В. И. ЗВОННИКОВ, М. Б. ЧЕЛЫШКОВА

СОВРЕМЕННЫЕ СРЕДСТВА ОЦЕНИВАНИЯ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ

Совет по Восточной федеральный Университет им. М.К. Амосова

Рекомендовано
Учебно-методическим объединением
по специальностям педагогического образования
в качестве учебного пособия для студентов высших учебных заведений,
обучающихся по педагогическим специальностям

3-е издание, стереотипное

ACADEMIA

Москва
Издательский центр «Академия»
2009

УДК 37.01(075.8)
ББК 74.00я73
З-437

Рецензенты:

академик РАО, доктор педагогических наук, профессор *Е. В. Бондаревская*;
доктор педагогических наук, профессор, главный научный сотрудник
МГ ОПУ им. М. А. Шолохова *И. И. Легостаев*

Библиотека Восточный Университет им. К. Тимирязова

Звонников В. И.

З-437 Современные средства оценивания результатов обучения : учеб. пособие для студ. высш. учеб. заведений / В. И. Звонников, М. Б. Челышкова. — 3-е изд., стер. — М. : Издательский центр «Академия», 2009. — 224 с.

ISBN 978-5-7695-6196-2

Учебное пособие посвящено истории, методам и средствам педагогического контроля. В нем содержатся теоретические и методические подходы к конструированию педагогических тестов, анализируются их функции и возможности применения. Рассматривается использование тестов на различных уровнях: в школьных контрольно-оценочных системах, на национальных экзаменах и в системах мониторинга качества образования.

Предназначено для студентов высших педагогических учебных заведений, а также для учителей школ, преподавателей вузов и ссузов, использующих в своей профессиональной деятельности педагогические тесты.

УДК 37.01(075.8)
ББК 74.00я73

Оригинал-макет данного издания является собственностью Издательского центра «Академия», и его воспроизведение любым способом без согласия правообладателя запрещается

© Звонников В. И., 2007
© Образовательно-издательский центр «Академия», 2007
© Оформление. Издательский центр «Академия», 2007

ISBN 978-5-7695-6196-2

ПРЕДИСЛОВИЕ

Данное учебное пособие адресовано студентам педагогических вузов — будущим учителям и тем, кому в своей работе придется создавать или применять тесты. Как правило, тесты нравятся ученикам и студентам. Высокий темп работы, атмосфера соревновательности мотивируют учащихся. Предельная стандартизация заданий и процедур тестирования создает равные условия педагогического контроля и способствует реализации права на объективную оценку.

Преподаватели относятся к тестам более сдержанно, в целом позитивно, поскольку отдают должное высокой точности и сопоставимости тестовых оценок, возможности повышения эффективности контроля и проведения сравнительных исследований в образовании. Но при этом каждый учитель понимает, что тестирование открывает путь для сравнения его работы с работой других учителей, формирования управленческих выводов об эффективности обучения и осуществления внешнего контроля за результатами образовательной деятельности, что по понятным причинам не нравится даже эффективно работающим учителям. Именно в этом некоторые психологи усматривают одну из причин неприятия тестов, которое усугубляется в тех случаях, когда преподаватель авторитарен, нетерпим к мнению других и уверен в непогрешимости собственных оценочных суждений.

На протяжении всей истории развития тестов отношение к ним не было однозначным. В различных странах в ориентации общественного мнения за или против тестов побеждают то их противники, то их сторонники. Потребность в тестах резко снижается в те периоды, когда общество не нуждается в объективных средствах для выявления наиболее подготовленных людей, когда протекционизм и взяточничество становятся решающими факторами при приеме в вузы или на работу.

Критика тестов, если она не сводится лишь к отрицанию, а носит конструктивный характер, оказывает позитивное влияние на их развитие. Она порождает теоретические и прикладные исследования, способствует развитию аппарата теории педагогических измерений. Если сопоставить упрощенные подходы к конструированию тестов начала XX в. и современные методы компьютерного моделирования тестов из банка калиброванных заданий, то можно заметить большой прогресс, благодаря которому ликвидированы многие недостатки тестовых методов, повышена их объективность и технологичность. Даже наиболее убежденные про-

Электронный архив библиотеки МГУ имени А.А. Кулешова

тивники тестов вынуждены согласиться с тем, что сегодня, когда приходится принимать множество управленческих решений в образовании, полный отказ от тестов невозможен.

Многие учителя далеки от абсолютизации роли тестирования в учебном процессе, поскольку всегда анализируют обоснованность тестовых оценок и правомерность их применения для различных ситуаций в обучении, что вполне оправданно. В о-п е р в ы х, тесты — это только инструмент, средство осуществления педагогического контроля, и, как любое средство, они могут приносить пользу, если применяются по назначению, или быть неуместными, когда их функциональное назначение не адекватно ситуации применения. В о-в т о р ы х, тесты могут быть сделаны хорошо или плохо. В последнем случае они не обеспечивают ни высокой объективности, ни сопоставимости, поскольку требования теории педагогических измерений не выполняются. В-т р е т ь и х, даже очень качественные тесты при неумелом их использовании представляют опасность. Необходимы специальные знания для корректного выбора теста из числа имеющихся — в соответствии с целями измерения; инструктирования тестируемых; подбора адекватных методов шкалирования результатов учащихся; выравнивания шкал по отдельным вариантам и правильной интерпретации тестовых баллов при использовании результатов в учебном процессе. Негативные последствия неумелого применения тестов нередко отмечаются в тех случаях, когда при сопоставлении результатов учащихся не учитываются важные факторы, влияющие на результаты тестирования, например предыстория развития конкретного учащегося и дополнительная информация о нем, релевантная целям измерения.

Несмотря на трудности использования тестов, было бы ошибкой считать, что тестирование – удел исключительно профессионалов, обеспечивающих информацией о качестве обучения органы управления образованием, и что учитель в своей повседневной работе вполне может обойтись без него. В России массовый интерес к тестам в педагогической среде был вызван введением Единого государственного экзамена (ЕГЭ). После начала эксперимента ментальность многих учителей стала изменяться, поскольку они почувствовали несомненную пользу независимой объективной информации о качестве подготовки учащихся. Желание иметь учеников с более высокими результатами по ЕГЭ стимулировало обращение большинства российских учителей, участвующих в эксперименте, к педагогическим тестам. Соответственно пришло понимание того, что для подготовки или выбора тестовых заданий педагогу необходимо специальное обучение методикам разработки и применения тестов.

Материал пособия как раз этим и полезен. В него включены основные сведения по педагогическому контролю, ориентирован-

Учебно-методический центр
Национального института
образования

ные на современные оценочные средства. Благодаря этому читатели смогут переосмыслить некоторые ставшие давно привычными положения дидактики, вернее, той ее части, в которой затрагиваются вопросы контроля знаний учеников, лучше понять достоинства тестовых методов, а главное — принять определенные новшества и перестроить соответственно собственную работу.

В учебном пособии рассматриваются общие положения теории педагогических измерений, даются основные определения и понятия, излагаются подходы к отбору содержания теста. Отдельный раздел посвящен формам тестовых заданий, которые сопровождаются примерами и рекомендациями по разработке заданий. Значительное внимание уделяется математико-статистическим методам обоснования качества тестов. В пособии содержится краткая информация о некоторых инновационных средствах аутентичного оценивания, к которым, в первую очередь, относят портфолио. Приводятся различные методы шкалирования данных тестирования.

Отдельный раздел посвящен проблемам использования тестовых технологий в образовании. Сферы их применения связываются с экспериментом по введению ЕГЭ и с мониторингом качества школьного образования.

Всего в пособии 15 глав, каждая из которых завершается и упражнениями для самостоятельного выполнения. В конце книги приведен список литературы для более подробного изучения затронутых вопросов. В целом в пособии делается акцент на современное состояние практики использования тестов в России.

Книга адресована студентам педагогических вузов, обучающимся по разным специальностям, а также школьным учителям, не имеющим зачастую достаточной математической подготовки, поэтому многие наиболее сложные разделы по современным моделям педагогического измерения и адаптивному тестированию приведены в максимально простой форме, без использования специальных терминов по статистике, но и без излишнего упрощения и примитивизации.

Авторы выражают глубокую благодарность Виктору Васильевичу Фирсову за помощь в написании этой книги.

Оренбургский государственный университет имени И.И. Дементьева

ГЛАВА

1

ПЕДАГОГИЧЕСКИЙ КОНТРОЛЬ В УЧЕБНОМ ПРОЦЕССЕ

1.1. Педагогический контроль, его структура и содержание

Общая характеристика педагогического контроля. Педагогический контроль представляет собой единую дидактическую и методическую систему проверочной деятельности, которая протекает при руководящей и организующей роли педагогов, носит совместный характер, объединяя преподавателей и учащихся, и направлена на оценку результатов учебного процесса. С помощью контроля можно оценить достижения учащихся и выявить пробелы в их знаниях, установить взаимосвязь между планируемыми, реализуемыми и достигнутыми уровнями образования, понять достоинства и недостатки новых методов обучения, сравнить работу преподавателей, дать руководителю учебного заведения объективную информацию для принятия управленческих решений и выполнить ряд других не менее важных задач [12; 28; 42; 47].

Основные компоненты контрольно-оценочной деятельности. К основным компонентам контрольно-оценочной деятельности относятся:

– выделение тем, разделов и т.д., выступающих в роли понятийных индикаторов;

– операционализация понятий путем формирования эмпирических индикаторов (вопросов, заданий и т.д);

– создание модели желаемых результатов контроля;

– проведение контрольных мероприятий;

– сличение модели и реальных ответов учащихся;

– формирование оценочных суждений и принятие на их основе решения о продолжении контроля или выставлении оценок.

Перечисленные компоненты всегда присутствуют в структуре контрольно-оценочной деятельности педагога. Они интерпретируются и раскрываются по-разному в зависимости от того, проводится контроль традиционными средствами или применяются тесты. При тестовом контроле операционализация понимается как процедура перехода от понятийных индикаторов к эмпирическим референтам, в роли которых выступают задания теста. Модель результатов предельно стандартизируется и задается в виде правил оценивания и ключа ответов, сличение же результатов тестируе-

Свободный
Федеральный университет
им. М.К. Аммосова

мых с ключом ответов проводится с минимизацией субъективного фактора автоматизированно либо с помощью экспертов.

Содержание контроля. Среди компонентов структуры контрольно-оценочной деятельности педагога наибольшую важность представляет выделение понятийных индикаторов, определяющее содержание контроля и предмет оценивания [7; 19; 47 и др.]. В повседневном учебном процессе реализуется контроль за усвоением знаний, который направлен на выявление и оценку результатов приобретения новых знаний учащимися, включая анализ этого процесса. Ему свойственен, как правило, пооперационный характер (Н. Ф. Талызина). Содержание контроля, предназначенного для оценки результатов учебного процесса, связано с трактовкой обученности, которую психологи рассматривают в контексте зависимости от обучаемости и совокупности свойств интеллекта (И. А. Зимняя). Педагоги же при оценке обученности ставят в центр внимания уровни освоения содержания обучения и способы учебной деятельности (В. М. Блинов и др.).

Содержание контроля не только отражает, чему учат и что хотят видеть в качестве результатов обучения, но и задает определенные приоритеты при обучении. Требования, предъявляемые к учебным достижениям в процессе контроля, неизбежно становятся ориентирами для учителя в его повседневной работе, особенно в тех случаях, когда контроль имеет внешний характер, а оценки используются для принятия административно-управленческих решений в образовании. Подобное влияние наблюдается сегодня в связи с развитием эксперимента по введению ЕГЭ [9]. Оно может иметь деструктивные последствия для образования, привести к дисбалансу в правильных акцентах при обучении. Например, при постоянном применении некачественных тестов, включающих в основном задания для оценивания фактов и понятий, возможно усиление внимания учителей к изложению фактологического материала вместо выявления сути фактов и законов, развития творческой деятельности учащихся и обучения их применению теоретических знаний на практике.

Стандартизация требований к освоению предметного содержания в контроле. Содержание контроля, нацеленного на результаты учебного процесса по отдельным предметам, задается в государственных образовательных стандартах (ГОС) обязательным минимумом содержания и требованиями к уровню его освоения выпускниками системы общего образования. В настоящий момент продолжается процесс совершенствования структуры ГОС. Происходит операционализация требований — приспособление формы представления требований к задачам измерения [8]. Также вводится уровневая дифференциация требований ГОС, обеспечивающая дополнительные стимулы для повышения продуктивности учебной деятельности (В. В. Фирсов) [67].

РЕПОЗИТОРИЙ ГГУ ИМЕНИ Ф. СКОРИНЫ

Опыт разработки и применения ГОС в других странах показал, что мотивация учащихся к реализации поставленных целей обучения растет, если четко выделены критерии достижений ученика и свидетельства его продвижения, задана и детально описана прогрессирующая последовательность уровней достижений и учащийся информирован о том, что нужно сделать, чтобы достичь более высоких оценок.

Термины и понятия педагогического контроля. В данном пособии предпочтение отдается термину «контроль учебных достижений». Учебные достижения рассматриваются как итоговые результаты обучения и являются синонимом термина «подготовленность». В общем случае под учебными достижениями иногда понимают не только подготовленность обучающихся в определенной предметной области, но и показатели сформированности личностных качеств обучаемого.

В педагогическом контроле используется также термин «оценивание» (Assessment). *Оценивание* — это процесс формирования оценки учебных достижений, в котором интегрируются и представляются в определенной шкале (шкалах) данные, полученные при тестировании, использовании портфолио, проведении экзаменов, выполнении практических работ учащимися, рейтинговании их результатов и т.д. [73].

Тестирование (Testing) — это процесс применения тестов, который является частью процесса оценивания. В данном пособии этот термин используется только в тех случаях, когда речь идет о предъявлении научно обоснованных тестов, обладающих необходимыми статистическими характеристиками и обеспечивающих высокое качество измерений.

Новый термин для нашей литературы — «тестирование административно-управленческого предназначения» (High-Stakes Testing) [91] связан с принятием наиболее ответственных решений в образовании. В России примером High-Stakes Testing можно считать тестирование, проводящееся при аттестации школ и вузов. Наиболее яркий пример — тестирование в ЕГЭ, которое служит для аттестации выпускников школ и отбора абитуриентов, анализа последствий образовательных реформ, сравнения качества подготовленности выпускников различных регионов России и отдельных районов внутри региона, эффективности деятельности учебных заведений и т.д.

