации:

$$\omega = \frac{12 \cdot S}{m^2(k^3 - k) - m\sum_{i=1}^{m} T_i} \qquad (3)$$

В свободную ячейку таблицы (D15) вводится формула для вычисления коэффициента конкордации: =12*D14/(D11^2*(D12^3-D12)-D11*H6).

В зависимости от степени согласованности мнений экспертов коэффициент конкордации может принимать значения от нуля (при отсутствии согласованности) до единицы (при полном единодушии).

Эти данные нужно внести в таблицу, которая примет вид, как показано на рис. 4:

	A	B	C	D	E	F	G	H
1		***Номер тестового задания***						***T***
2	***Эксперт***	***A1***	***A2***	***A3***	***A4***	***A5***	***T1=***	12
3	***1***	6	2	3	3	2	***T2=***	0
4	***2***	5	2	3	4	1	***T3=***	24
5	***3***	7	3	3	3	2	***T4=***	12
6	***4***	4	3	2	4	3	***Сумма***	**48**
7	Сумма рангов	22	10	11	14	8		
8	Отклонение	9	-3	-2	1	-5		
9	Квадрат отклонения	81	9	4	1	25		
10								
11	Количество экспертов m=			4				
12	Количество тестовых заданий k=			5				
13	Среднее значение суммы рангов			13				
14	Сумма квадратов отклонения			120				
15	Коэффициент конкордации			0,83333		8		
16	Значение критерия хи-квадрат			13,3333		9,4877		

Рис. 4. **Расчетная таблица для вычисления коэффициента конкордации**

Если ω < 0,2 – 0,4, значит согласованность экспертов слабая, если ω > 0,6 – 0,8, то согласованность экспертов сильная. После каждого тура опроса экспертов осуществляется контроль и проверяется согласованность их мнений до тех пор, пока степень этой согласованности окажется удовлетворительной. Слабая согласованность обычно является следствием следующих причин:

- в рассматриваемой группе экспертов действительно отсутствует общность мнений;
- внутри группы существуют коалиции с высокой согласованностью мнений, однако, обобщенные мнения коалиций противоположны.

Полученное значение ω=0,833. Так как величина коэффициента конкордации существенно отличается от нуля, можно считать, что между мнениями экспертов имеется существенная связь.

Использовать коэффициент конкордации можно после оценки его значимости, которая возможна по вычисленному значению критерия хи-квадрат:

$$\chi^2 = \frac{12 \cdot S}{m \cdot k(k+1) - \frac{1}{k-1}\sum_{i=1}^{m} T_i} \qquad (4)$$

Для его вычисления в ячейку D16 вводится формула для вычисления χ^2: =12*D14/(D11*D12*(D12+1)-1/(D12-1)*H6). Полученное значение χ^2=13,333.

Гипотеза о наличии согласия может быть принята, если при заданном числе степеней свободы табличное значение χ^2 меньше расчетного для уровня значимости α=0,05. Табличное значение можно определить с помощью функции ХИ2ОБР(уровень значимости; степени_свободы). Количество степеней свободы определяется как k – 1.В данном случае для вычисления табличного значения используется формула: =ХИ2ОБР(0,05;4). Вычисленное значение равно 9,48. В связи с тем, что табличное значение критерия меньше расчетного, можно с 95% доверительной вероятностью утверждать, что мнение исследователей относительно степени значимости тестовых заданий является согласованным.

Также можно воспользоваться функцией ХИ2РАСП(проверяемое значение; степени_свободы). Эта функция возвращает вероятность того, что проверяемое значение меньше табличного, то есть вероятность нулевой гипотезы. Введем в свободную ячейку таблицы формулу =ХИ2РАСП(D16;4). Вычисленное значение будет равно 0,00976. Так как эта величина меньше уровня значимости α=0,05, нулевую гипотезу о несогласованности мнений экспертов можно отвергнуть.

Так как мнение экспертов согласовано, можно построить диаграмму рангов для рассматриваемых заданий, откладывая по одной оси номера заданий, а по другой – соответствующие суммы рангов. Чем меньше сумма рангов данного задания, тем выше его значимость. По результатам ранжирования часть заданий можно исключить из дальнейшего рассмотрения. Диаграмма рангов приведена на рис. 5.

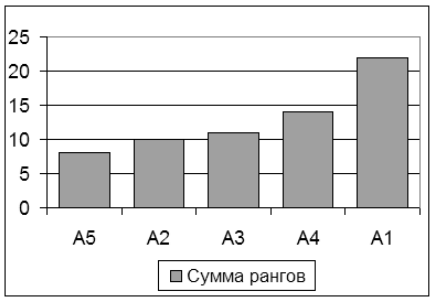

Рис. 5. **Диаграмма рангов**

Далее возникает вопрос: имеются ли какие-либо критерии, позволяющие определить, сколько же тестовых заданий нужно исключить?

Согласованность мнений экспертов относительно важности каждого

i-го тестового задания оценивается по коэффициенту вариации, который вычисляется по формуле:

$$C_i = \frac{S_i}{\overline{R_i}} \times 100 \qquad (5)$$

где C_i – коэффициент вариации мнений экспертов,
S_i– среднеквадратическое отклонение по каждому *i*-му тестовому заданию,

$$S_i = \sqrt{\frac{\sum_{j=1}^{m} (\overline{R_i} - R_{ij})^2}{m-1}} \qquad (6)$$

где $\overline{R_i}$ – средний по всем экспертам ранг *i*-го тестового задания;
R_{ij} – ранг *i*-го тестового задания, проставленный *j*-м экспертом;
m – число экспертов.

Для вычислений нужно добавить в таблицу 5 строк после девятой строки и ввести в них следующие формулы для расчета коэффициента вариации и коэффициентов весомости. На рис. 6 формулы показаны для столбца B, далее они копируются в столбцы C:F.

	A	B	C	D	E	F	G
1		***Номер тестового задания***					
2	***Эксперт***	***A1***	***A2***	***A3***	***A4***	***A5***	***T1=***
3	***1***	6	2	3	3	2	***T2=***
4	***2***	5	2	3	4	1	***T3=***
5	***3***	7	3	3	3	2	***T4=***
6	***4***	4	3	2	4	3	***Сумма***
7	Сумма рангов	=СУММ(B3:B6)	=СУМ	=СУМ	=СУМ	=СУМ	
8	Отклонение	=B7-D18	=C7-$D	=D7-$D	=E7-$D	=F7-$D	
9	Квадрат отклонения	=B8^2	=C8^2	=D8^2	=E8^2	=F8^2	
10	Средний ранг	=СРЗНАЧ(B3:B6)	=СРЗН	=СРЗН	=СРЗН	=СРЗН	
11	СКО	=СТАНДОТКЛОН(B3:B6)	=СТАН	=СТАН	=СТАН	=СТАН	
12	Коэффициент вариации	=B11/B10*100	=C11/C	=D11/D	=E11/E	=F11/F	**Сумма:**
13	1/(сумма рангов)	=1/B7	=1/C7	=1/D7	=1/E7	=1/F7	=СУММ(B13:F13)
14	Коэффициент весомости	=B13/G13	=C13/$	=D13/$	=E13/$	=F13/$	

Рис. 6. **Расчетная таблица для вычисления коэффициента вариации**

Чем больше значение коэффициента вариации C_i, тем меньше согласованность мнений экспертов в отношении значимости *i*-го задания. Интерпретация полученного коэффициента вариации: при C_i < 10% согласованность мнений экспертов считают высокой; при 10% < C_i< 15% – выше средней; при 15% < C_i < 25% – средней; при 25 < $C_i \leq$ 35% – ниже средней и при C_i > 35% – низкой.