Электронный архив библиотеки МГОУ им. Н.К. Крупской

1.2. Виды контроля в учебном процессе

Общие подходы к классификации. По традиционной классификации видов педагогического контроля в обучении выделяются входной, текущий и итоговый контроль [42]. Процессы самоконт-

роля (самооценки) в эту классификацию не включаются. Их рассмотрением обычно занимается педагогическая психология, поскольку они связаны с механизмом перехода внешних контролирующих воздействий учителя во внутреннее состояние учащегося.

Входной контроль. Входной контроль в практике школьного обучения систематически не проводится. Он используется лишь в случае отбора учащихся при конкурсном зачислении в профильные классы или является инициативой педагогов, уделяющих большое внимание индивидуализации учебного процесса. В условиях личностно ориентированного, развивающего обучения входной контроль помогает построить индивидуальные траектории освоения нового материала для наиболее слабых или наиболее сильных учащихся, при отказе от традиционной ориентации на гипотетического среднего ученика [48; 58]. Наиболее эффективным средством осуществления входного контроля, который чаще всего носит характер экспресс-диагностики, являются педагогические тесты.

Текущий контроль. Цель текущего контроля — следить за ходом обучения. Его осуществление позволяет преподавателю получить оперативную информацию о ходе учебного процесса для его своевременной коррекции и перестройки в нужном направлении. Наибольший интерес представляют данные о динамике усвоения каждым обучаемым нового материала, степени рациональности его мыслительных процессов или алгоритмов при выполнении заданий, так как при правильно организованном учебном процессе учитель должен контролировать не только содержание выполняемых учащимися действий, но и их свойства. Получение подобной информации возможно лишь при выявлении причин затруднений и ошибок учащихся, которые анализируются в ситуациях, когда текущий контроль приобретает явно выраженный диагностический характер. Повышение эффективности и усиление диагностического характера обратной связи в текущем контроле становятся возможными в тех случаях, когда на помощь учителю приходят компьютеры и диагностические тесты.

Итоговый контроль. Итоговый контроль (поэтапный, рубежный, заключительный контроль) предназначен для оценки учебных достижений после завершения определенного этапа обучения, прохождения раздела или всего учебного курса. Обычно формой итоговой оценки обучаемого является либо его отметка на экзамене (устном или письменном), либо результат выполнения теста. Сравнительный и прогностический анализ результатов итогового контроля дает учителю важную информацию, необходимую для улучшения своей работы в будущем. Данные анализа позволяют выявить систематические проблемы в подготовке учащихся и осуществить управленческие действия по коррекции процесса

Северо-Восточный федеральный университет им. М.К. Аммосова

обучения, если его результаты не согласуются с поставленными целями.

Итоговый контроль может быть внешним или внутренним. Внешний итоговый контроль проводят не зависимые от школы структуры, например при государственной аттестации или в ЕГЭ. В образовании под *аттестацией* понимается процедура установления соответствия уровня и качества подготовки выпускников зафиксированной документально системе требований к уровню и качеству образования. Роль общепризнанной нормы играют требования ГОС или других нормативных документов, действующих в условиях отсутствия стандартов. Обоснованность аттестационных оценок достигается репрезентативным отображением требований ГОС в содержании контролирующих материалов, роль которых во многих странах выполняют тесты. Внутренний итоговый контроль проводят сами учителя, например во время школьных выпускных экзаменов.

1.3. Функции контроля

Общие замечания. В большинстве учебников для системы педагогического образования в качестве основных функций педагогического контроля выделяются контролирующая, диагностическая, обучающая, воспитывающая и мотивирующая функции. В ходе исторического развития педагогической науки и появления представлений о контроле как составляющей управления качеством образования добавились информационная, сравнительная и прогностическая функции.

Контролирующая функция. Контролирующая функция является основной для итогового контроля и вряд ли нуждается в развернутых пояснениях в силу очевидности состава своих компонентов для различных уровней управления образованием. Она предполагает осуществление систематического контроля за результатами обучения, определение состояния усвоенных знаний, умений и навыков и находит свое отражение в оценке учебных достижений. По сфере приложения результатов контролирующая функция может быть связана с различными уровнями управления качеством образования.

Диагностическая функция. Диагностическая функция наиболее полно реализуется в текущем контроле. Активизация роли диагностической функции является важным условием повышения качества учебного процесса. Благодаря детальному анализу причин и характера затруднений учащихся педагогическая диагностика открывает новые возможности в индивидуализации обучения, поскольку каждый ученик приступает к изучению нового материала только после устранения всех пробелов в знаниях, препят-

Северо-Восточный федеральный университет им. М.К. Аммосова

ствующих усвоению следующих разделов курса. Диагностическая функция текущего контроля осуществляется с помощью традиционных средств (контрольных работ, опросов и т.д.) или специальных диагностических тестов, о которых будет рассказано в разделе 5.2 [28; 54].

Обучающая функция. Обсуждения, возникающие в классе при ответах учащихся на вопросы преподавателей, самоконтроль и самооценка учеников при подготовке к контрольным работам и опросам выполняют, хотя и не в полной мере, обучающую функцию, поскольку традиционные средства и методы контроля лишь частично обеспечивают возможность ее реализации. Решающая роль здесь, как и во многих других ситуациях, принадлежит педагогическим тестам.

Особенно эффективно обучающий потенциал тестов позволяют реализовать автоматизированные контрольно-обучающие программы, опирающиеся на банки калиброванных заданий, адаптивные алгоритмы для выдачи обучающих заданий и современные модели теории педагогических измерений [60; 62; 63]. Сочетание известных оценок трудности тестовых заданий, помещенных в банк, и прогностических возможностей моделей, определяющих вероятность правильного ответа учащегося в зависимости от соотношения между его подготовленностью и трудностью контрольных заданий, открывает широкие возможности для индивидуализации процесса усвоения знаний обучаемыми без роста трудозатрат со стороны педагога.

С помощью контрольно-обучающей программы диагностируются трудности, с которыми приходится сталкиваться учащимся при выполнении заданий, устанавливаются их причины, а затем на основе анализа этих причин учащимся предъявляются разъясняющие образцы правильных действий и рекомендуются индивидуальные программы коррекции, дифференцированные в соответствии с результатами диагностики.

Воспитывающая и мотивирующая функции. Воспитывающая функция контроля проявляется в становлении таких позитивных качеств личности учащегося, как интерес к знаниям, умение систематически работать, навыки самоконтроля и самооценки. Ученики изучают предмет глубже и серьезнее, если заранее известно, что по нему будет осуществляться постоянный контроль.

Воспитывающая функция контроля призвана играть ведущую роль в формировании мотивационной основы учебной деятельности учащихся. В повседневном учебном процессе мотивация к обучению при контроле создается не всегда и не везде. Если контроль объективный и оценки преподавателя справедливы, то у учащихся возникают дополнительные стимулы к усвоению новых знаний. В случае преобладающего субъективизма в оценках педагога систематическая проверка знаний, как правило, приводит к отрица-

Учебно-методический центр Гомельского государственного университета имени Ф. Скорины

тельным результатам. У школьников появляются недоброжелательное отношение к преподавателю и полное нежелание учиться.

Росту мотивации учебной деятельности способствуют тесты и отметочные (балльные) критерии оценивания. Критерии задаются на описательном уровне и содержат совокупность требований к учебным достижениям для каждой оценки. Они не только побуждают учащихся к более высоким достижениям, но и вселяют в них уверенность в объективности педагога и «прозрачности» процесса выставления оценок.

Развивающая функция. Контроль укрепляет память и тренирует мышление, формирует умения и навыки применения знаний на практике, словом, способствует осуществлению развивающей функции обучения. Полнота реализации развивающей функции контроля тесно связана с характером проверочных заданий, их содержанием и уровнем деятельности, необходимой для их выполнения. Узкопредметная направленность заданий в сочетании с воспроизведением знаний в знакомой ситуации вряд ли будет способствовать активизации развивающей функции контроля. Задания на применение знаний в измененной или незнакомой ситуации, наоборот, заставляют учащегося анализировать, обобщать, оценивать и привлекать элементы творчества при решении поставленных проблем.

Информационная функция. Информационная функция присуща самой природе контроля вне зависимости от его вида, сферы приложения результатов и средств осуществления. В силу очевидности она не нуждается в дополнительных пояснениях. Необходимо лишь отметить, что полнота реализации информационной функции напрямую зависит от степени объективности данных контроля, которая должна расти по мере повышения ответственности принимаемых по результатам контроля управленческих решений. Например, высокий уровень объективности требуется при использовании данных контроля в случаях аттестации, идентификации проблем в системе обучения, связанных с введением инноваций, недостатками в методах преподавания, искажениями в пропорциях учебных планов, просчетами авторов школьных учебников и др.

Сравнительная функция. Сравнительная функция контроля проявляется при сопоставлении данных тестирования по школе с нормами районного или регионального уровня для выявления отставания отдельных школ по ряду показателей качества образования. Подобное сопоставление будет корректным при условии проведения специального шкалирования, статистической коррекции данных тестирования, выравнивания шкал, формирования репрезентативных выборок учащихся и учета социально-экономических факторов, влияющих на состав учащихся, кадровый подбор преподавателей и обстановку в районе расположения школы.

Гомельский государственный университет имени Ф. Скорины

Некорректность сопоставления может обернуться серьезными негативными последствиями в тех случаях, когда его результаты используются для принятия решений по вопросам о размерах финансирования отдельных школ, продвижении преподавателей по службе, размерах оплаты их труда или освобождении от должности и т.д.

Прогностическая функция. Прогностическая функция контроля предназначена для выявления способности к усвоению нового материала и неизбежно отражает воздействие предшествующего обучения, и потому ее можно предсказать по результатам контроля. Если для получения результатов контроля использовались лишь традиционные средства, то спрогнозировать вероятную успешность обучения того или иного испытуемого только на их основе будет невозможно. Однако задача становится вполне разрешимой, если применить тесты, прогностическая способность которых была заранее подтверждена специальными аналитическими методами [1; 5; 22]. Например, контрольные измерительные материалы (КИМ) ЕГЭ обладают высокой прогностической способностью, поскольку позволяют не только выделить лучших абитуриентов в момент приема в вуз, но и достоверно предсказать успешность их дальнейшего обучения.

Северо-Восточный федеральный университет им. М.К. Аммосова

1.4. Принципы контроля

Общая характеристика. Направляющая и регламентирующая роль по отношению к процессу контроля принадлежит дидактическим принципам (научности и эффективности, иерархической организации, систематичности, объективности, справедливости и всесторонности), которыми должны руководствоваться педагоги в своей контрольно-оценочной деятельности [54].

Принцип научности. Принцип научности предписывает использование при контроле научно обоснованных средств, проверочных процедур и методов анализа данных контроля для оценивания подготовленности учащихся, что можно выполнить далеко не всегда. Например, как оценить качество контрольной работы, сплошь и рядом применяемой в повседневном учебном процессе? Для этого приходится привлекать не теоретические знания, а здравый смысл и обычный педагогический опыт. То содержание контроля, которое нравится одному преподавателю, может не подойти другому, так как у каждого из нас есть свои представления о том, что необходимо проверять, по каким критериям делать выводы и оценивать результаты проверки.

В отличие от традиционных средств педагогические тесты разрабатываются по специальной технологии, опирающейся на теорию педагогических измерений. Они имеют научно обоснованные

критерии качества. Принцип научности предполагает проверку точности и устойчивости данных тестирования (*критерий надежности*) и адекватности данных поставленной цели измерения (*критерий валидности*). Подробно вопросы качества результатов педагогических измерений рассматриваются в главе 9.

Принцип иерархической организации. Принцип иерархической организации нацеливает на построение определенной иерархии знаний, умений и навыков при отборе содержания контроля. Преподавателю приходится сталкиваться с тем, что все проверить просто невозможно. Поэтому принцип иерархической организации предполагает отбор наиболее значимых, укрупненных элементов содержания для отображения их в тестах или в традиционных оценочных средствах.

Принцип систематичности. Принцип систематичности педагогического контроля находится в определенной зависимости от плановости последнего. Неравномерное увеличение частоты проверок, их неожиданность создают дополнительное нервное и эмоциональное напряжение у учащегося. В равной степени и уменьшение числа контрольных проверок имеет негативные последствия. Опыт свидетельствует о том, что отрицание необходимости контроля, отказ от совершенствования форм и методов контроля приводят к ухудшению качества обучения.

Принципы объективности и справедливости. Принципы объективности и справедливости довольно тесно связаны между собой, поскольку объективность является необходимым условием справедливости. Для реализации принципов в практике обучения необходимо ввести представление об объективных оценках. В теории педагогических измерений понятие «объективная оценка» замещается понятием «истинная оценка», которое характеризует уровень подготовленности учащегося в момент проведения контроля и не содержит ошибочного компонента. Поэтому можно говорить о большей или меньшей объективности оценок в зависимости от величины ошибочного компонента измерений (более подробно связь ошибочного и истинного компонентов рассматривается в главе 4).

Первый путь к повышению объективности контроля связан с формированием коллегиальной экспертной оценки. Успех при этом определяется качественным составом проверяющей комиссии и четкостью соблюдения инструкций экспертами. Другой путь повышения объективности результатов контроля, который не нашел применения в практике в силу низкой эффективности, предполагает создание определенных эталонов усвоения знаний, умений и навыков обучаемых по каждому предмету [28; 29]. И наконец, третий путь связан с тестированием, в котором полнота реализации принципа объективности контроля зависит от надежности тестов, стандартизации условий проведения тестирования, тех-

ФГБОУВПО «Костромской государственный университет им. Н.А. Некрасова»

нологии выставления оценок и всей совокупности факторов, влияющих на величину ошибки измерения. Особенно эффективно проблема объективизации решается в современной теории тестов, согласно которой на основе математических моделей измерения осуществляется переход от сырых баллов испытуемых к наиболее правдоподобным оценкам, дающим оптимальное приближение к истинным компонентам измерения [54; 70; 72; 74].

Принцип всесторонности. Принцип всесторонности подчеркивает необходимость тщательного отбора содержания контроля, которое должно репрезентативно отражать содержание учебных программ и видов проверяемой учебной деятельности. Задания, предназначенные для проверки, должны охватывать по возможности максимально широкий круг вопросов, подлежащих контролю, и не дублировать друг друга.