Если величина ω говорит о хорошей согласованности мнений экспертов (ω >0,6), то всех *n* тестовых заданий выделяют наиболее значимые задания, для которых значение коэффициента весомости больше, чем величина *1/n*.

Итоговая расчетная таблица приведена на рис. 7.

	A	B	C	D	E	F	G	H
1		***Номер тестового задания***						***T***
2	***Эксперт***	***A1***	***A2***	***A3***	***A4***	***A5***	***T1=***	12
3	***1***	6	2	3	3	2	***T2=***	0
4	***2***	5	2	3	4	1	***T3=***	24
5	***3***	7	3	3	3	2	***T4=***	12
6	***4***	4	3	2	4	3	***Сумма***	**48**
7	Сумма рангов	22	10	11	14	8		
8	Отклонение	9	-3	-2	1	-5		
9	Квадрат отклонения	81	9	4	1	25		
10	Средний ранг	5,50	2,50	2,75	3,50	2,00		
11	СКО	1,29	0,58	0,50	0,58	0,82		
12	Коэффициент вариации	23,5	23,1	18,2	16,5	40,8	**Сумма:**	
13	1/(сумма рангов)	0,05	0,10	0,09	0,07	0,13	0,43	
14	Коэффициент весомости	0,11	0,23	0,21	0,17	0,29		

Рис. 7. **Расчетная таблица для вычисления коэффициента конкордации**

Коэффициент вариации показывает, что для заданий *A1–A4* согласованность мнений экспертов средняя, а для пятого задания – слабая. Коэффициент весомости показывает значимость задания. Расчетные коэффициенты сравнивают со значением *1/n*, где *n* – количество заданий, подвергающихся экспертизе. В примере сравнение коэффициента весомости со значением 1/5=0,2 показывает, что для дальнейшего использования будут отобраны тестовые задания, занимающие первые три места (с наименьшей суммой рангов) – *A5, A2, A3*.

1.2. Задания для практического выполнения

1. Создать базу из 15-20 тестовых заданий по одному из разделов школьного курса информатики или математики (указать класс, тему, раздел учебного материала).
2. Выполнить первичный отбор тестовых заданий на основе экспертного оценивания их качества, привлекая в качестве экспертов студентов своей группы. В процедуре оценивания должно участвовать не менее 5 экспертов. Для каждого эксперта требуется рассчитать показатель качества (Кос) по отклонению от средней оценки экспертной группы. Эксперты с низким показателем качества должны быть исключены из группы или заменены.
3. Оценить согласованность мнений экспертов с использованием коэффициента конкордации и оценки его значимости.
4. Определить согласованность мнений экспертов относительно важности каждого тестового задания, вычислить коэффициенты весомости. Построить диаграмму рангов тестовых заданий. На основе результа-

тов экспертизы исключить из базы тестовых заданий те задания, которые не прошли экспертный отбор.

5. Создать тест с использованием тестовых заданий, прошедших экспертный отбор. Провести тестирование студентов группы (подгруппы). Результаты представить в виде тестовой матрицы.

1.3. Контрольные вопросы по теме

1. Что представляет собой ранговая оценка?
2. Что показывает и как строится диаграмма суммы рангов?
3. Каким образом производится оценка качества экспертов? Какое существует условие для принятия эксперта в группу?
4. По какому критерию вычисляется согласованность мнений экспертов в группе? В каком диапазоне изменяется данный критерий?
5. Как определяется значимость коэффициента конкордации? Как интерпретируется его значение?
6. Что делать, если мнения экспертов оказались слабо согласованы? По каким причинам это может произойти?
7. Как определяется согласованность мнений экспертов относительно важности каждого тестового задания?
8. Для чего применяется и как рассчитывается коэффициент весомости заданий? В каком случае коэффициент весомости считается значимым?
9. Как определить, сколько тестовых заданий нужно исключить из рассмотрения (если число исключаемых заданий не определено заранее)?
10. Каким образом оценку объектов, данную экспертами в баллах, можно перевести в ранговую?

Тема 2. Определение квалиметрических характеристик тестовых заданий

2.1. Теоретические сведения

Тест обладает составом, целостностью и структурой. Взаимосвязь заданий, их принадлежность общему измеряемому фактору – это и есть целостность теста. Каждое задание теста выполняет отведенную ему роль, ни одно из заданий теста не может быть изъято без потери качества измерения. Способ связи заданий между собой образует структуру теста. В основном, это так называемая факторная структура, в которой каждое задание связано с другими через общую вариацию тестовых результатов. [2]

Оценивание качества тестовых заданий проводится путем вычисления:

- статистических характеристик индивидуальных баллов испытуемых;
- значений трудности заданий;

- значений коэффициентов корреляции тестового задания с итоговым баллом по тесту;
- параметров дискриминативности тестовых заданий;
- коэффициента корреляции задания с заданием (проверка теста на гомогенность).

Далее проводится анализ соответствия эмпирических показателей с рекомендуемыми критериями.

В педагогических измерениях к основополагающим понятиям относятся понятие **истинного балла** (true score) – параметра испытуемого и термины «**сырой балл**» и «**наблюдаемый балл**», которые получаются простым суммированием оценок по отдельным заданиям теста. Часто истинный балл называют константой испытуемого в момент измерения, не зависящей от средства измерения.

При одномерных измерениях каждому испытуемому можно поставить в соответствие только один истинный балл. Наблюдаемых баллов может быть столько, сколько было используемых для измерения этой переменной тестов.

Так как любые результаты тестирования всегда содержат в себе ошибочные компоненты измерения, главной целью при создании или применении педагогического теста является получение наиболее точной оценки параметра подготовленности испытуемых. Поэтому при создании тесты проходят процесс научного обоснования качества, который нацелен на улучшение характеристик заданий для повышения точности тестовых баллов. Этот процесс основывается на математико-статистическом аппарате классической или современной теории тестов (Item Response Theory). Современная теория достаточно сложна и требует больших усилий при обработке и интерпретации данных для коррекции тестов. Классическую теорию используют значительно чаще, особенно при небольших выборках в 50–100 человек на каждый вариант теста[1].

Математико-статистическая обработка обычно проводится с помощью специального программного обеспечения, но для того, чтобы понять смысл некоторых показателей качества теста, выполним эти вычисления в среде Ecxel.

Условимся каждый правильный ответ испытуемого на задание считать равным 1 баллу, а каждый неправильный ответ или пропуск задания приравнивать к 0 баллов. Тогда профиль ответов испытуемого будет иметь вид последовательности из единиц и нулей. Наиболее адекватной формой представления наблюдаемых результатов выполнения теста служит матрица, в которой сведены воедино профили ответов респондентов и профили заданий теста (столбцы из оценок всех отвечавших на вопросы теста по каждому заданию теста).

[1] Если такое количество данных сразу собрать не удается, то их нужно накапливать на протяжении нескольких лет. Особенно нежелательны небольшие выборки при разработке итогового теста.