1.5. Психолого-педагогические аспекты педагогического контроля

Общие замечания. Психолого-педагогические аспекты контроля связаны с анализом уровня сформированности навыков самоконтроля и самооценки у обучаемых, являющихся важным показателем качества учебной деятельности [19]. Формированию навыков самоконтроля в значительной степени препятствуют авторитарные методы проведения проверок, когда ведущая роль в контроле принадлежит преподавателю, а подчиненная и пассивная отводится ученику. Альтернативой авторитарности служат идеи сотрудничества преподавателя и учеников, базирующиеся на помощи обучаемому в осознании им своих задатков, склонностей и способностей в условиях превращения социально значимых целей обучения и воспитания в личностно значимые.

Самоконтроль и самооценка. Под *самоконтролем* понимаются действия обучаемых, проявляющиеся в навыках осуществления контроля за результатами собственной деятельности и коррекции ее в процессе выполнения учебных заданий [19; 46]. Результат процесса самоконтроля — *самооценка*, которая может быть и завышенной, и заниженной в зависимости от психологических особенностей человека. Хотя уровень сформированности навыков самоконтроля педагоги относят к показателям полноценности обучения, они не умеют их отслеживать и оценивать. Трудность здесь состоит в том, что эти навыки раскрываются при анализе механизма перехода контрольно-оценочных процессов, осуществляемых учителем, во внутреннее состояние учащегося, которое характеризуется новыми действиями по самоконтролю и самооценке в структуре учебной деятельности и плохо поддается внешней оценке.

Полесский государственный университет

С точки зрения П. П. Блонского, процесс перехода внешних контрольно-оценочных процессов во внутренние навыки самоконтроля имеет четыре стадии, на которых уровень внешнего проявления навыков самоконтроля снижается и теряет внешний характер проявления по мере продвижения учащегося в усвоении учебного материала [42]. Участие преподавателя в формировании навыков самоконтроля вполне реально на первой стадии, когда учащийся еще не усвоил материал. Вторая стадия предполагает наличие сформированных навыков самоконтроля. На ней учащийся репродуцирует усвоенный материал и контролирует свои знания совместно с педагогом. Третья стадия отличается ростом уровня усвоения, благодаря которому самоконтроль приобретает выборочный характер, а участие педагога уменьшается, становится минимальным к четвертой стадии полного усвоения учебного материала. На четвертой стадии самоконтроль теряет всякие внешние проявления, становясь интерпсихическим свойством обучаемого. Таким образом, основные усилия педагога по формированию навыков самоконтроля у обучаемых должны быть сосредоточены на первой и второй стадиях путем выбора оптимального режима контроля и средств его осуществления.

Сотрудничество педагога и обучаемых в контроле. Идеи сотрудничества с обучаемым как отдельная область исследований появились в педагогической науке в середине 80-х гг. XX в. в качестве перспективного направления инноваций. Сотрудничество связано с гуманистической идеей помощи обучаемому, которая особенно трудно реализуется в педагогическом контроле. В условиях сотрудничества текущий контроль должен влиять на характер учебной деятельности, ее интенсивность, направленность и обеспечивать потребность в добровольном обращении обучаемых за помощью преподавателя для решения поставленных задач.

Необходимые условия для возникновения сотрудничества в процессе текущего контроля. Создание необходимых условий сотрудничества педагога и обучаемых в текущем контроле связано с анализом мотивационной сферы учебной деятельности. Главной причиной снижения мотивации к обучению можно считать отсутствие специальных механизмов оптимизации трудности учебных задач, отвечающих принципам доступности и вместе с тем достаточно высокой трудности. В свое время эти принципы были положены в основу дидактической системы Л. В. Занкова. Эксперимент по ее внедрению показал, что при нарушении принципа доступности учащийся пытается механически списать готовые решения во время контроля или полностью отказывается от поставленных перед ним учебных задач. Недостаточная же трудность учебных заданий приводит к неполной реализации развивающей функции контроля, что исключает потребность в сотрудничестве с педагогом.

Учебно-методический комплекс
Электронный архив библиотеки
МГУ имени А.А. Кулешова

Противоречивость между принципами доступности и достаточно высокой трудности является условной, если под доступностью понимать не легкость, а работу на пределе возможности обучаемого, когда задания слишком трудны для самостоятельной работы, но вполне могут быть выполнены с помощью педагога. Обращение за помощью создает основу для сотрудничества, поэтому необходимые условия сотрудничества педагога и обучаемых в текущем контроле связаны с оптимизацией трудности контрольных заданий в индивидуализированном режиме, с учетом отличий в подготовке учащихся.

Вопрос в том, как реализовать это технологически, связан с наличием банка тестовых заданий, имеющих устойчивые, статистические оценки трудности (см. главу 9), и специальных шкал, позволяющих сопоставить уровень подготовленности учащихся и трудность тестовых заданий благодаря введению общих единиц измерения (см. главу 12).

Зоны развития личности обучаемого. Психологическая трактовка основных положений педагогического контроля в условиях сотрудничества учителя и учащихся связана с концепцией Л. С. Выготского о зонах развития личности [16]. Зоне актуального развития личности обучаемого должны соответствовать задания, с которыми он может справиться самостоятельно. Оценка знаний школьника, полученная на основе таких заданий в итоговом контроле, является соответствующим показателем уже освоенного им учебного материала.

Если оценки достаточно объективны, то при выполнении освоенных заданий необходимости в помощи преподавателя обычно не возникает.

Задания, соответствующие зоне ближайшего развития, обучаемый может выполнить правильно только с помощью педагога, благодаря сотрудничеству с которым у обучаемого расширяется область знаний и умений, расширяются границы освоенных заданий за счет перехода в их число новых заданий, относившихся ранее к зоне ближайшего развития ученика.

Таким образом, введенные Л. С. Выготским зоны развития личности обучаемого помогают наметить пути формирования отношений сотрудничества между преподавателем и учеником в процессе текущего контроля.

Необходимым условием возникновения таких отношений между учеником и преподавателем является подбор контрольных заданий оптимальной трудности, соответствующих зоне ближайшего развития каждого обучаемого.

Практические методы отбора заданий, соответствующих по трудности зоне ближайшего развития каждого школьника, рассматриваются в ряде исследований по адаптивному тестированию и в пособии по разработке педагогических тестов [59; 60].

Федеральный государственный бюджетный университет имени К. А. Аммосова

Практические задания и вопросы для обсуждения

1. Вспомните свое обучение в средней школе. Имелись ли факторы, которые отрицательно повлияли на ваше отношение к тестам? Были ли эти негативные факторы связаны непосредственно с самими тестами или у вас вызывали неприятие результаты тестирования и способы их интерпретации учителем на уроках?

2. Отметьте некоторые социальные тенденции (например, изменение демографической ситуации, рост народонаселения, развитие рыночных отношений, конкуренция за рабочие места), повышающие важность введения тестирования в школах.

3. Вы только что узнали, что директор школы в качестве эксперимента готов выплатить финансовое поощрение учителю, если результаты тестирования в его классе будут улучшаться на протяжении трех лет. Как, по вашему мнению, относится к этому эксперименту: а) учитель; б) родители ученика; в) учащийся? Обоснуйте свою точку зрения.

4. Как вы считаете, будет ли возрастать роль тестов в российских школах на протяжении ближайших десяти лет? Аргументируйте свой ответ.

5. Какие функции контроля вы считаете приоритетными для повышения качества обучения? Следует ли с помощью контроля активизировать принуждение к обучению?

6. Ниже представлены ситуации, выявленные с помощью контроля. Какой вид контроля вы бы использовали для их определения? В каких случаях лучше использовать тесты, а в каких — традиционные средства контроля?

- Из всего класса только 10 учащихся достигли по своим результатам необходимой скорости чтения, а 15 пока не научились хорошо читать.
- Иванов на уроке математики получил отметку «5», а Петров — «2».
- Сидоров с трудом различает звонкие и глухие согласные.
- Мой метод преподавания не эффективен для этого класса.
- Петя нуждается в моей помощи для развития навыков устного счета.
- Сидорова необходимо перевести в группу успевающих учеников.
- У моего коллеги, преподавателя русского языка, класс оказался сильнее моего.
- Миша медленно усваивает новый материал.

Какой вид контроля вы используете для принятия каждого из 8 решений? Когда лучше использовать тесты, а в каких случаях можно полагаться на традиционные средства контроля?

Сибирский Федеральный Университет им. М.К. Аммосова

ГЛАВА

2

КОНТРОЛЬ, ОЦЕНКИ И ЭВАЛЮАЦИЯ В ОБРАЗОВАНИИ: РАЗВИТИЕ И СОВРЕМЕННОЕ СОСТОЯНИЕ

2.1. Исторические аспекты развития контроля и оценки в образовании

Становление контроля и оценки в образовании. Контроль в том или ином виде всегда присутствует в обучении. В процессе исторического развития образовательной практики менялись лишь формы и средства осуществления проверок, приоритеты в оценках и приемы их выставления, интенсивность проведения контрольных мероприятий, меры воздействия на учащихся, а также акценты при интерпретации результатов контроля в образовании. Истории педагогического контроля посвящены многочисленные публикации российских и зарубежных авторов [37; 45; 77]. В данном пособии приводится лишь краткий обзор основных этапов становления контроля в образовании и выделяются ключевые моменты в истории его развития.

Отдельные теоретические представления о контроле сложились довольно давно, в конце XVIII — начале XIX в. Они касались в основном проверки и оценки репродуктивных знаний учащихся, за воспроизведение которых по образцу, предложенному педагогом, выставлялись оценки. В целом в XVIII и XIX вв. контроль рассматривался исключительно в контексте принуждения к обучению, подводил итог определенным его результатам и акцентировал воспитательные функции оценок.

В XIX в. во многих странах усилилось внимание к личности обучаемого, стала острой проблема справедливости оценок. В частности в России для совершенствования средств и методов контроля на начальном этапе обучения для отсева малоспособных детей предлагались специальные испытания (Прокопович), осуществлялись проверка условий самостоятельной работы учащихся во внеучебное время (С. С. Татищев), оценка внимательности учащегося (К. Д. Ушинский), среди учащихся проводился еженедельно самоанализ ошибок и затруднений (К. М. Новиков) и др. [37].

Контроль и оценка знаний в отечественном образовании начала XX в. Педагогические воззрения в первые годы XX в. в России характеризовались нарастанием гуманистических тенденций. Акценты с оценки результатов обучения сместились на процесс приобретения знаний, настойчивость учащихся и динамичность осво-

Северо-Восточный федеральный университет им. М. К. Аммосова

ения ими нового учебного материала. В контроле стали учитываться индивидуальные психологические характеристики учеников, их подготовленность к началу обучения, семейные условия и социально-экономическая среда. Результаты контроля и дополнительную информацию об учащемся рекомендовалось выражать в оценочных суждениях и отметках.

Стремление к гуманизму было характерно и для послереволюционной школы, хотя идеи демократизации отношений учителя и учащихся в те годы нередко доводились до абсурда. В частности, согласно предписаниям органов управления образованием учитель утрачивал свои контролирующие функции, превращался в советчика и старшего товарища обучаемых. Согласно Постановлению Наркомпроса от 31(18) мая 1918 г. роль контроля в образовании сводилась к нулю, вплоть до полной отмены баллов, экзаменов и индивидуальных проверок обучаемых. Несмотря на директивные документы Наркомпроса, многие педагоги, заботившиеся о качестве обучения, пытались всячески сохранить текущий контроль, приспособив его к официальной позиции органов управления образованием. Правоту сторонников контроля в образовании подтвердила сама жизнь. Отмена экзаменов и проверок в 1918 г. привела к снижению качества знаний учащихся, ухудшению дисциплины и мотивации учебной деятельности.

Контроль и оценка знаний в 20—60-е гг. XX в. В начале 20-х гг. XX в. наметилось некоторое смягчение официальной позиции по отношению к педагогическому контролю: стала допускаться проверка знаний учащихся с помощью письменных работ и собеседований, была введена практика проведения зачетов и применения тестов. Создание в 1922 г. Особой Коллегии по учету работы школ при научно-педагогической секции ГУС (Государственного ученого совета) Наркомпроса и проведение в 1923 г. специальной конференции по учету педагогической работы в школе послужили сигналом того, что в школах появились затруднения, вызванные отменой оценок. Начиная с 1926 г., учителям разрешалось высказывать оценочные суждения, но только в словесной форме, а наиболее приемлемыми формами контроля считались дневники учащихся, дискуссии, рефераты, коллективные отчеты и тесты, зачеты, вопросы и письменные контрольные работы.

Усиление позиций педагогического контроля наметилось в 30-е гг. XX в. и продолжалось вплоть до 50-х гг. XX в. на фоне утверждения административного стиля руководства во всех социальных сферах, в том числе и в образовании. Широкое распространение в школе получили идеи авторитарности, согласно которым в контрольно-оценочной деятельности педагога стала доминировать функция принуждения к обучению. В педагогической литературе тех

Северо-Восточный федеральный университет им. М.К. Аммосова

лет рекомендовалось усиливать ведущую роль учителя, улучшать способы работы учеников под руководством педагога, акцентировать контролирующие функции учета знаний [46; 50; 72]. Однако качество знаний от этого не повышалось. Установка на ужесточение контроля порождала, с одной стороны, вполне естественное противодействие учащихся, а с другой — приводила к утверждению антигуманизма, формализму, безответственному отношению к реальным результатам обучения.

Контроль и оценка знаний во второй половине XX в. С начала 60-х вплоть до начала 90-х гг. XX в. на страницах многих периодических изданий по проблемам образования развернулось обсуждение существующей балльной системы оценивания в школе, в котором преобладала критика отметок. Педагоги и ученые отмечали формальный характер традиционной системы контроля, фетишизацию отметок, отсутствие объективности цифровых баллов и процентоманию, характерную для отчетов школ о своей работе [25; 47; 58].

Причины субъективизма балльной системы в публикациях тех лет обычно связывались не с отсутствием стандартизированных средств контроля, а с неоднозначностью описания целей обучения и требований к уровням усвоения знаний. Ученые предлагали различные пути совершенствования контроля, основанные на введении научно обоснованных нормативов результатов усвоения, типологии знаний, специальных показателей успеваемости, как правило, слишком субъективных и надуманных, чтобы быть действительно полезными в учебном процессе. В основном эти подходы были пригодны лишь для проверки простейших уровней учебной деятельности и не затрагивали творческие уровни ее осуществления [59].

В 60-е гг. XX в. стремление к объективизации оценок подготовленности учащихся в определенной степени способствовало распространению программированного контроля [7; 17; 54]. В зависимости от вида обучающих программ (линейные, разветвленные, адаптивные) в программированном обучении использовались особые приемы проверки и коррекции результатов обучения. В силу отсутствия в те годы педагогических тестов и навыков по их разработке при программированном контроле проверялись наиболее простые виды учебной деятельности, задания имели упрощенный вид и предполагали выбор одного или нескольких готовых ответов, а скрытые психологические составляющие процесса усвоения, понимание материала, логика умозаключений учащихся, коммуникативные способности оставались за рамками проверок. Несмотря на недостатки, в целом программированный контроль был определенным шагом вперед по пути стандартизации требований к результатам учебного процесса. Тем не менее к концу 80-х гг. XX в. он сошел на нет, что было связано с появлением во многих вузах

Государственный университет имени К.Д. Ушинского

нашей страны первых персональных компьютеров (ПК) и неофициальным снятием запрета на тесты.

Современные тенденции контроля и оценки знаний. Начало XXI в. совпало с экспериментом по введению Единого государственного экзамена в нескольких регионах России (2001), вызвавшего острые дискуссии по поводу тестов среди педагогов и ученых. В качестве отклика на этот эксперимент в школах и вузах в широких масштабах стали разрабатываться и применяться педагогические тесты. Распространение тестов в России совпало по времени с периодом интенсивного внедрения в учебный процесс ПК, открывающих новые возможности для контроля, самоконтроля и самооценки на основе программно-инструментальных средств и контрольно-обучающих программ.

В целом современный педагогический контроль носит эклектический характер и характеризуется совмещением привычных оценочных средств с новыми, использующими мультимедийные и Интернет-технологии без анализа многих дидактических, технологических и психологических проблем.

2.2. Традиционные средства контроля, оценки и отметки

Традиционные формы и средства контроля. В школе к традиционным средствам контроля относятся письменные или устные поурочные опросы, домашние задания и экзамены. *Устные поурочные опросы* обычно применяются в текущем контроле. Они предполагают получение ответов учащихся на вопросы учителя и обладают достоинствами, поскольку легки в организационном плане, обеспечивают оперативную обратную связь в процессе коррекции усвоения знаний учащимися, стимулируют обсуждения в классе и развивают коммуникативные компетенции. Недостатком устных опросов является фрагментарность охвата учащихся, поскольку за урок учитель может опросить не более 4—5 человек. К *письменным поурочным опросам* относятся контрольные работы, которые подводят итоги определенного периода обучения.

Особой формой контроля является *домашняя работа*, обсуждение результатов которой в классе оказывает обучающее воздействие, особенно в тех случаях, когда задания допускают нестандартные решения. В итоговом контроле обычно используют *устные* или *письменные экзамены*, как правило, вызывающие значительные эмоциональные и физические перегрузки у школьников, привыкших добросовестно учиться.

Достоинства и недостатки традиционных контрольно-оценочных средств. Разработка традиционных контрольно-оценочных средств обычно не вызывает затруднений у педагогов, поскольку она опирается на обширную методическую базу и легко осуществима.

Полесский государственный университет имени А.Н. Сахарова

К тому же необходимую подготовку к использованию привычных опросов и экзаменов учителя получают из собственного опыта школьных лет. Традиционный контроль не требует предварительных финансовых вложений, для его проведения не нужны дорогостоящие компьютеры, программное обеспечение и тесты.

Недостатки традиционных контрольно-оценочных средств значительно перевешивают достоинства. К этим недостаткам относятся отсутствие связи традиционных средств контроля с современными технологиями обучения, обеспечивающими развитие вариативности и доступности для учащихся образовательных программ, низкая эффективность в условиях массового обучения, субъективизм и несопоставимость результатов контроля. Несмотря на эти недостатки, многие учителя, даже те, кто привык добросовестно трудиться, ратуют за использование традиционных контрольно-оценочных средств. Говорят, что преподаватель на экзамене сам выставляет себе отметки, но мало кто способен беспристрастно оценить собственную работу. Поэтому контроль — достаточно консервативная сфера практической педагогики, хотя многие учителя в своих выступлениях обращают внимание на то, что в существующую традиционную систему контроля необходимо внести изменения.

Оценки и отметки. Проверочная деятельность учителя завершается выставлением оценок. По сложившейся традиции в учебном процессе слово «оценка» означает некий результат. В более широком значении под этим словом понимается не только конечный результат, но и процесс формирования оценки. Чтобы избежать путаницы, в контексте данного пособия в последнем случае используется термин «оценивание».

Оценивание является необходимым компонентом процесса контроля, результаты которого имеют большое значение для учащихся и их родителей, поскольку школьные оценки влияют в той или иной степени на будущее ребенка и вносят элемент соревнования в отношения учащихся. Казалось бы, такие аргументы должны вызывать у педагога стремление к максимальной объективности и беспристрастности. Однако зачастую этого не происходит, например, в тех случаях, когда оценки ставятся в спешке или зависят от личностных отношений учителя и ученика, посещаемости занятий, поведения учащегося на уроках и т.д.

Для придания оценки максимальной объективности и адекватности поставленной цели контроля необходимо сосредоточиться на предмете оценивания и минимизировать влияние других факторов, смещающих оценочные суждения. Конечно, в реальности на каждую оценку, выставленную традиционным путем, оказывают влияние различные факторы, поэтому такие оценки нельзя использовать для сравнения результатов работы учителей, интерпретировать их в управлении качеством образования.

Гомельский государственный университет имени Ф. Скорины

Педагогические оценки нередко ошибочно отождествляют с отметками. Следует помнить, что оценка выражает результат, а отметка служит для установления численных аналогов оценочных суждений. Например, по установившейся в нашей школе пятибалльной шкале отметок удовлетворительные знания оцениваются «тройкой», отличные — «пятеркой». На самом деле эти баллы не имеют четкого педагогического смысла и не дают количественной характеристики ответа учащегося. Во многих странах вместо численных аналогов используются буквенные символы (*A*, *B*, *C* и т.д.), с помощью которых устанавливается место результата каждого учащегося в группе контролируемых учеников.

Характеристика процесса оценивания. Процесс оценивания основан на сравнении, которое может иметь различный характер в зависимости от того, что выбрано в качестве базовой системы при выставлении оценок. Такой системой могут быть:

1) результаты других учащихся;
2) требования программы или ГОС;
3) априорные оценки способностей учащегося;
4) объем затраченного учащимся труда и его прилежание в освоении учебного материала.

В первом случае при выставлении оценки проводится сравнение подготовленности каждого учащегося с результатами всего класса или определенной группы учеников, после чего учащиеся ранжируются на группы, внутри которых все имеют одинаковые оценки. Обычно в классе учитель руководствуется именно такой логикой. Например, если во время устного опроса большинство учеников дает слабые ответы, на «тройку», более сильный ответ учащегося на фоне предыдущих всегда в глазах учителя заслуживает «четверки» или «пятерки».

Во втором случае, при сравнении подготовленности учащегося с установленными требованиями к учебным достижениям, результаты остальных учеников не играют никакой роли, а оценка выставляется в зависимости от процентного соотношения выполненных требований и полного объема требований, планируемых к усвоению. Полученный для каждого учащегося процент сравнивается с критериями, установленными экспертным или эмпирическим путем. По результатам сравнения в зависимости от полученного процента выставляются оценки. Хотя на словах такой процесс кажется достаточно простым и объективным, он трудноосуществим на практике, поскольку разработать эталонные наборы требований для всех школ и каждого урока нереально.

В третьем случае достижения учащегося сопоставляются с его потенциальными возможностями, интуитивно оцененными учителем. Те ученики, чьи способности, по мнению педагога, вы-

Мозырский государственный педагогический университет имени И.П. Шамякина

соки, а достижения ниже возможностей, получают низкие оценки. Ученики с низким потенциалом, демонстрирующие в процессе контроля такие же достижения, что и более способные, получают более высокие оценки. Такой подход кажется многим педагогам очень привлекательным, поскольку, по их мнению, мотивирует учеников к повышению уровня учебных достижений. На самом деле он несправедлив, субъективен и служит обычно причиной конфронтаций в классе.

В четвертом случае в качестве основы для сравнения вместо способностей выбираются усилия, затраченные учащимися на приобретение новых знаний, интенсивность учебной деятельности и прилежание. По сравнению с предыдущим такой подход еще более несправедлив, так как направлен против ярких одаренных детей и снижает мотивацию самых способных учащихся к получению высоких оценок. Ученикам, которые склонны к упорному труду, учителя обычно завышают оценки, руководствуясь простой логикой — чем больше затраченные усилия, тем выше оценка. Тем же, кто легко усваивает материал, оценки занижаются, тогда как другие вознаграждаются за то, что истратили больше времени на усвоение того же или меньшего объема учебного материала.

Современные тенденции в оценочных процессах. Отсутствие в нашей стране стандартизированных тестов, фиксирующих на многие годы в единых шкалах требования к подготовленности учащихся и задающих некоторые нормы оценок, привело к девальвации существующей пятибалльной шкалы [61]. Согласно данным исследования, проведенного А. Г. Шмелевым путем опроса сотен респондентов, в сознании педагогов чаще присутствует идеализированная шкала, в которой отметки выставляются при сопоставлении планируемого и достигаемого уровня усвоения учебного материала. Применимая в школах шкала выглядит намного нейтральнее и дает возможность учителям, за небольшим исключением (в сильных профильных классах), ставить минимальное количество «двоек».

Введение в 2001 г. ЕГЭ опровергло несколько идеализированные представления о качестве российского образования. Объективные данные ЕГЭ по большинству регионов России показали, что существующая граница между «двойкой» и «тройкой» намного ниже, чем субъективные представления о ней, поскольку вместо абсолютной успеваемости во многих школах появилось до 20 % двоечников. В целом опыт ЕГЭ можно оценить как позитивный. Совмещение субъективных оценочных суждений учителей и объективных данных тестирования со временем неизбежно приведет к выставлению более обоснованных отметок в школах и будет стимулировать учащихся к повышению уровня учебных достижений.

Электронный архив библиотеки МГПУ имени А. А. Кулешова

2.3. Контроль и оценка в современном образовании, основные инновационные тенденции

Условия обновления контрольно-оценочной системы в школьном образовании. Инновационные тенденции, характерные для современного образования, затрагивают не только процесс образования, но и контрольно-оценочную систему, выдвигая повышенные требования к ее эффективности. Для обновления контрольно-оценочной системы необходимо:

– минимизировать субъективизм в итоговом контроле и перейти к расширенному использованию стандартизированных тестов;

– снизить долю авторитарности и принуждения в текущем контроле, создать условия для самоконтроля и самооценки учащихся;

– отказаться от преимущественной ориентации текущего и итогового контроля на оценку результатов заучивания, деятельности по образцу, алгоритмических знаний и перейти к инновационным измерителям, обеспечивающим оценку компетентностей, способностей к творческой и практической деятельности;

– заменить привычную ориентацию на «среднего ученика» индивидуализированными методами коррекции учебной деятельности в процессе текущего контроля, систематически использовать входной контроль;

– снизить долю традиционных письменных проверок за счет введения аутентичных форм оценивания, предполагающих использование в контроле релевантных, значимых для учащихся, оценочных средств: тестов практических умений, ситуационных заданий и портфолио.

Основные инновационные тенденции в контроле. В последнее десятилетие наблюдается усиление связи между контролем и обучением. Целевые установки, определяющие результаты образования, задаются в терминах измеряемых результатов. В свою очередь процесс обучения строится так, чтобы активизировать обучающие и развивающие функции контроля за счет оптимизации содержания и трудности учебных задач, подбираемых для текущего контроля в индивидуальном режиме. Контроль приобретает все большее значение, он меняет свой характер и объединяет традиционные функции по проверке и оценке результатов обучения с функциями управления качеством всего учебного процесса.

В системе оценивания результатов обучения происходят значительные изменения, которые характеризуются переходом от бихевиористской точки зрения к когнитивной и проявляются в смещении акцентов с преимущественной оценки результатов обучения на компоненты процесса получения результата, с пассивного ответа на заданный вопрос на активное конструирование содержания ответа, с оценки отдельных, изолированных умений на интегрированную и междисциплинарную оценку. В контроле зна-

Электронный архив библиотеки МГУ имени А.А. Кулешова

чительно усилилось внимание к метапознанию, предполагающему формирование межпредметных знаний, умений переноса знаний из одного предмета в другой и общеучебных умений. При оценивании результатов обучения изменился контекст расшифровки понятий «знающий» и «умеющий». Вместо прежнего приоритета фактологии и алгоритмических умений на первое место вышли умения применять знания в нестандартных или практических ситуациях.

В современном контроле измерения стали органической частью образовательного процесса, важнейшим средством получения информации, широко используемой в управлении качеством образования. На фоне постоянно растущей роли тестов пришло осознание ограниченности количественных методов, благодаря чему в педагогическом контроле стала развиваться так называемая смешанная методология, строящаяся на сочетании количественных и качественных оценок. Соответственно появилось новое поколение измерителей, обеспечивающих вместе с традиционными средствами контроля и тестами многомерные аутентичные (комплексные, многогранные) оценки, охватывающие результаты учебной деятельности и в школе, и во внеучебное время. Приоритет статических оценок, фиксирующих уровень подготовленности обучаемых в момент контроля, сменился в последнее время преобладанием динамического анализа изменений качества подготовленности учащихся, основанного на повсеместно разрабатываемых и внедряемых системах мониторинга качества образования [68; 77].

Портфолио и тесты для оценки практической деятельности учащихся. В современном контроле появились новые виды измерителей, выявляющих позитивную динамику изменений подготовленности, активность учащихся в усвоении новых знаний, рост их компетентности, а также степень освоения коммуникативных и интеллектуальных умений. В первую очередь к числу таких измерителей следует отнести *портфолио* (рабочие папки), содержащие целевые подборки работ учащегося по одной или нескольким учебным дисциплинам и составленные учителем в сотрудничестве с учащимся. Участие ученика в отборе работ является важным фактором положительной мотивации учебной деятельности, стимулирующим стремление к самооценке своих достижений. Поэтому многие преподаватели видят в портфолио эффективное средство развития у школьников навыков критического мышления и получения реальных самооценок. Несмотря на индивидуализированный подход при выборе заданий, результаты выполнения которых нуждаются в основном в экспертных оценках, портфолио обеспечивают достаточно объективную информацию о качестве учебных достижений. Это связано с тем, что процесс их проверки предельно стандартизируется, четко определяются критерии оценки

Учебно-методический центр Федеральный институт развития образования

достижений, вырабатываемые в сотрудничестве с учащимися, тщательно обеспечиваются свидетельства самостоятельной работы учеников.