Пример матрицы наблюдаемых результатов, полученной при выполнения N (N = 30) студентами n (n = 10) заданий теста при дихотомических оценках (1 или 0) по заданиям приведен в табл. 1.

Справа в вертикальном столбце содержатся индивидуальные баллы студентов X_i (i = 1, 2, …, N), которые получаются суммированием единиц по горизонтали в каждом профиле ответов. Сложение единиц в столбцах по профилям ответов на n заданий теста позволяет получить числа Y_j (j = 1, 2, …, n), соответствующие количеству правильных ответов на каждое задание. С помощью матрицы можно выполнить ряд расчетов, интерпретация результатов которых позволяет сделать важные выводы относительно качества заданий теста и получить достаточно точные оценки параметра испытуемых в том случае, если тест соответствует определенным критериям качества.

Для анализа будем использовать упорядоченную матрицу, в которой задания ранжированы по нарастанию трудности и баллы испытуемых расположены по убыванию или возрастанию сверху вниз (фрагмент упорядоченной матрицы приведен на рис. 8).

	A	B	C	D	E	F	G	H	I	J	K	L
1	Номера испытуемых i	Номер задания теста										Индивидуальные баллы (множество X_i)
2		**3**	**7**	**5**	**2**	**4**	**6**	**8**	**1**	**10**	**9**	
3	19	0	0	0	0	0	0	0	0	0	0	0
4	16	0	0	0	0	0	0	0	1	0	0	1
5	24	0	0	0	0	0	0	0	0	0	1	1
6	7	0	0	1	0	0	1	0	0	0	0	2
7	22	0	0	0	0	0	0	0	1	0	1	2
8	29	0	0	0	1	0	0	0	0	0	1	2
9	6	0	0	0	1	1	0	0	0	0	1	3
10	11	0	1	0	0	0	1	0	0	0	1	3
11	15	0	0	0	0	0	0	1	0	1	1	3
12	23	0	0	0	0	0	0	0	1	1	1	3
13	26	0	0	0	0	0	0	1	1	1	0	3
14	3	1	0	1	0	0	1	0	1	0	0	4
15	10	0	1	0	0	0	1	0	0	1	1	4
16	13	0	0	1	0	1	0	1	1	0	0	4
17	25	0	0	0	1	0	0	1	0	1	1	4
18	30	0	0	1	0	1	0	0	1	0	1	4
19	1	1	0	0	1	0	0	1	0	1	1	5

Рис. 8. **Упорядоченная матрица исходных данных (фрагмент)**

Таблица 1

Пример матрицы наблюдаемых результатов выполнения теста

Номера испытуемых i	Номер задания теста										Индивидуальные баллы (множество X_i)
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
1	0	1	1	0	0	0	0	1	1	1	5
2	1	1	1	1	1	1	1	0	0	1	8
3	1	0	1	0	1	1	0	0	0	0	4
4	0	1	0	1	0	1	0	1	1	1	6
5	1	1	0	1	1	0	0	1	1	1	7
6	0	1	0	1	0	0	0	0	1	0	3
7	0	0	0	0	1	1	0	0	0	0	2
8	1	1	1	1	1	1	1	1	1	1	10
9	0	1	0	1	1	1	1	1	0	0	6
10	0	0	0	0	0	1	1	0	1	1	4
11	0	0	0	0	0	1	1	0	1	0	3
12	1	1	0	1	0	0	1	0	1	1	6
13	1	0	0	1	1	0	0	1	0	0	4
14	1	0	0	1	1	0	1	1	1	1	7
15	0	0	0	0	0	0	0	1	1	1	3
16	1	0	0	0	0	0	0	0	0	0	1
17	1	0	0	0	0	1	1	0	1	1	5
18	1	0	0	1	1	1	0	1	0	0	5
19	0	0	0	0	0	0	0	0	0	0	0
20	1	0	1	0	0	1	0	0	1	1	5
21	1	1	0	1	1	0	0	0	1	1	6
22	1	0	0	0	0	0	0	0	1	0	2
23	1	0	0	0	0	0	0	0	1	1	3
24	0	0	0	0	0	0	0	0	1	0	1
25	0	1	0	0	0	0	0	1	1	1	4
26	1	0	0	0	0	0	0	1	0	1	3
27	0	1	0	0	0	1	0	1	1	1	5
28	0	1	1	1	0	0	0	0	1	1	5
29	0	1	0	0	0	0	0	0	1	0	2
30	1	0	0	1	1	0	0	0	1	0	4
Число правильных ответов (множество Yi)	16	13	6	13	11	12	8	12	21	17	129

Для больших выборок испытуемых (50 студентов и более) до того, как выполнить графическую интерпретацию, формируют частотное распределение баллов (табл. 2).

Частота тестового балла – это количество испытуемых, имеющих данный тестовый балл.

Таблица 2

Частотное распределение баллов

Балл Xi	0	1	2	3	4	5	6	7	8	10
Частота (f)	1	2	3	5	5	6	4	2	1	1

В табл. 2 содержатся только различные индивидуальные баллы испытуемых, взятые из последнего столбца матрицы эмпирических результатов выполнения теста (см. рис. 1) и расположенные в порядке возрастания вместе с числом их повторений (f). Сумма всех частот равна числу студентов в группе.

По ряду частотного распределения можно получить графическое представление результатов тестирования в виде гистограммы. Для рассматриваемого примера (см. табл. 2) гистограмма приведена на рис. 9. Середина столбца совмещается с серединой интервала разряда, который выбран длиной в 1 балл.

Рис. 9. **Столбчатая гистограмма распределения баллов табл. 3**

Для дальнейшего анализа данных оцениваются меры центральной тенденции в распределении результатов тестирования, которые предназначены для выявления той точки, вокруг которой в основном группируются все результаты выполнения теста.

Мода – это такое значение, которое встречается наиболее часто среди результатов выполнения теста. Если все значения баллов встречаются одинаково часто, принято считать, что моды у распределения нет.

Среднее выборочное (**среднее арифметическое**) определяется суммированием всех значений совокупности и последующим делением на их количество. Для индивидуальных баллов X_1, X_2, …, X_N группы N испытуемых среднее значение $\bar{X}$ будет:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N} \qquad (7)$$

В отличие от моды, фиксирующей одно или несколько значений, на величину среднего влияют значения всех результатов распределения. Таким образом, среднее арифметическое характеризует все распределение в целом. Оно обобщает индивидуальные особенности составляющих распределения на основе уравнивания отдельных значений рассматриваемой величины. В нашем примере значение среднего оказалось равным 4,3.

Меры центральной тенденции полезны при оценке качества теста, если есть результаты апробации теста на репрезентативной выборке студентов. Обычно считают, что хороший нормативно-ориентированный тест обеспечивает нормальное распределение индивидуальных баллов репрезентативной выборки испытуемых, если среднее значение баллов находится в центре распределения, а остальные значения концентрируются вокруг среднего по нормальному закону, т.е. примерно 70% значений находятся в центре, а остальные сходят на нет к краям распределения, как на рис. 10.