На сегодняшний день в сфере образования существует определенная типология портфолио, представленная в отечественных и зарубежных работах [68]. Первый вид — *рабочее портфолио* — включает работы учащегося за определенный период времени, которые показывают произошедшие изменения в его знаниях. Второй вид — *протокольное портфолио* — в документальной форме отражает все виды учебной деятельности и подтверждает самостоятельность работы учащегося. В этот вид портфолио могут включаться черновики готовых работ ученика. Третий вид — *процессное портфолио* — предназначено для демонстрации достижений учащегося на различных этапах процесса обучения. Четвертый вид — *итоговое портфолио* — обычно используется для получения суммарной оценки знаний и умений учащегося, усвоенных по основным предметам учебной программы. В последнем случае в портфолио обычно включается самая лучшая завершенная работа ученика, выбранная им совместно с учителем. Формы представления материалов портфолио могут быть различными. Нередко используются аудиовизуальные средства, такие как фотографии, видеозаписи, электронные версии работ учащегося.

Сторонники портфолио обычно относят их к средствам аутентичного оценивания и в качестве позитивной аргументации приводят их высокую валидность, адекватность современным требованиям к качеству образования. Однако портфолио, как и тесты, не решают всех проблем оценивания качества образования, поскольку имеют недостатки. Они дороги, требуют больше времени при использовании по сравнению с тестами и вызывают сомнения в надежности.

К числу новых форм измерителей относятся *тесты*, которые разрабатываются для оценки практической деятельности учащихся (Performance assessment) [25; 67]. Такие тесты позволяют выявить уровень освоения практических навыков с помощью экспериментальных заданий деятельностного характера, в результате выполнения которых получается некоторый материальный продукт, оцениваемый экспертами в стандартизированной шкале баллов. Многие из тестов практических умений не отвечают по своим характеристикам требованиям теории педагогических измерений. Тем не менее они имеют высокую валидность и вызывают большой интерес у учеников. Экспериментальные задания обычно применяются в процессе текущего контроля, но не влияют на принятие административных решений в образовании, поэтому низкая точность оценок не является проблемой. В случае неудачи ученик может заново выполнить тесты и добиться успеха.

Полесский государственный университет
Восточный
им. М.А. Маркова

Автоматизированный контроль. В последнее десятилетие интенсивно развиваются новые компьютерные технологии, позволяющие автоматизировать процесс текущего и итогового контроля на основе использования программно-инструментальных средств [2; 31; 39; 43]. Нередко контролирующие программы совмещают с обучающими программами, при этом используют диалог учителя с обучаемым для проверки или коррекции учебной деятельности с помощью дополнительной информации, восполняющей обнаруженные пробелы в знаниях учащихся. Современные инструментальные системы контроля и оценки знаний имеют, как правило, дружественный интерфейс, поддерживают различные формы заданий и позволяют реализовывать сценарии проведения контроля, используют работу с текстом, неподвижными и анимированными изображениями, звуком, видео и т.д.

Отдавая предпочтение тем или иным инновациям, нужно всегда стремиться к многогранной оценке качества результатов обучения и пониманию целесообразности использования новшеств в учебном процессе. Например, информация, полученная о подготовленности учащегося с помощью средств автоматизированного контроля, должна обязательно подкрепляться дополнительными данными об особенностях его памяти, воображения, мышления и речи. Следует принимать во внимание уровень подготовленности ученика к работе на компьютере, его коммуникативные способности (умение вести диалог, дискуссию, вербально выражать свои взгляды и мысли, общаться и сотрудничать со своим сверстниками и учителями и т.д.).

Северо-Осетинский государственный университет им. М.К. Аммосова

2.4. Контрольно-оценочная система в школе

Понятие о системе, ее свойства и задачи. Общий замысел современной контрольно-оценочной системы состоит в создании совокупности методик, процедур, измерителей, программно-педагогических средств, взаимодействующих как единое целое в процессе проверки результатов обучения, оценивания состояния объектов контроля, анализа данных контроля, их интерпретации и выработки корректирующих воздействий в целях повышения качества обучения. В том или ином виде контрольно-оценочная система существует в любой школе. Однако ее современный вариант предполагает опору на базы данных школьного мониторинга, наличие специальных методик проведения контроля и оценивания результатов учащихся, банка инновационных измерителей, стандартизированных шкал, программно-инструментальных средств, а также подготовки коллектива учителей по педагогическим измерениям и использованию данных контроля в управлении качеством образования.

Свойства и задачи системы. Современная контрольно-оценочная система должна обладать целостным функционально-структурным строением, сочетающим традиционные и инновационные методы контроля.

Создание такой системы в школе предполагает налаживание и поддержку всех необходимых информационных потоков для управления качеством обучения, охват пользователей с различными уровнями доступа, среди которых — учащиеся, их родители, учителя и администрация школы.

К основным задачам контрольно-оценочной системы можно отнести:

– получение объективной информации об уровне и качестве индивидуальных учебных достижений учащихся в целях коррекции учебного процесса;

– получение объективной текущей и прогностической информации о качестве обучения для муниципальных и других органов управления образованием;

– обеспечение возможности индивидуализации учебного процесса на основе результатов контроля, реализации личностно ориентированной, развивающей и других инновационных технологий обучения без необоснованного роста трудозатрат со стороны педагога;

– сбор и анализ объективной информации о подготовленности обучающихся для выставления итоговых оценок при переходе на следующую ступень обучения;

– поддержку развития новых форм, методов и средств контроля, адекватных компетентностному подходу, аутентичной, сбалансированной и интегральной оценке учебных достижений учащихся;

– обеспечение возможности самоконтроля, самокоррекции и самооценки для учащихся;

– создание и поддержку функционирования школьной системы мониторинга качества образования.

Программное обеспечение и банк тестовых заданий. По своему функциональному назначению и особенностям конструирования программные средства внутришкольной системы можно разделить на инструментальные оболочки для проведения и оценки результатов компьютерного тестирования, обучающие компьютерные программы и специальные пакеты для экспертизы качества тестов, их генерации и статистического обоснования их качества.

Особое значение при создании внутришкольной системы контроля имеет компьютерный банк тестовых заданий. Он включает в себя калиброванные тестовые задания и стандартизированные по форме представления, обладающие устойчивыми статистическими характеристиками.

Северо-Восточный федеральный университет им. М.К. Аммосова

2.5. Эвалюация в образовании

Что такое эвалюация? Термин «эвалюация» (*evaluation*) не распространен в отечественной педагогической науке, хотя он широко применяется в большинстве зарубежных стран, имеющих высокоразвитые системы управления качеством образования. Впервые он был использован в России в книге «Педагогическое образование в университете: контекстно-биографический подход», изданной в 2001 г. и показывающей эволюционный путь становления этого понятия в образовательных системах ряда стран.

Современная трактовка эвалюации. Согласно современной трактовке, термин «эволюция» не связывается лишь с процессом получения оценок, а рассматривается гораздо шире — как интегративная категория оценочно-аналитической деятельности в управлении качеством образования. Эвалюация в образовании охватывает спектр теоретико-методологических и практических работ по систематическому исследованию ценности и позитивности качеств обучаемых, анализируемых на основе единой методологии, сочетания количественных и качественных методов для отслеживания характера и динамики изменений. Таким образом, эвалюация в образовании включает все направления оценочно-аналитической деятельности, среди которых:

– постановка целей, определение методологического подхода (как правило, с опорой на динамические методы анализа изменений характеристик обучаемых);

– разработка логических или математических моделей;

– выбор методов сбора и анализа информации (с преимущественной опорой на теорию педагогических измерений и статистические методы, которые не исключают широкого применения качественных экспертных оценок);

– разработка дизайна исследования;

– определение методов обработки и интерпретации данных в процессе анализа для принятия управленческих решений в целях повышения качества образования.

Так как эвалюация находится на стыке педагогики, психологии, экономики, менеджмента и других наук, то ее теория носит междисциплинарный характер. Применение количественных методов в эвалюации предполагает использование математико-статистического аппарата, метаанализа, дисперсионного анализа, факторного анализа и т.д. Теория управления качеством образования привносит в эвалюацию синергетический и процессный подходы к анализу объектов анализа в эвалюации. При использовании качественных методов сбора информации привлекаются эксперты, применяются анкеты и интервью. Оценивание эффективности деятельности образовательных учреждений обычно включает анализ соотношения финансовых затрат на образование и их

Электронный архив библиотеки МГУ имени А.А. Кулешова

отдачи, поэтому при рассмотрении эффективности вложений используются различные стоимостные модели и количественные методы анализа социально-экономических объектов. Связь эвалюации с менеджментом качества образования неизбежно влечет за собой применение стандартов качества ИСО 9000 и EQUIS.

Становление и развитие эвалюации в образовании в наши дни связано с глобальными переменами, произошедшими в XXI в. Сегодня функции, методы, процедуры и технологии менеджмента широко внедряются в различные направления профессиональной деятельности, в том числе и в образование. Во всем мире в сфере образования утверждается парадигма теории управления, включающая системный подход, приоретет стратегического управления развитием социально-экономических систем, мониторинг и бенчмаркинг, что позволяет планировать эффекты нововведений, управлять качеством образования, оценивать риски принятия управленческих решений, эффективно распределять человеческие ресурсы и внедрять новые информационные технологии.

Эти тенденции проявляются во многих странах, обладающих высокоразвитыми системами образования. В российских вузах повсеместно создаются системы менеджмента качества образования, базирующиеся на эвалюации и охватывающие все аспекты образования. В США и Канаде организованы ассоциации оценщиков — American Evaluation Association (AEA) и Canadian Evaluation Society (CES), объединяющие педагогов, управленцев и других работников социальной сферы из разных стран, проводятся ежегодные конференции, издается научная литература, энциклопедии и журналы по проблемам эвалюации в образовании. В целом эвалюация является необходимым слагаемым современных образовательных систем, без нее не могут быть решены многие современные проблемы образования.

Федеральный государственный университет им. М.А. Шолохова

Практические задания и вопросы для обсуждения

1. Охарактеризуйте достоинства и недостатки традиционных средств контроля.

2. Чем отличается автоматизированный тестовый контроль от программированного контроля?

3. Представьте, какой будет средняя школа через 10 лет. Как вы считаете, в каких направлениях будут развиваться оценочные средства? Какие инновации, по вашему мнению, приживутся в образовании?

4. Оцените достоинства и недостатки аутентичного оценивания. Стоит ли учителю наблюдать детей во внеучебное время и интересоваться их достижениями вне школы?

5. Вы замечаете, что одни из учеников из класса систематически не успевает. Вы склонны отнести его к категории детей с педагогической запущенностью и просить директора школы о переводе этого ученика в

специальный класс. Однако директор замечает, что сначала нужно собрать дополнительные данные, свидетельствующие о необходимости такого перевода, а уже потом ставить этот вопрос перед администрацией школы. Какие данные, по вашему мнению, следует собирать и какие оценочные средства будут наиболее эффективными при решении этого вопроса?

6. Между родителями ученика I класса и учителем состоялся следующий диалог.

Р о д и т е л и. Наш сын Миша очень хорошо читает. Обычно у него по чтению одни «пятерки». Почему в этой четверти вы поставили ему «четыре»? Он стал хуже читать?

У ч и т е л ь. Нет, что вы, ваш сын читает лучше всех в классе. Я нередко прошу его прочесть отрывок из книги при объяснении нового материала в классе.

Р о д и т е л и. Тогда почему у него «четверка», а у его соседки по парте Маши «пятерка», хотя Миша говорит, что она читает по складам?

У ч и т е л ь. Миша часто смотрит в окно во время объяснения нового материала или мешает Маше, когда читают другие, более слабые ученики. А Маша всегда прилежно работает в классе и дома. Ее родители говорят, что она тратит весь день на подготовку уроков. Поэтому в этой четверти я поставил ей «пять».

Проанализируйте ситуацию с точки зрения Миши и его родителей. Должны ли родители согласиться с таким решением учителя? Какие оценочные системы использует педагог? Прав ли он?

Северо-Восточный федеральный университет им. М.К.Аммосова

ГЛАВА

3

РАЗВИТИЕ ПЕДАГОГИЧЕСКОГО ТЕСТИРОВАНИЯ В РОССИИ И ЗА РУБЕЖОМ

3.1. Исторические предпосылки современного тестирования в отечественном образовании

Тесты в России в конце XIX первой половины XX в. История тестов, пришедших на смену донаучным формам и обыденным представлениям педагогов о проверочных средствах, началась в России, как и в других странах, в конце XIX — начала XX в. В этот период тесты широко применялись сторонниками экспериментальной педагогики, которые всячески пропагандировали их наравне с другими психометрическими методиками в педологии. Видные специалисты того времени (А. П. Болтунов, А. Ф. Лазурский, А. П. Нечаев, Ф. Е. Рыбаков и др.) в своих исследованиях личности обучаемого широко использовали тесты [37].

На волне стремления к прогрессу в России в первые годы советской власти был востребован потенциал, накопленный российской наукой в образовании. У истоков советской педологии, в рамках которой развивались тестовые методики, стояли известные ученые того времени (М. Я. Басо, Л. С. Выготский, А. П. Болтунов, П. П. Блонский, К. Н. Корнилов, А. П. Нечаев, А. Р. Лурия и др.), посвятившие ряд своих работ оценке результатов обучения и развития детей. Однако период активной деятельности педологов и тестологов в советской России оказался коротким. В 1936 г. вышло известное постановление ВКП(б) «О педологических извращениях в системе Наркомпросов» [40], в котором тесты были объявлены вредным методом. Многие годы, вплоть до начала 90-х гг. XX в., отношение к тестам было негативным. Педагоги-практики вернулись к тестам в период прекращения массовых политических репрессий. Уже к началу 60-х гг. XX в. многие учителя, заботящиеся не о формальных показателях, а о реальном качестве учебного процесса, стали использовать при контроле наборы заданий в тестовой форме, которые чаще всего назывались по-разному, но только не словом «тест».