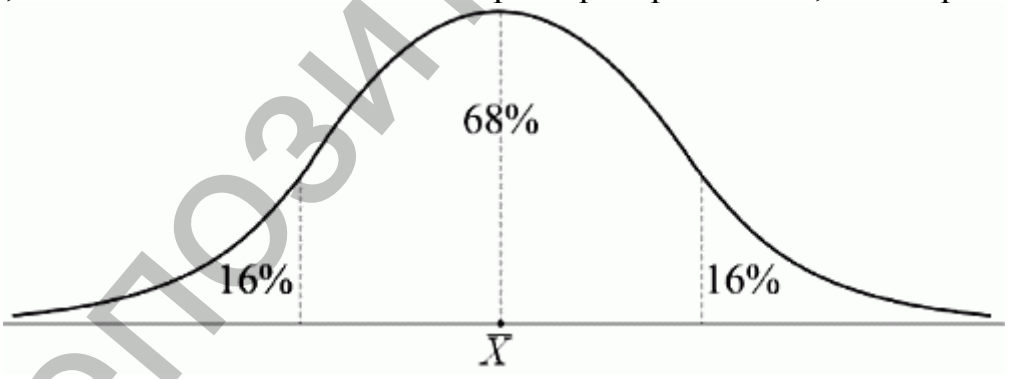

Рис. 10. **Нормальная кривая распределения индивидуальных баллов**

Правильно сконструированный нормативно-ориентированный тест на репрезентативной выборке студентов должен обеспечивать близкое к симметричному распределение индивидуальных баллов, когда мода и среднее значение примерно равны, а остальные результаты расположены вокруг среднего по нормальному закону.

Дальнейший анализ данных тестирования связан с оцениванием мер изменчивости в распределении индивидуальных баллов. Характеристика

изменчивости указывает на особенности разброса эмпирических данных вокруг среднего значения баллов. Отдельные значения индивидуальных баллов могут быть тесно сгруппированы вокруг своего среднего балла либо, наоборот, сильно удалены от него. Для отражения характера рассеяния отдельных значений вокруг среднего используют различные меры: размах, дисперсию и стандартное отклонение.

Размах измеряет на шкале расстояние, в пределах которого изменяются все значения показателя в распределении. Например, для распределения индивидуальных баллов в табл. 3 размах равен 9 – 1 = 8. Вариационный размах легко вычисляется, но используется крайне редко при характеристике распределения баллов по тесту.

Подсчет **дисперсии** основан на вычислении отклонений $X_i - \bar{X}$ (i = 1, 2, …, N) каждого значения показателя от среднего арифметического в распределении. Для индивидуальных баллов значения отклонений несут информацию о вариации совокупности значений баллов N студентов, поскольку отражают меру неоднородности результатов по тесту. Совокупность с большей неоднородностью будет иметь большие по модулю отклонения, наоборот, для однородных распределений отклонения должны быть близки к нулю. Знак отклонения указывает место результата студента по отношению к среднему арифметическому по тесту. Для студента с индивидуальным баллом выше среднего значение разности $X_i - \bar{X}$ будет положительно, а для тех, у кого результат ниже $\bar{X}$, отклонение $X_i - \bar{X}$ меньше нуля.

Чтобы отрицательные и положительные слагаемые не уничтожали друг друга, каждое отклонение возводят в квадрат и находят сумму квадратов отклонений. Эта сумма будет большой, если результаты тестирования отличаются существенной неоднородностью, и малой в случае близких результатов испытуемых по тесту.

$$\sum_{i=1}^{N}(X_i - \bar{X})^2 = (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_N - \bar{X})^2 = {d_1}^2 + {d_2}^2 + \cdots + {d_N}^2 \quad (8)$$

В нашем случае получили значение этой суммы, равное 140,3.

Величина суммы зависит также от размера выборки испытуемых, выполнявших тест, поэтому для сопоставимости мер изменчивости распределений, отличающихся по объему, каждую сумму делят на N – 1, где N – число студентов, выполнявших тест. Определяемая таким образом мера изменчивости называется **исправленной дисперсией**. Она обычно обозначается символом ${S_x}^2$ и вычисляется по формуле

$${S_x}^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1} \qquad (9)$$

Значение исправленной дисперсии для наших данных ≈4,838.

Кроме дисперсии, для характеристики меры изменчивости распределения удобно использовать еще один показатель вариации, который назы-

вается **стандартным отклонением** и вычисляется путем извлечения квадратного корня из дисперсии:

$$S_x = \sqrt{S_x{}^2} \quad (10)$$

Значение стандартного отклонения в рассматриваемом примере равно 2,19953.

В оценке качества тестов дисперсия играет важную роль. Низкий показатель дисперсии указывает на плохое качество нормативно-ориентированного теста, поскольку не обеспечивает высокий дифференцирующий эффект. Излишне высокое значение дисперсии, характерное для случая, когда все испытуемые отличаются по числу выполненных заданий, также требует переработки теста из-за существенного отличия вида распределения баллов от планируемой нормальной кривой.

Использование стандартного отклонения как меры вариации особенно эффективно для нормального распределения баллов испытуемых, поскольку в этом случае можно прогнозировать процент данных, лежащих внутри одного, двух и трех стандартных отклонений, откладываемых от центра распределения. В любом нормальном распределении приблизительно 68% площади под кривой лежит в пределах одного стандартного отклонения, откладываемого влево и вправо от среднего (т.е. $\bar{X} \pm 1 \cdot S_x$); 95% площади под кривой расположено в пределах двух S_x откладываемых слева и справа от среднего ($\bar{X} \pm 2 \cdot S_x$); 99,7% площади под кривой – в пределах трех S_x по обе стороны от $\bar{X}$ ($\bar{X} \pm 3 \cdot S_x$).

Кривая распределения индивидуальных баллов, получаемых на репрезентативной выборке, носит неслучайный характер. Она является следствием подбора трудности заданий теста. При смещении в сторону легких заданий большая часть респондентов выполнит почти все задания теста и получит высокие индивидуальные баллы. В случае приоритетного подбора трудных заданий в распределении индивидуальных баллов получится всплеск вблизи начала горизонтальной оси. При оптимальной трудности теста, когда распределение оценок параметра трудности заданий имеет вид нормальной кривой, автоматически возникает нормальность распределения индивидуальных баллов репрезентативной выборки испытуемых. Этот факт позволяет считать полученное распределение устойчивым по отношению к генеральной совокупности и определить репрезентативные нормы выполнения теста.

Углубленный анализ качества теста, позволяющий сделать выводы о направлениях коррекции содержания отдельных заданий, связан с вычислением показателей связи между результатами испытуемых по отдельным заданиям теста. При оценке качества заданий важно понять, существует ли тенденция, когда одни и те же студенты добиваются успеха в какой-либо

паре заданий теста либо состав испытуемых, добивающихся успеха, полностью меняется при переходе от одного задания теста к другому. Ответ на вопрос о существовании связи между двумя наборами данных получают с помощью **корреляции**.

Для вычисления коэффициента корреляции используется величина, которая называется **коэффициентом корреляции Пирсона** r_{xy}:

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left(\sum_{i=1}^{N}(X_i - \bar{X})^2\right)\left(\sum_{i=1}^{N}(Y_i - \bar{Y})^2\right)}} \qquad (11)$$

Преобразованный коэффициент Пирсона для дихотомических данных вычисляется по формуле

$$\varphi_{jl} = \frac{p_{jl} - p_j p_l}{\sqrt{p_j q_j \cdot p_l q_l}}, \qquad (12)$$

где p_{jl} – доля испытуемых, выполнивших правильно оба задания с номерами j и l, т.е. доля тех, кто получил 1 балл по обоим заданиям; p_j – доля испытуемых, правильно выполнивших j-е задание, $q_j = 1 - p_j$; p_l – доля испытуемых, правильно выполнивших l-е задание теста, $q_l = 1 - p_l$.