Промежуток времени с 60-х гг. до начала 90-х гг. XX в. был для отечественной школы периодом постепенной либерализации, когда учителя изыскивали различные резервы в борьбе за повышение качества обучения на фоне снижения авторитарного характера учебного процесса. Тесты в это время существовали полулегально,

официально запрет на них не был отменен, но уже предпринимались попытки диссертационных исследований по проблемам тестирования, появлялись работы педагогов-новаторов, в которых тесты и анкетные опросы использовались как инструментарий для подтверждения эффективности нововведений. В это время тесты разрабатывались без должного знания теории педагогических измерений, на основе опыта и здравого смысла, поэтому чаще всего тестами их было можно назвать лишь условно.

Столь же необоснованными по надежности и валидности были наборы заданий с выбором ответов, создаваемые практически повсеместно на рубеже 60—70-х гг. XX в. в связи с интенсивным развитием программированного обучения. Только к середине 80-х гг. XX в. в нашей стране в научных и методических работах преподавателей стали утверждаться основные положения теории педагогических измерений, декларирующие необходимость эмпирической верификации качества измерений и статистического анализа характеристик тестов [42].

В целом рассматриваемый временной промежуток в развитии тестов был периодом прогресса, когда на смену представлениям о тесте как о простом наборе заданий пришло научное понимание этого термина. В этот период появляются научные издания по тестовой и смежной с ней проблематике, учебные пособия и многочисленные статьи отечественных авторов, как поддерживающие, так и осуждающие тесты в образовании.

Развитие тестов в конце XX в. и в наши дни. Новая история тестов в России началась в 90-е гг. XX в., когда стали больше говорить не о руководстве, а о научно обоснованном управлении учебным процессом, в информационном обеспечении которого важная роль по праву принадлежит тестам [3]. Новое понимание возможностей тестов в образовании способствовало росту научных исследований. В 90-е гг. XX в. появляются работы по проблемам измерений и тестирования в образовании, защищаются многочисленные кандидатские и докторские диссертации, издаются монографии, учебные пособия, журналы, проводятся конференции и симпозиумы.

Последнее десятилетие XX в. в нашей стране совпало с периодом бурного развития структур, занимающихся практической работой по созданию и применению тестов. К основным событиям этого периода можно отнести открытие в 1990 г. первой в стране кафедры педагогических измерений в Исследовательском центре проблем качества подготовки специалистов (директор Н. А. Селезнева) Московского института стали и сплавов; создание при Московском государственном университете в этом же году Центра тестирования «Гуманитарные технологии» (проект по компьютерному аттестационному тестированию для старшеклассников «Телетестинг» — научный руководитель А. Г. Шмелев); привлече-

ние в 1991 г. Центра качества образования Института общего среднего образования РАО к участию в сравнительных международных исследованиях по оценке учебных достижений (директор Центра — Г. С. Ковалева); открытие в 1995 г. Центра тестирования выпускников общеобразовательных учреждений, преобразованного впоследствии в Федеральный центр тестирования со статусом государственного учреждения Министерства образования и науки Российской Федерации; основание в 1998 г. государственной системы тестирования иностранных граждан по русскому языку как иностранному при поддержке ведущих вузов Москвы, Санкт-Петербурга и других городов России. Перечень этот можно было бы смело продолжить. Начиная с 90-х гг. XX в. практически во всех регионах России создаются центры тестирования и аттестации учащихся, методические лаборатории по диагностике и центры качества образования.

Значимым событием в области подготовки кадров по педагогическим измерениям стало открытие в 2001 г. на факультете повышения квалификации Российского университета дружбы народов (декан Т. М. Балыхина) кафедры тестологии (заведующая кафедрой М.Б. Челышкова). На этой кафедре впервые в нашей стране началась реализация профессиональной образовательной программы, рассчитанной на 1480 часов, для получения дополнительной квалификации «Тестолог (специалист в области педагогических измерений)».

Для повышения качества тестовых материалов, разрабатываемых в России, в 2000 г. при Исследовательском центре проблем качества подготовки специалистов Московского института стали и сплавов (технологического университета) был открыт Центр сертификации педагогических тестовых материалов (ПТМ) (директор В. И. Звонников) и создан Координационный совет Минобразования России (в настоящий момент — Минобрнауки) по вопросам сертификации качества педагогических тестовых материалов.

Коренные изменения в отношении учителей к тестам произошли в 2001 г. в связи с началом эксперимента по введению единого государственного экзамена, благодаря которому тесты получили официальное признание в России. За годы эксперимента значительно повысилась степень доверия к результатам тестирования со стороны органов управления образованием, образовательных учреждений, самих учителей, которые убедились в высокой объективности и обоснованности баллов ЕГЭ, в их высокой прогностичности при отборе абитуриентов вузов. В 2002 г. был открыт Федеральный институт педагогических измерений (директор А. Г. Ершов), приоритетным направлением деятельности которого является научное, методическое и организационное сопровождение процесса создания контрольных измерительных материалов (КИМ) для ЕГЭ.

Северо-Восточный федеральный университет имени М.К. Аммосова

В целом, подводя итоги развития тестирования в России, можно сказать, что на сегодняшний день среди ученых-педагогов в нашей стране наконец появилось понимание того, что теория педагогических измерений — наука, обладающая своей методологией, методами и аппаратом, необходимым для разработки качественных педагогических тестов. Сегодня большими тиражами издаются инновационные работы по педагогическим тестам отечественных ученых и зарубежных авторов, а также сборники материалов КИМ ЕГЭ, специальные журналы по тестовой проблематике. Таким образом, можно считать, что на данный момент в России сформировалось сообщество профессионалов — специалистов по разработке и применению тестов.

3.2. Развитие тестирования в зарубежных странах

Становление тестов в психологии, образовании и армии. Появление педагогических тестов за рубежом нередко связывают с именем французского врача и психолога А. Бине (*A. Binet*). Его работа по диагностике интеллектуальных способностей, вышедшая в 1905 г., считается точкой отсчета в становлении основных научных подходов к измерениям в психологии и образовании [69]. Конечно, были и более ранние попытки создания подобных тестов. Британские исследователи Ф. Гальтон (*F. Galton*) и Дж. Кэттелл (*J. Cattell*) в 1890 г. использовали термин «тест на интеллект» (mental test) [1]. Им же принадлежит заслуга введения в научный оборот специальной характеристики качества теста (power of discrimination), указывающей на его способность дифференцировать испытуемых по измеряемой переменной.

Немного позднее идеи А. Бине были использованы немецким психологом и философом В. Штерном (*W. Stern*), который предложил специальный коэффициент для оценивания интеллекта — коэффициент IQ. Этот термин используется до сих пор, правда, несколько в ином контексте. В США работы А. Бине по измерениям продолжил Л. Терман (*L. Terman*). Он создал новые тесты (Стэнфорда-Бине), представляющие собой оригинальную модификацию работы его предшественника [1].

Достижения второго десятилетия XX в. в сфере измерений были связаны с проблемой распределения большого количества призывников по различным родам войск в США во время Первой мировой войны. Правительство Соединенных Штатов организовало «мозговой центр» из специалистов — составителей тестов, статистиков и экспертов по измерениям — и направило его работу на решение задач, продиктованных военной необходимостью. В этом центре свои усилия по отбору и распределению новобранцев объединили такие известные специалисты, как Э. Торн-

Полесский государственный университет
Электронный архив

дайк (*E. Thorndike*), Р. Годдард (*R. Goddard*), П. Йеркс (*P. Yerkes*) и Л. Терман (*L. Terman*). Результатом их работы стали знаменитые батареи Альфа и Бета тестов для армии, которые после стали использовать (вплоть до настоящего времени) в качестве примеров при обучении разработчиков тестов в образовании.

Первые стандартизированные измерители в образовании США были созданы для оценки качества почерка, правописания и выполнения арифметических действий в конце 20-х гг. XX в. Тогда же в Америке появилась батарея тестов SAT, разработанных по инициативе Совета колледжей и предназначенных для отбора абитуриентов [1]. Несколько позже, в конце 40-х гг. XX в., были созданы батареи тестов достижений, реализующие идеи многомерных измерений и обеспечивающие сопоставимость результатов по разным школьным предметам.

Развитие классической (традиционной) теории педагогических измерений и тестирования в XX в. В 1904 г. англичанин Ч. Спирмен (*C. Spearman*) опубликовал фундаментальный теоретический труд по исследованию общих интеллектуальных способностей. Используя школьные оценки по различным предметам, Ч. Спирмен применил к данным тестирования аппарат новой для того времени теории корреляции и выявил примеры ковариации оценок школьников. Тем самым он заложил основы научных подходов к обоснованию качества тестов, соединив в своем исследовании теорию физических измерений, корреляционные методы и накопленный его предшественниками-психологами опыт оценивания способностей детей.

Прообразом научных положений теории педагогических измерений послужила далекая от образования работа Н. Р. Кэмпбелла «Основы физики» (1920), благодаря которой был разработан теоретический аппарат для анализа качества измерений, а обыденное представление о тесте и его научное определение стали заметно различаться. Для обоснования качества педагогических измерений, их надежности и валидности была создана классическая теория тестов, получившая впоследствии название традиционной, и базирующаяся на концепции параллельных измерений и теории корреляции. На основе этой теории в 30—40-е гг. XX в. интенсивно разрабатывались количественные методы для анализа качества тестовых заданий, строились стандартизованные тесты учебных достижений и осваивались методы шкалирования результатов выполнения тестов.

В конце 40-х гг. XX в. увлечение чистой теорией сменилось пониманием важности правильного применения измерений в сфере психологии и образования. Исследователи осознали тот факт, что к тестируемым нужно относиться бережно и осторожно и что любая ошибка в оценках может привести к необратимым последствиям. В этой связи в теории педагогических измерений активи-

Электронный архив библиотеки МГУ имени А.А. Кулешова

зировался поиск эффективных методов повышения надежности результатов тестирования, стали разрабатываться методы факторного, дисперсионного и корреляционного анализа данных, получили широкое развитие методы формирования репрезентативных выборок, необходимых для стандартизации тестов.

Значимым событием конца 40-х гг. XX в. для развития теории и практики измерений в образовании стало создание в 1947 г. в США Службы образовательного тестирования — Educational Testing Service (ETS). Сегодня эта организация имеет представительства практически во всех странах мира. Из числа тестов, созданных ETS, наибольшую известность в связи с расширением программ международного обмена в обучении приобрел TOEFL, выявляющий уровни владения различными видами речевой деятельности для обучения в англоязычных странах.

Создание современной теории тестов. В истории тестов были этапы подъемов и спадов. В частности период ожесточенной критики педагогических тестов наблюдался в США в конце 60-х гг. XX в. Широкая публичная дискуссия по проблемам использования тестирования совпала по времени с научной критикой, вызванной недостатками классической теории тестов. Высказывались опасения в том, что тестирование служит инструментом подавления инициативы и творчества учащихся, слишком упрощенно трактует такие сложные конструкты, какими являются учебные умения, не объективно оценивает учебные достижения. Приводились и другие аргументы против широкого использования тестов. Ученые-тестологи занимались в основном конструктивной научной критикой традиционных методов разработки тестов, что привело к построению новой теории — Item Response Theory (IRT), которую нередко в наши дни называют современной теорией тестов.

Неоспоримые преимущества IRT, связанные с возможностью прогноза надежности измерений, увеличения эффективности тестирования и получением оценки параметров подготовленности учащихся, не зависящих от трудности заданий теста, стали широко применяться на практике с конца 80-х гг. XX в. В это время были разработаны алгоритмы оценивания параметров испытуемых и заданий на основе математических моделей IRT, создано программное обеспечение и стали широко использоваться ПК. Сочетание аппарата классической теории и IRT при разработке тестов открыло новые возможности повышения качества педагогических измерений и применения тестирования в образовании.

Современный период в истории развития тестов. Современный период в истории развития тестов характеризуется интенсивным развитием теории IRT, созданием новых моделей и методик ее применения, внедрением в тестирование компьютерных технологий, адаптивного тестирования, различных инноваций в области

Оливерова Ростовский государственный университет им. М.А. Шолохова

разработки и применения тестов. В настоящий момент наметился ряд направлений исследований, нацеленных на расширение возможностей педагогических измерений, разработку инновационных измерителей и повышение качества тестов. К их числу относится создание новых моделей педагогического измерения, инновационных форм тестовых заданий для проверки творческих и практико-ориентированных аспектов подготовленности учащихся, методов калибровки тестовых заданий и методик компьютерного моделирования тестов, обеспечивающих планируемую точность измерений.

Большое внимание уделяется психолого-педагогическим проблемам тестирования (проблемам тревожности, мотивации), развитию специальных процедур для выявления размерности пространства измерений, созданию специальных методик, позволяющих обоснованно использовать результаты педагогических измерений в управлении качеством образования.

3.3. Тестирование в психологии и в образовании

Взаимосвязь психологических и педагогических измерений. Взаимодействие психологов и педагогов в области разработки тестов имеет свои исторические причины и научные предпосылки. В научном плане оно предопределено близостью методологий измерений в эмпирических науках и единством методик разработки тестов, общностью объекта исследования, тесной взаимосвязью и частичным пересечением субъектов измерения. Исторические аспекты этого взаимодействия обусловлены особенностями становления теории педагогических измерений, сложившимися за рубежом в 20—40-е гг. XX в. В то время потребность в стандартизированных педагогических тестах постоянно увеличивалась, а сообщество тестологов в сфере образования практически еще не сложилось, поэтому психологи зачастую привлекались к работе над педагогическими тестами.

В наши дни область совместных работ педагогов и психологов расширяется. Во многих учебных заведениях в России и за рубежом работают психологи, создаются специальные службы консультирования школьников и студентов. В их функции входит не только психологическая поддержка учащихся в проблемных ситуациях, но и участие в проведении мониторинга, анкетных опросов и в подготовке учебного заведения к государственной аттестации. Педагоги, в свою очередь, проявляют интерес к оцениванию способностей обучающихся, когнитивных аспектов подготовленности, психологических особенностей учащихся в усвоении знаний и нередко выполняют свои исследования на стыке психодиагностики и измерений в образовании.