Результаты подсчета значений коэффициента корреляции между всеми заданиями для примера матрицы сведены в табл. 3. В Excel для её построения использовалась опция **Корреляция** пакета **Анализ данных**. Соответствующие настройки окна **Корреляция** приведены на рис. 11.

Таблица 3

Коэффициенты корреляции заданий

	1	2	3	4	5	6	7	8	9	10
1	1									
2	-0,26	1								
3	0,134	0,235	1							
4	0,279	0,457	0,067	1						
5	0,434	0,033	0,138	0,591	1					
6	-0,05	-0,03	0,272	-0,03	0,226	1				
7	0,111	0,081	0,075	0,233	0,167	0,431	1			
8	-0,05	0,247	-0,07	0,247	0,226	0,028	-0,03	1		
9	-0,17	0,279	-0,04	-0,01	-0,41	-0,21	0,066	-0,06	1	
10	0,126	0,357	0,269	0,086	-0,17	0,027	0,223	0,302	0,455	1

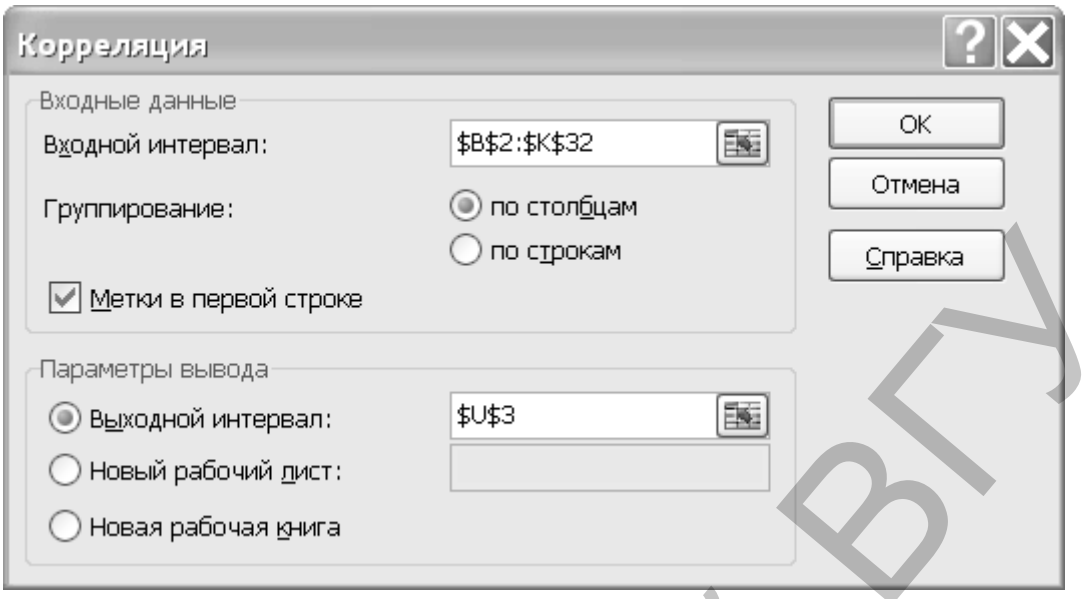

Рис. 11. **Настройки диалогового окна Корреляция для вычисления коэффициентов корреляции заданий теста**

Отрицательные значения коэффициента корреляции указывают на определенный просчет разработчиков в содержании заданий, которые рекомендуется из теста удалить. Наиболее распространенная причина появления отрицательной корреляции – отсутствие предметной чистоты содержания – нередко встречается при разработке самых разных тестов.

Анализ полученных в табл. 3 значений указывает на наличие ряда довольно высоких значений коэффициента корреляции ($\varphi_{5,4}$ = 0,591; $\varphi_{4,2}$ = 0,457; $\varphi_{10,9}$ = 0,455), которые могут получить различную трактовку в зависимости от вида разрабатываемого теста. Для тематических тестов высокая корреляция между заданиями неизбежна, так как задания тематического теста имеют слабо варьирующее исходное содержание, что вполне объяснимо назначением теста. Однако для итоговых тестов высокой корреляции между заданиями по возможности стараются избегать, поскольку вряд ли имеет смысл включать в итоговый тест несколько заданий, оценивающих одинаковые содержательные элементы. Поэтому в итоговых аттестационных тестах обычно стремятся к невысокой положительной корреляции, когда значения коэффициента варьируют в интервале (0; 0,3), и каждое задание привносит свой специфический вклад в общее содержание теста.

С помощью подсчета значений **точечного бисериального коэффициента корреляции** можно оценить **валидность** отдельных заданий теста. Бисериальный коэффициент корреляции используется в том случае, когда один набор значений распределения задается в дихотомической шкале, а другой – в интервальной. Под эту ситуацию подпадает подсчет корреляции между результатами выполнения каждого задания (дихотомическая шкала) и суммой баллов испытуемых (интервальная или квазиинтервальная шкала) по заданиям теста.

Формула для вычисления значения точечного бисериального коэффициента r_{pbis}, имеет вид:

$$r_{pbis} = \frac{(\bar{X}_1)_j - (\bar{X}_0)_j}{S_x} \sqrt{\frac{(N_1)_j \cdot (N_0)_j}{N(N-1)}}, \quad (13)$$

где $(\bar{X}_1)_j$ – среднее значение индивидуальных баллов испытуемых, выполнивших верно j-е задание теста; $(\bar{X}_0)_j$ – среднее значение индивидуальных баллов испытуемых, выполнивших неверно j-е задание теста; S_x – стандартное отклонение по множеству значений индивидуальных баллов; $(N_1)_j$ – число испытуемых, выполнивших верно j-е задание теста; $(N_0)_j$ – число испытуемых, выполнивших неверно j-е задание теста; N – общее число испытуемых, $N = N_1 + N_0$.

Значения бисериального коэффициента корреляции десяти заданий с суммой баллов по тесту, рассчитанные с помощью Excel для данных матрицы и отсортированные по убыванию, приводятся в табл. 4.

Таблица 4

Значения коэффициента бисериальной корреляции

Номер задания	4	10	2	7	5	3	8	6	1	9
r_{pbis}	0,6563	0,5879	0,5319	0,5089	0,5022	0,4316	0,4216	0,3587	0,3460	0,1917

В целом задание можно считать **валидным**, когда значение $(r_{bis})_j = 0{,}5$ или выше этого числа. В тесте из рассматриваемого примера валидными оказались задания с номерами 4, 10, 2, 7 и 5. Остальные задания этого теста имеют бисериальный коэффициент со значением, меньшим 0,5. Оценка валидности задания позволяет судить о том, насколько оно пригодно для работы в соответствии с общей целью создания теста. Если эта цель – дифференциация студентов по уровню подготовки, то валидные задания должны четко отделять хорошо подготовленных от слабо подготовленных испытуемых тестируемой группы. В тесте из примера самую хорошую дифференциацию по уровню подготовки дает задание № 4, так как у него самый высокий показатель бисериального коэффициента – 0,6563.