Гомельский государственный университет им. Ф. Скорины

Тенденция вхождения психологов в педагогические измерения в нашей стране наметилась в середине 90-х гг. XX в. В это время возрос интерес к педагогическим тестам, и отсутствие тестологов в образовании ощутилось особенно остро. Образовавшийся кадровый вакуум в значительной степени заполнили психологи. Попутно, сами того не желая, они привнесли в образовательную среду устойчивые представления о тестах как о наборе довольно тривиальных по содержанию заданий, носящих косвенный характер и имеющих форму с выбором ответа. Преодолению искаженного представления о тестах способствует эксперимент по введению ЕГЭ. Однако, несмотря на то что он начался еще в 2001 г., многим пользователям до сих пор неизвестны широкие возможности педагогических тестов.

Различия между психологическими и педагогическими тестами. Несмотря на имеющиеся точки соприкосновения, между педагогическими и психологическими тестами есть существенные различия. Подробно они представлены в книге А. Анастази [1]. В частности автор обращается внимание на то, что педагогические тесты оценивают результаты усвоения общего для всех обучаемых программного курса, а психологические — отражают индивидуальное восприятие жизненного опыта, различающегося по социальным и экономическим условиям его приобретения.

Результаты тестирования в психологии используются в основном для прогнозирования успешности какой-либо деятельности в зависимости от оценок личностных характеристик и склонностей или носят рекомендательный характер при выявлении личностных проблем. Педагогические тесты в итоговом контроле дают количественные оценки уровня подготовленности обучаемых, которые нередко служат основой для принятия административно-управленческих решений в образовании, поэтому к качеству результатов педагогических измерений и их сопоставимости должны предъявляться повышенные требования, обеспечивающие корректность управленческих выводов и административных решений.

Внешние различия между тестами в психологии и в образовании проявляются в формулировках заданий, которые в педагогических тестах всегда носят прямой, а в психологических тестах, чаще всего, косвенный характер. Например, в психодиагностике встречаются ситуации, когда неправильные ответы испытуемых расцениваются как положительные характеристики их личности, а правильные, наоборот, свидетельствуют о необходимости принятия коррекционных мер. Иногда границы между тестами стираются, особенно в тех случаях, когда педагогические тесты выступают в качестве предикторов успешности профессиональной деятельности или нацелены на проверку логического мышления и творческих способностей учащихся. На существующее сходство

Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования

между измерителями в психологии и в образовании обычно указывает высокая корреляция, которую можно оценить по эмпирическим результатам выполнения тестов.

3.4. Обзор современных отечественных и зарубежных исследований по проблемам тестирования в образовании

Исследования в России в конце XX — начале XXI в. Развитие научных исследований по проблемам педагогических измерений в России тесно связано с расширением сферы их приложения, поэтому период активизации исследований по тестовой проблематике совпал с началом эксперимента по введению ЕГЭ и повсеместным созданием в вузах России систем менеджмента качества образования.

В конце XX — начале XXI в. появились исследования по тестам и смежной проблематике В. С. Аванесова, Т. М. Балыхиной, Н. Ф. Ефремовой, И. И. Легостаева, А. А. Макарова, Е. А. Михайлычева, В. Г. Наводнова, В. П. Панасюка, В. Ю. Переверзева, И. Д. Рудинского, М. Б. Челышковой и др.

С конца 90-х гг. XX в. вплоть до настоящего времени из года в год растет число монографий, учебных пособий, статей и методических работ по вопросам измерений и тестирования в образовании.

В целом современные исследования отечественных ученых по проблемам измерения и тестирования в образовании осуществляются по пяти направлениям. К ним относятся: 1) разработка научных подходов к совершенствованию содержания измерителей в контексте современных трактовок качества учебных достижений; 2) разработка теоретико-методологического и 3) методического обеспечения процесса конструирования надежных и валидных тестов; 4) научное обоснование процедур применения тестов; 5) научное обоснование шкалирования данных тестирования и их интерпретация для применения результатов педагогических измерений в управлении качеством образования.

Несомненный интерес представляют также анализ и психолого-педагогическая интерпретация латентных характеристик подготовленности выпускников, развитие моделей содержательной преемственности тестовых материалов для итогового контроля на различных ступенях обучения (предшкольное, начальное, основное, среднее, высшее), необходимость разработки которых резко возросла в наше время в связи с созданием Общероссийской системы оценки качества образования (ОСОКО).

К перечисленным направлениям исследований примыкает работа по формированию репрезентативных выборок учащихся для проведения мониторинговых исследований, для выравнивания шкал по параллельным вариантам тестов, разработке моделей и

Гомельский государственный университет им. Ф. Скорины

технологий информационного сопровождения тестирования, хранения и автоматического формирования тестов и др.

Проведение исследований по выделенным направлениям в нашей стране, безусловно, позволит значительно повысить качество используемых в образовании измерителей и тем самым будет способствовать повышению качества образования.

Основные направления современных зарубежных исследований по проблемам измерения и тестирования. В современных зарубежных исследованиях в теории педагогических измерений к приоритетным направлениям исследовательских работ относятся:

– развитие методологии сочетания качественных и количественных методов при эвалюации в образовании;

– оптимизация методов интеграции данных педагогических измерений, полученных с помощью инновационных форм заданий тестов;

– развитие параметрических и непараметрических моделей IRT;

– разработка методик, алгоритмов и математико-статистического аппарата теории педагогических измерений для создания программного обеспечения и практического использования новых моделей IRT;

– исследования психологических и этических проблем в тестировании;

– разработка научно-методических подходов к построению уровневых шкал в образовании, разработка новых методов выравнивания;

– развитие специальных процедур и подходов для выявления размерности пространства измерений и ее адекватности поставленным целям тестирования и т.д.

Многие зарубежные исследования посвящены аппарату и методам теории IRT, которая широко используется в практике тестирования. Работу педагогов значительно облегчают многочисленные программно-инструментальные и программно-педагогические продукты, реализующие алгоритмы IRT для оценки результатов испытуемых и конструирования новых тестов. К числу наиболее интересных, созданных мировым лидером в компьютерном тестировании Assessment Systems Corporation (ASC), можно отнести такие программы, как RASCH, RASCAL, Quest, ConQuest, а также программы XCALIBRE, ASCAL, LOGIMO, MSP, PARELLA и многие другие [76; 79]. Некоторые из разработок корпорации ASC, например MieroCAT, CAT, позволяют реализовывать адаптивные варьирующие алгоритмы с переменным шагом и осуществлять процессы генерации адаптивных тестов.

Таким образом, в XXI в. за рубежом теория IRT заняла лидирующее положение в сфере конструирования и применения педагогических тестов, а результаты, полученные такими исследователями, как Ф. Лорд (F. M. Lord), М. Новик (M. Novick), Е. Самеджим

Северо-Восточный федеральный университет им. М.К. Аммосова

(E. Samejima), Д. Вэйс (D. Weiss), Б. Райт (B. Wright), Ури (Urry) и др., внедрены в практику разработки и применения тестов [82; 83].

3.5. Тесты и учителя

Причины затруднений учителей при работе с тестами. Использование тестовых методик предполагает наличие определенных условий, которые имеются далеко не во всякой школе. Прежде всего требуются сами тестовые задания, разработка которых не может быть инициативной работой одного учителя в ущерб своему свободному времени. Желательно, чтобы в создании тестов принимали участие почти все преподаватели школы, чтобы они обменивались своими достижениями, совместно преодолевали трудности и получали благодаря поддержке администрации денежное вознаграждение за свою дополнительную работу. Необходимо также иметь программно-инструментальное обеспечение для ведения баз данных тестирования, хранения банков заданий, выдачи заданий учащимся в компьютерной форме или распечатки заданий на бланках и обработки данных тестирования. Как уже отмечалось, перечисленные условия есть далеко не во всех учебных заведениях, поэтому нередки случаи, когда по возвращении в школу после очередного повышения квалификации по тестовым методикам учитель сразу забывает все, чему научился, и включается в повседневный учебный процесс.

Считается, что педагогам с гуманитарной базовой подготовкой трудно освоить содержание курса по педагогическим измерениям, в котором используется математико-статистическое обоснование качества тестов. Если вы недостаточно сильны в математике и боитесь не освоить методики разработки и применения тестов, то ваши опасения необоснованны. Многолетний опыт преподавания курса по теории и практике конструирования тестов, накопленный авторами этой книги при работе с преподавателями средней и высшей школы, свидетельствует о том, что отсутствие специальных знаний по математике не является препятствием в освоении этого курса.

При изложении материала в нем используется основной набор математических действий — сложение, вычитание, умножение и деление, а простейшие понятия математической статистики легко усваиваются всеми педагогами вне зависимости от базового образования по мере изложения методик разработки и применения тестов.

Какие тесты необходимо разрабатывать и применять всем учителям? Без сомнения, математико-статистический аппарат, используемый профессионалами в полном объеме при разработке

Электронный архив библиотеки МГУ имени А.А. Кулешова

тестов для административно-управленческих решений, довольно сложен, но он учителям и не нужен. В основном в учебном процессе используются тесты для входного и текущего контроля, не нуждающиеся в серьезном статистическом обосновании. От разработчиков таких тестов требуется лишь владение методикой отбора содержания, знание требований к тестовым формам и простейших показателей дескриптивной статистики, необходимых для выполнения основных требований к качеству теста. Поскольку на тесты для текущего контроля приходится основная доля по времени и объему использования, то будет верным считать, что учитель —

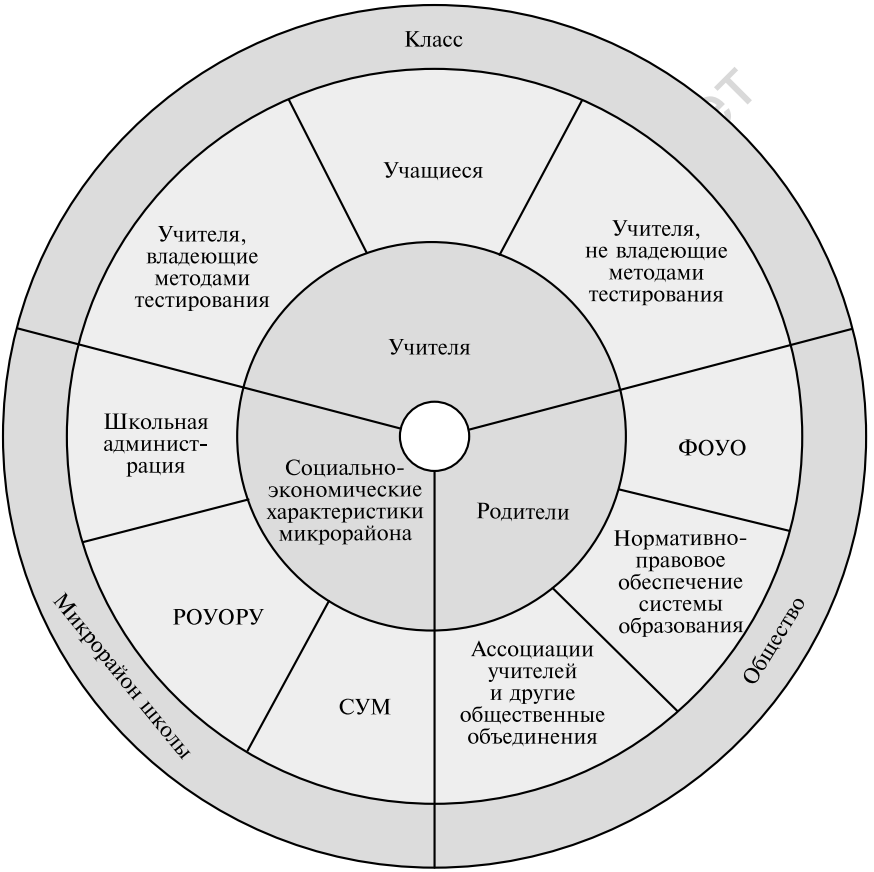

Рис. 1. Важнейшие факторы, влияющие на интерпретацию результатов педагогического тестирования:

РОУОРУ — руководители органов управления образованием районного уровня; СУМ — сообщество учителей микрорайона; ФОУО — федеральные органы управления образованием

ключевая фигура в разработке и применении педагогических тестов.

Факторы, влияющие на интерпретацию результатов тестирования. Помимо разработки тестов для текущего контроля учителю также нужно обязательно овладеть методикой применения педагогических тестов, интерпретацией их результатов, в первую очередь, для разъяснения родителям учеников проблем и трудностей, с которыми сталкиваются их дети. Такая интерпретация не может иметь изолированный, полностью беспристрастный характер. На нее оказывают влияние многочисленные факторы и сопутствующие элементы социально-экономической среды, окружающей учащихся, учителя, класс и в целом школу (см. рис. 1).

Результаты тестирования влияют на отношения педагога с учащимися и их родителями и на его статус в профессиональном сообществе. Поэтому умения применять педагогические тесты, шкалировать и интерпретировать результаты их выполнения необходимы каждому учителю для самооценки эффективности своей работы, выявления в ней слабых мест и поиска факторов профессионального роста. Игнорирование или отрицание потребности в тестировании не отменят эту потребность, поэтому, учитывая, что педагог работает в условиях ограниченного времени, правильнее было бы начать осваивать разработку и применение тестов еще в студенческие годы, до начала педагогической деятельности.

Северо-Восточный федеральный университет им. М.К. Аммосова

Вопросы для обсуждения

1. Каково ваше отношение к тестам в образовании? Под влиянием каких факторов оно сложилось?
2. Какие периоды можно выделить в отечественной и зарубежной истории развития тестов? Каковы их отличительные черты?
3. Каковы приоритетные направления исследований в теории педагогических измерений? Какие из них, по вашему мнению, являются первоочередными для развития тестирования в школах России?
4. В чем различие педагогических и психологических тестов?

ГЛАВА

4

ПЕДАГОГИЧЕСКИЕ ИЗМЕРЕНИЯ. КОМПОНЕНТЫ И УРОВНИ ИЗМЕРЕНИЙ

4.1. Основные понятия теории педагогических измерений

Измерение в образовании, латентные переменные. Согласно наиболее распространенному определению, введенному в 1946 г. американским психологом С. Стивенсом, измерение — это процедура приписывания чисел некоторым характеристикам объектов в соответствии с определенными правилами [6; 22]. Данное определение — результат формального обобщения опыта количественных измерений, широко применяемых в физике и других естественных науках, однако на протяжении многих лет его брали за основу и в эмпирических науках.