В оценке валидности задания решающую роль играет разность $(\bar{X}_1)_j - (\bar{X}_0)_j$, которая находится в числителе дроби формулы для вычисления бисериального коэффициента корреляции. Чем выше значение этой разности, тем выше вклад данного задания в общую дифференциацию испытуемых. Значения, близкие к нулю, указывают на низкую дифференцирующую способность заданий теста. В том случае, когда в разности доминирует значение $(\bar{X}_0)$, а не $(\bar{X}_1)$, задание следует удалить из теста. В нем побеждают слабые испытуемые, а сильные выбирают неверный ответ либо пропускают задание при выполнении теста. Таким образом, подлежат удалению все задания, у которых $r_{bis} < 0$.

Оценка **трудности** тестовых заданий получается по формуле

$$p_j = \frac{R_j}{N}, \quad (14)$$

где p_j – доля правильных ответов на j-е задание; R_j – количество студентов, выполнивших j-е задание верно; N – число студентов в тестируемой группе; j – номер задания теста, j = 1, 2, …, n. Трудность задания нередко выражают в процентах, тогда оценку, полученную по этой формуле, умножают на 100%.

Долю правильных ответов на задание pj естественно интерпретировать как легкость задания, в то время как трудность скорее ассоциируется с долей неправильных ответов q_j, которая находится путем вычитания p_j из единицы: $q_j = 1 - p_j$. Однако по сложившейся традиции в классической теории тестов за трудность задания принимается именно доля p_j.

Подбор заданий по трудности в тесте удобно оценить с помощью гистограммы, приведенной на рис. 12.

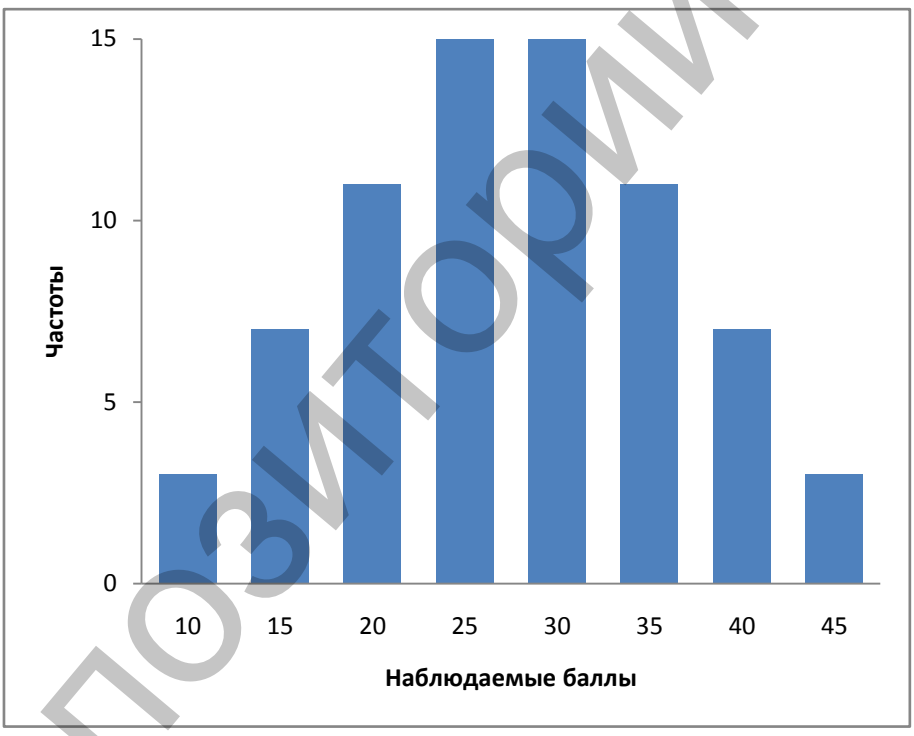

Рис. 12. **Гистограмма нормативно-ориентированного теста, хорошо сбалансированного по трудности**

В хорошо сбалансированном по трудности нормативно-ориентированном тесте есть несколько самых легких заданий со значениями $p > 0$. Есть несколько самых трудных с $p > 1$. Остальные задания по значениям p занимают промежуточное положение между этими крайними ситуациями и имеют в основном трудность 60–70%.

Сложность (трудность) всего теста рассчитывается аналогично как доля правильных ответов на задания всего теста. Допустимый диапазон значения p_j составляет 0,4–0,7.

2.2. Задания для практического выполнения

1. Опишите свой тест: нормативно-ориентированным или критериально-ориентированным он является.
2. Результаты прохождения теста, разработанного вами на предыдущем занятии, представьте в виде матрицы наблюдаемых баллов.
3. Рассчитайте индивидуальные баллы и количество правильных ответов на каждое тестовое задание.
4. Используя пакет Анализ данных, приведите описательную статистику по результатам Вашего теста. Интерпретируйте результаты.
5. Постройте гистограмму распределения частот набранных баллов по тесту.
6. Определите, является ли Ваш тест унимодальным.
7. Определите меру неоднородности результатов тестирования (дисперсию).
8. Проанализируйте наличие связи между заданиями Вашего теста.
9. Оцените валидность заданий теста, используя бисериальный коэффициент корреляции.
10. Произведите анализ трудности тестовых заданий и всего теста математическими методами.
11. Сформируйте текстовый отчет о результатах анализа качества вашего теста.

2.3. Контрольные вопросы по теме

1. Что понимается под нормативно-ориентированным тестом? Какие тесты называются критериально-ориентированными?
2. Что понимается под матрицей наблюдаемых результатов?
3. Что такое профиль испытуемого? В каком случае имеет место инвертированный профиль?
4. Что отражает частотное распределение баллов?
5. Что такое «карман» при построении гистограммы в Ecxel? Как повлияет изменение значений кармана (например, их сокращение вдвое) на вид гисторгаммы? Как можно интерпретировать результаты в этом случае?
6. Дайте определения понятиям мода, медиана, среднее арифметическое, размах, дисперсия. Какие соглашения приняты в статистике по использованию моды?
7. Поясните, что значит «унимодальный тест».
8. Считается, что хороший нормативно-ориентированный тест обеспечивает нормальное распределение индивидуальных баллов репрезентативной выборки испытуемых. Как это понимать?
9. М.Б. Челышкова [10] считает, что если среднее арифметическое примерно равно утроенному стандартному отклонению, т.е. $\bar{X} \approx 3S_x$, то

можно считать дисперсию оптимальной, а распределение тестовых баллов близким к нормальному. Всегда ли справедливо это утверждение? Ответ поясните.

10. Какие выводы можно сделать по таблице значений коэффициента корреляции между всеми заданиями теста? Одинаковы ли эти выводы для нормативно-ориентированных и критериально-ориентированных тестов?
11. Для чего используется точечный бисериальный коэффициент? Какую оценку тестовых заданий можно сделать, имея таблицу значений бисериальных коэффициентов?
12. Как оценивается трудность тестового задания?
13. Как оценивается трудность всего теста?

Тема 3. Оценка качества теста как измерительного инструмента

3.1. Теоретические сведения и пример выполнения задания

После формирования банка тестовых заданий из них составляется тест нужной длины, который должен удовлетворять определенным критериям надежности и валидности. Понятия надежности и валидности педагогического теста чрезвычайно важны, поскольку именно они характеризуют тест как измерительный инструмент, устойчивость его к действию помех. Надежность теста характеризуется воспроизводимостью его результатов.

Надежность как устойчивость результатов теста определяют с помощью повторного тестирования посредством того же теста, который проводится через некоторое время (например, через 2 недели) или эквивалентной его формы (параллельного теста). Вычисляется корреляционный коэффициент, показывающий степень совпадения результатов тестирования при повторной проверке знаний. Согласованность результатов можно измерять коэффициентом корреляции Пирсона $\boldsymbol{r}$. Если значения коэффициента $\boldsymbol{r}$ попадают в интервал 0,80-0,89, то говорят, что тест обладает хорошей надежностью, а если этот коэффициент не меньше 0,90, то надежность можно назвать очень высокой. Недостатком метода повторных измерений является необходимость дважды проходить один и тот же тест теми же испытуемыми.

Надежность как внутренняя согласованность определяется связью каждого тестового задания с общим результатом, то есть эта характеристика теста указывает на степень однородности состава заданий с точки зрения измеряемого качества.

При применении *метода расщепления* тестовую матрицу разбивают на две половины, состоящие из заданий с четными и нечетными номерами. Коэффициент корреляции $\boldsymbol{r_{1/2}}$ Пирсона между двумя совокупностями сум-

марных баллов результатов может служить оценкой надежности всего теста. Оценку надежности полного теста можно делать с использованием коэффициента корреляции $r_{1/2}$, по формуле Спирмена-Брауна, чтобы получить коэффициент, который относится ко всему тесту целиком, а не к отдельным его частям:

$$\rho_1 = \frac{2r_{1/2}}{1+r_{1/2}} \qquad (15)$$

Другой способ оценки надежности расщепленного теста основан на формуле Рюлона:

$$\rho_2 = 1 - \frac{S_d^2}{S_x^2} \qquad (16)$$

где S_x^2 – дисперсия суммарных баллов результата (индивидуальные баллы в тестовой таблице), а S_d^2 – дисперсия разностей между результатами каждого испытуемого по обеим половинам теста.

Еще один метод определения надежности, основанный на однократном предъявлении теста, носит имя Кьюдера-Ричардсона. Он использует данные о выполнении испытуемыми каждого задания. Коэффициент надежности Кьюдера-Ричардсона вычисляется по следующей формуле:

$$\rho_3 = \frac{N}{N-1}\left(1 - \frac{\sum_{j=1}^{n} p_i q_i}{S_x^2}\right) \qquad (17)$$

где p_j – доля правильных ответов на j-ое задание, т.е. кол-во правильных ответов, деленное на кол-во студентов;

q_j – доля неправильных ответов на j-ое задание, т. е. кол-во неправильных ответов, деленное на кол-во студентов (q_j=1 – p_j);

S_x^2 – дисперсия суммарных баллов результата (индивидуальные баллы в тестовой таблице).

Рассмотрим на примере оценку надежности теста. Для этого в исходную таблицу необходимо добавить: столбцы для расчета балла за задания с четными и нечетными номерами (столбцы M и N); столбец разностей между результатами каждого испытуемого по обеим половинам теста (столбец O); строки для расчета доли правильных и неправильных ответов, и их произведения (строки 34-36).

Ячейки с формулами заполняются в соответствии с табл. 5.

Таблица 5

Формулы для расчетов

Показатель	Ячейка	Формула
Сумма произведения доли правильных и неправильных ответов по всем испытуемым	L36	=СУММ(B36:K36)
Коэффициент корреляции Пирсона r1/2	L38	=КОРРЕЛ(M3:M32;N3:N32)
Коэффициент надежности по формуле Спирмена-Брауна	L39	=2*L38/(1+L38)
Дисперсия суммарных баллов	L40	=ДИСП(L3:L32)
Дисперсия разностей между результатами половин теста	L41	=ДИСП(O3:O32)
Коэффициент надежности по формуле Рюлона	L42	=1-L41/L40
Коэффициент надежности по формуле Кьюдера-Ричардсона	L43	=30/(30-1)*(L40-L36)/L40

Расчетная таблица приведена на рис. 13.
Результаты расчетов отражены в табл. 6.

Таблица 6

Результаты расчетов основных показателей для оценки надежности теста

Показатель	Значение
Коэффициент корреляции Пирсона $r_{1/2}$	0,56
Коэффициент надежности по формуле Спирмена-Брауна	0,72
Дисперсия суммарных баллов	6,23
Дисперсия разностей между результатами половин теста	1,86
Коэффициент надежности по формуле Рюлона	0,70
Коэффициент надежности по формуле Кьюдера-Ричардсона	0,69

В данном примере значения коэффициента надежности как внутренней согласованности по формулам Спирмена-Брауна, Рюлона и Кьюдера-Ричардсона колеблются около 0,7, что говорит о средней надежности теста. Каждый из рассмотренных методов включает различные источники ошибки и поэтому подвержен различным искажениям. Например, ретестовая надежность в большей степени подвержена эффектам научения, нежели надежность по внутренней согласованности. Также ретестовая надежность снижается с увеличением промежутка времени между тестом и ретестом. Методы расщепления теста имеют серьезные преимущества по сравнению с ретестовым и методами параллельных форм главным образом

благодаря отсутствию необходимости в повторном обследовании. Таким образом снимается влияние многих посторонних факторов, в частности, тренировки, запоминания решений и т.д. К недостаткам методов расщепления теста относится невозможность проверить устойчивость результатов теста спустя определенное время.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	№ п/п	Номер задания теста										Индиви-дуальные баллы	Балл за задания с четным номером	Балл за задания с нечетным номером	Разность
2		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**				
3	1	0	1	1	0	0	0	0	1	1	1	5	3	2	1
4	2	1	1	1	1	1	1	1	0	0	1	8	4	4	0
5	3	1	0	0	0	1	0	0	0	0	0	2	0	2	-2
6	4	0	1	0	1	0	1	0	1	1	1	6	5	1	4
7	5	1	1	0	1	1	0	0	1	1	1	7	4	3	1
8	6	0	1	0	1	0	0	0	0	1	0	3	2	1	1
9	7	0	0	0	0	1	1	0	0	0	0	2	1	1	0
10	8	1	1	1	1	1	1	1	1	1	1	10	5	5	0
11	9	0	1	0	1	1	1	1	1	0	0	6	4	2	2
12	10	0	0	0	0	0	1	1	0	1	1	4	2	2	0
13	11	0	0	0	0	0	1	1	0	1	0	3	1	2	-1
14	12	1	1	0	1	0	0	1	0	1	1	6	3	3	0
15	13	1	0	0	1	1	0	0	1	0	0	4	2	2	0
16	14	1	0	0	1	1	0	1	1	1	1	7	3	4	-1
17	15	0	0	0	0	0	0	0	1	1	1	3	2	1	1
18	16	1	0	0	0	0	0	0	0	0	0	1	0	1	-1
19	17	1	0	0	0	0	0	1	0	0	0	2	0	2	-2
20	18	1	0	1	1	1	1	1	1	0	0	7	3	4	-1
21	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	20	1	0	1	1	0	0	0	0	1	1	5	2	3	-1
23	21	0	1	0	0	0	0	0	0	1	0	2	1	1	0
24	22	1	0	0	0	0	0	0	0	1	0	2	0	2	-2
25	23	1	0	0	0	0	0	0	0	1	1	3	1	2	-1
26	24	0	0	0	0	0	0	0	0	1	0	1	0	1	-1
27	25	0	1	0	0	0	0	0	1	1	1	4	3	1	2
28	26	1	0	0	0	0	0	0	1	0	1	3	2	1	1
29	27	0	1	1	1	1	1	1	1	1	1	9	5	4	1
30	28	0	1	1	1	0	0	0	0	1	1	5	3	2	1
31	29	0	1	0	0	0	0	0	0	1	0	2	1	1	0
32	30	1	0	0	1	1	0	0	0	1	0	4	1	3	-2
33	***Yi***	16	15	10	18	16	15	17	20	29	25				
34	***p***	0,53	0,50	0,33	0,60	0,53	0,50	0,57	0,67	0,97	0,83				
35	***q***	0,47	0,50	0,67	0,40	0,47	0,50	0,43	0,33	0,03	0,17	**Сумма:**			
36	***pq***	0,25	0,25	0,22	0,24	0,25	0,25	0,25	0,22	0,03	0,14	**2,099**			

Рис. 13. **Расчетная таблица для оценки надежности теста**

Коэффициент надежности теста позволяет определить стандартную ошибку измерения

$$S_E = S_x \cdot \sqrt{1-\rho} \qquad (18)$$

Полученное значение стандартной ошибки измерения можно использовать для построения доверительного интервала, в пределах которого с некоторой доверительной вероятностью находится истинное значение тестового балла испытуемого. Для построения доверительного интервала первоначально выбирается уровень ошибки. Как правило, используется 5-процентный уровень, при этом значению $\alpha = 0{,}05$ соответствует табличное значение t- распределения Стьюдента равное 1,96.Тогда доверительный интервал имеет вид:

$$\left(x_i - 1.96 \cdot S_E, x_i + 1.96 \cdot S_E\right) \qquad (19)$$

Валидность теста показывает, насколько хорошо тест делает то, для чего он был создан. Различают следующие основные *виды валидности*: *содержательную* – степень соответствия теста программам обучения и образовательным стандартам; *критериальную* – степень соответствия результатов тестирования внешнему, не относящемуся к тесту критерию; *квалиметрическую* – степень связи результатов математической обработки результатов тестирования и их интерпретации.

Количественно валидность чаще всего оценивается с помощью величины коэффициента корреляции между тестовой оценкой и некоторым независимым внешним критерием, то есть оценкой эксперта (преподавателя). Например, для теста, разработанного для оценки знаний, это корреляция между результатами теста и, например, семестровыми экзаменационными отметками. В этом случае в качестве независимого критерия берутся оценки, выставленные при традиционной проверке знаний студентов без использования тестов.

Важно проверить валидность теста по распределению. В валидном тесте результаты тестирования распределяются по нормальному закону. В соответствии с данным критерием подавляющая доля заданий в тесте должна быть средней трудности, но также обязательно должны присутствовать легкие и трудные задания. Если нормальность распределения результатов тестирования не выполняется, и тест не является валидным, необходимо заменить часть тестовых заданий, что достигается путем варьирования числа легких и трудных заданий в тесте.

Для определения оптимального времени тестирования проводится эксперимент на равнозначных группах испытуемых, не менее трех. Априори время тестирования определяется как трехкратная величина времени, затраченного профессионалом на ответы заданий теста. Апостериорно эта величина определяется по наибольшей дисперсии полученных результатов среди групп, выполнявших тест за разное время.

3.2. Задания для практического выполнения

1. Выполнить оценку надежности теста, разработанного на предыдущих занятиях с использованием формул Спирмена-Брауна, Рюлона и Кьюдера-Ричардсона.

2. Оценить ошибку измерения и построить доверительные интервалы для тестовых баллов испытуемых.

3. Проверить соответствие нормальному законураспределения тестовых баллов испытуемых.

4. Определить оптимальное время тестирования.

3.3. Контрольные вопросы по теме

1. Чем характеризуется надежность теста?
2. Какие виды показателей надежности вы знаете?
3. Что представляет собой и каковы недостатки метода повторных измерений? Параллельного теста?
4. Что представляет собой метод расщепления теста? В чем его преимущество?
5. Как рассчитать оценку надежности по формулам Спирмена-Брауна, Рюлона или Кьюдера-Ричардсона? Как интерпретируются полученные результаты?
6. Каким образом выполняется построение доверительного интервала для тестового балла испытуемого?
7. Что означает понятие «валидность теста»? Какие различают виды валидности?
8. Как оценивается количественно валидность теста?
9. Как оценивается валидность теста по распределению?
10. Как определяется оптимальное время тестирования?

СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ

1. Аванесов В.С. Композиция тестовых заданий / В.С. Аванесов. – М.: Центр тестирования, 2002. – 240 с.
2. Аванесов В.С. Теория и методика педагогических измерений [Электронный ресурс] // http://testolog.narod.ru/Theory.html.
3. Гласс Дж., Стенли Дж. Статистические методы в педагогике и психологии / пер. с англ. Л.И. Хайрусовой. – М., 1976. – 410 с.
4. Звонников В.И. Современные средства оценивания результатов обучения: учебное пособие / В.И. Звонников, М.Б. Челышкова. – М.: «Академия», 2007. – 224 с.
5. Звонников В.И., Челышкова М.Б. Контроль качества обучения при аттестации: компетентностный подход: учебное пособие / В.И. Звонников, М.Б. Челышкова. – М.: Университетская книга; Логос, 2009. – 272 с.
6. Звонников В.И., Челышкова М.Б. Современные средства оценивания результатов обучения / В.И. Звонников, М.Б. Челышкова. – М.: Издательский центр «Академия», 2008. – 224 с.
7. Майоров А.Н. Теория и практика создания тестов для системы образования. – М.: «Интеллект-Центр», 2002. – 296 с.
8. Поддубная Л.М. Задания в тестовой форме для автоматизированного контроля знаний студентов / Л.М. Поддубная, А.О. Татур, М.Б. Челышкова. – М.: МИФИ, 1995. – 80 с.
9. Поддубный А.В., Панина И.К., Ащепкова Л.Я. Методические основы разработки и использования педагогических тестов [Электронный ресурс] // http://www.dvgu.ru/umu/pedtest/Main.htm.
10. Челышкова М.Б. Теория и практика конструирования педагогических тестов: Учебное пособие. – М.: Логос, 2002. – 432 с.

Учебное издание

БУЛГАКОВА Наталья Валентиновна
ЧИРКИНА Анна Александровна

МЕТОДЫ МОНИТОРИНГА КАЧЕСТВА УЧЕБНОГО ПРОЦЕССА

Методические рекомендации

Технический редактор *Г.В. Разбоева*
Компьютерный дизайн *И.В. Волкова*

Подписано в печать 2014. Формат 60x84 $^{1}/_{16}$. Бумага офсетная.
Усл. печ. л. 2,03. Уч.-изд. л. 1,42. Тираж экз. Заказ .

Издатель и полиграфическое исполнение – учреждение образования
«Витебский государственный университет имени П.М. Машерова».

Свидетельство о государственной регистрации в качестве издателя,
изготовителя, распространителя печатных изданий
№ 1/255 от 31.03.2014 г.

Отпечатано на ризографе учреждения образования
«Витебский государственный университет имени П.М. Машерова».
210038, г. Витебск, Московский проспект, 33.