По мере развития педагогики, психологии и социологии возникла потребность во введении не только количественных, но и качественных оценок для величин, отличающихся по степени проявления того или иного свойства. Качественные оценки являются менее точными по сравнению с количественными в силу применяемых способов и инструментов измерения. Например, классифицирующие понятия в образовании («знающий», «подготовленный» и др.), которые дифференцируют обучающихся по уровню знаний и играют важную роль в учебном процессе, определяются субъективно учителем или группой учителей. Нередко качественные оценки выражают с помощью чисел, которые выбирают на основе экспертных суждений и соглашений. Приписываемые числа могут трактоваться по-разному. Так, в традиционном педагогическом контроле у каждого учителя есть свои представления о том, за что нужно ставить «5», «4», «3» и т.д.

Неоднозначность оценивания в образовании усугубляется латентным (скрытым, исключающим возможность непосредственного измерения) характером измеряемых переменных. В силу латентности оцениванию подвергаются не сами характеристики обученности и обучаемости, а их эмпирические референты — наблюдаемые признаки измеряемых характеристик. Выбор последних происходит интуитивно, поэтому их соответствие латентным характеристикам нуждается в доказательстве на основе экспертного и статистического анализа эмпирических результатов измерения.

Электронный архив библиотеки МГПУ имени А.А. Кулешова

Современная трактовка понятия «педагогическое измерение». Современная теория измерений появилась в 80-х гг. XX в. Она строится на более строгой аксиоматической основе [22; 34]. В соответствии с новыми представлениями, измерение трактуется как конструирование числовой функции, осуществляющей изоморфное отображение некоторой эмпирической структуры в соответствующим образом подобранную числовую структуру.

Изоморфизм — важное понятие математики, которое определяет ряд условий взаимно однозначного отображения двух множеств с сохранением их свойств в процессе такого отображения. Хотя это понятие впервые появилось в высшей алгебре, в наше время оно используется довольно широко, хотя и не вполне строго, например в педагогических измерениях. Поскольку эмпирическая структура и строящаяся по результатам оценивания числовая структура (шкала) изоморфны, имея шкалу, можно, не обращаясь непосредственно к измеряемым объектам, восстановить все их свойства, характерные для эмпирической структуры.

Компоненты процесса педагогических измерений. Процесс педагогических измерений включает:

– выбор предмета измерения (латентных характеристик объектов) и их числа;

– выбор эмпирических референтов (наблюдаемых характеристик объектов);

– выбор измерительных процедур;

– конструирование и использование измерительных инструментов;

– выбор шкалы (если измеряемая переменная одна) или шкал (если измеряют более одной переменной при многомерных измерениях);

– построение отображения результатов измерения на шкалу (шка́лы в случае многомерных измерений) по определенным процедурам и правилам;

– обработку, анализ и интерпретацию результатов измерения.

В силу неизбежности ошибок измерения оцениваемые характеристики объектов могут принимать более или менее точные значения, поэтому эти характеристики принято называть переменными измерения. Любые отклонения от стандартизированных условий измерения, обработки, анализа и интерпретации полученных результатов увеличивают ошибки измерения, которые представляют наибольшую опасность в эмпирических науках в силу латентного характера переменных. Поэтому так важен анализ устойчивости и точности (надежности) результатов тестирования, что выгодно отличает тесты от традиционных оценочных средств [22; 46; 60].

Еще одна характеристика качества результатов тестирования — валидность — отражает адекватность эмпирических результатов поставленным целям измерения [22; 60]. В силу многогранности це-

Северо-Восточный федеральный университет им. М.К. Аммосова

лей анализ валидности должен быть многоаспектным, но в любом случае важное место занимает доказательство адекватности эмпирических референтов концептуально выделенной переменной (переменных) измерения (конструктная валидность).

Измерительный инструмент. Измерительный инструмент включает два компонента. Первый компонент — само измеряющее устройство, роль которого в педагогических измерениях чаще всего, но не всегда выполняет тест. В самом обобщенном виде под тестом можно понимать совокупность контрольных заданий в стандартизированной форме, обладающих необходимыми системообразующими статистическими характеристиками и обеспечивающих обоснованные оценки концептуально выделенной переменной (переменных) измерения с высокой объективностью. Таким образом, в самом определении теста заложены требования к его качеству, отсутствующие в традиционных оценочных средствах.

Второй компонент измерительного инструмента — заранее подготовленная шкала, которая служит для фиксации результатов измерения и на которой откладываются оценки (количественные или качественные) измеряемой переменной. В процессе упорядочения оценок каждому элементу совокупности наблюдаемых эмпирических данных ставится в соответствие определенный балл, устанавливающий положение наблюдаемого элемента на шкале, где можно размещать сырые (первичные) баллы (результаты суммирования оценок по отдельным заданиям теста) или производные баллы, получающиеся в результате преобразования первичных оценок для повышения сопоставимости и удобства интерпретации результатов учащихся.

Шкала с отложенными оценками переменной является целью измерения. При измерениях с высокой надежностью и валидностью она адекватно отображает оцениваемые характеристики и представляет их без существенных искажений. В зависимости от количества оцениваемых характеристик объекта можно говорить об *одномерных* (одна переменная) или *многомерных* (более одной переменной) *измерениях*. Соответственно по результатам измерения строится одна шкала или несколько шкал, число которых в последнем случае обычно бывает равно числу переменных измерения.

Обработка и анализ данных измерения. Последний компонент процесса педагогических измерений, включающий обработку, анализ и интерпретацию данных, служит для выявления обеспечиваемого качества результатов измерения, коррекции тестов и представления полученных данных в форме, удобной для интерпретации и сравнения. Благодаря сопоставимости тестовых баллов, достигаемой в процессе обработки, по результатам педагогических измерений можно выстраивать качественный анализ результатов учащихся, проводить мониторинг и принимать обоснованные управленческие решения в образовании.

Электронный архив библиотеки МГУ имени А.А. Кулешова

4.2. Объективность педагогических измерений

Может ли быть абсолютная объективность? Появление первых стандартизованных тестов в образовании вызвало массовую позитивную реакцию, поскольку первоначально они рассматривались как средство получения объективных оценок подготовленности обучаемых, преодолевающее субъективизм традиционных оценочных средств. По мере развития теории педагогических измерений и накопления опыта применения тестов пришло понимание того, что абсолютная объективность — это недостижимая характеристика результатов любых, в том числе и педагогических, измерений в силу существования ошибочных компонентов, неизбежно смещающих оценки. Поэтому при использовании тестов можно говорить лишь о высокой или низкой объективности, степень проявления которой связана с величиной надежности теста.

Наиболее полно трактовка термина «объективность измерений» представлена в исследованиях Е. Вебстера (E. Webster) [22], предложившего восемь толкований этого понятия. Три из них — процедурная объективность, классическая (традиционная) объективность и инвариантная (специфическая) объективность — непосредственно относятся к педагогическим измерениям.

Процедурная объективность. Под процедурной объективностью (первая трактовка термина «объективность измерений») понимается независимость результатов тестирования от субъективных суждений педагога, использующего тест. Эта независимость обеспечивается благодаря равенству условий тестирования, использованию для одной группы тестируемых параллельных (совпадающих по трудности и другим характеристикам) вариантов теста, стандартизации процедуры проверки результатов и максимальной ее автоматизации, исключающей влияние педагога на оценки.

Сведение всех видов объективности только к процедурной недопустимо, поскольку при таком подходе не выдвигается никаких требований к качеству теста. В этом случае может создаться впечатление, что для получения объективных данных о подготовленности испытуемых достаточно перейти от традиционных экзаменов к любым, в том числе некачественным, тестам, устранив влияние педагога на оценку тестирования путем автоматизации процедуры подсчета баллов испытуемых.

Классическая, или традиционная, объективность. Второе, углубленное, понимание объективности измерений рассматривается в классической теории тестов и основывается на понятиях «сырой балл» и «истинный балл», отличающихся друг от друга на величину ошибки измерения.

Сырой (первичный, наблюдаемый, индивидуальный) *балл* получается простым суммированием результатов испытуемого по

отдельным заданиям теста. При дихотомической оценке результатов по заданиям (1 или 0) индивидуальный балл равен количеству правильно выполненных заданий теста. *Истинный балл* в классической теории отождествляется с абсолютно объективной оценкой свойств испытуемого, свободной от влияния любых ошибок измерения. В отличие от сырого балла, который меняется в зависимости от теста и способа подсчета результата испытуемого, истинный балл трактуется как не зависящая от средств измерения константа, характеризующая оцениваемое свойство испытуемого в момент измерения, но меняющаяся в процессе обучения.

Согласно основной аксиоме классической теории тестов, любой наблюдаемый балл равен сумме истинного балла и ошибки измерения. В тех случаях, когда ошибка измерения не превышает выбранных пределов точности измерений, говорят о высокой объективности результатов тестирования, а оценки испытуемых принимают за их истинные баллы. Таким образом, углубленное понимание объективности измерений требует оценивания величины ошибки измерения, на размер которой влияют не только условия проведения тестирования, но и качество теста.

Инвариантная, или специфическая, объективность. Третья трактовка объективности основана на современной теории конструирования тестов — Item Response Theory (IRT). Преимущества IRT, позволяющие оценить подготовленность обучаемых независимо от трудности заданий теста, приводят к достижению так называемой инвариантной объективности измерений, которая предпочтительнее объективности, обеспечиваемой классической теорией тестов [60].

Для достижения специфической объективности необходима подгонка данных тестирования к требованиям моделей теории IRT и длительная серьезная работа над тестом. Поэтому на практике тестологи часто сталкиваются с тем, что эффект инвариантной объективности либо реализуется со слишком большими затратами, либо не реализуется вообще в силу недостаточно высокого качества теста.

4.3. Размерность пространства измерений, одномерные и многомерные конструкты, латентные переменные

Концептуальные и реальные переменные измерения, конструкты. Измерение начинается с постановки цели, в соответствии с которой выбирают одну (одномерный случай) или несколько (многомерный случай) переменных. В последнем случае для обозначения измеряемых характеристик часто используют обобщающий термин — «конструкт». Каждый разработчик теста уверен в том,

Северо-Восточный федеральный университет им. М.К. Аммосова

что он ясно представляет себе измеряемые характеристики и способен на основании своего педагогического опыта точно подобрать задания, обеспечивающие оценивание конструкта. Многие тестологи так и остаются в полном заблуждении относительно того, что на самом деле измеряет тест, поскольку не проверяют соответствие задуманного конструкта и реальных результатов измерения.

Анализ такого соответствия является необходимым этапом оценивания валидности измерения. Нередко его пытаются провести априорно, до начала тестирования, экспертным путем, и в результате, как правило, получают недостоверную информацию. Для корректного оценивания валидности необходим статистический анализ эмпирических данных тестирования, поэтому понять, что же мы на самом деле измеряем, можно лишь после применения теста.

Сложность процедуры установления размерности пространства измерений увеличивает проблемы, связанные с неоднозначной трактовкой многих конструктов в образовании и в других социальных науках (рис. 2). Каждый педагог вкладывает в оценивание учебных достижений свое ви́дение оптимального набора переменных измерения. Неоднозначность трактовки конструкта усугубляется по мере продвижения от начальных ступеней образования к более высоким ступеням, когда содержание большинства учебных курсов приобретает междисциплинарный характер.

Операционализации. Операционализация заключается в придании оцениваемым латентным характеристикам подготовленности учащихся формы, удобной для фиксации определенными правилами измерения. При педагогическом измерении в качестве таких

Рис. 2. Пример неоднозначной трактовки конструкта и ошибки в выводах о связи переменных

Сетевое электронное издание федерального университета им. М.К.Аммосова

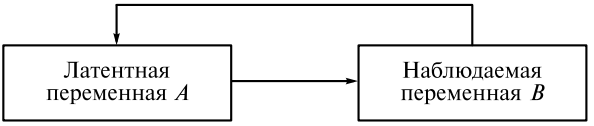

Рис. 3. Связь между латентной и наблюдаемой переменными

характеристик подготовленности обучаемых обычно выступают знания, умения, навыки, компетентности и т.д.

В процессе операционализации происходит выделение набора эмпирических индикаторов, в роли которых выступают задания теста. Количество правильно выполненных заданий, подсчитанное и преобразованное по определенным правилам, дает основание для присвоения испытуемому определенного места на шкале переменной измерения.

Визуализация результатов педагогического измерения. Визуализация — геометрическая интерпретация связи между латентной переменной A (одномерной или многомерной) и наблюдаемой переменной B — показана на рис. 3.

Стрелки на рисунке указывают характер связи между переменными. Латентная переменная A является первопричиной, порождающей множество наблюдаемых результатов выполнения теста. Однако при измерениях всегда ставят обратную задачу — по наблюдаемым результатам тестирования найти достаточно точные оценки латентных переменных.

Взаимосвязь результатов измерения и положения испытуемого на шкале переменной для одномерного случая представлена на рис. 4. Каждая оценка переменной измерения для учащихся из тестируемой группы соответствует одной из точек оси. В свою очередь каждая точка определяет положение испытуемого или группы испытуемых с одинаковым тестовым баллом, полученным по результатам выполнения теста.

На изображенной оси более высокие баллы располагаются правее, а более низкие — левее. Крайний слева результат отражает случай, когда испытуемый выполнил правильно лишь несколько заданий теста. Противоположной ситуации, когда ученик выполнил все или почти все, соответствует крайняя правая точка на оси

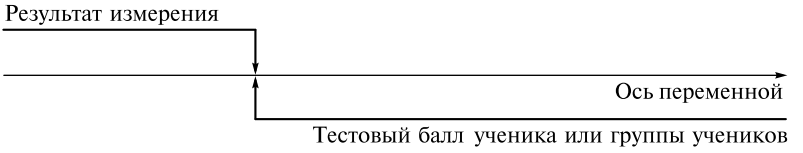

Рис. 4. Геометрическая интерпретация результата тестовых измерений

Восточный университет им. Г.Р. Державина

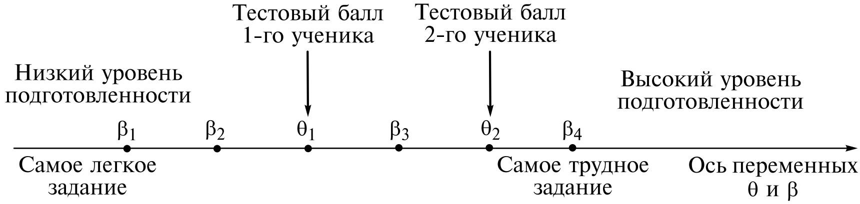

Рис. 5. Соотношение между трудностью заданий и подготовленностью учащихся:

β_j — уровень трудности j-го задания, $j = 1, 2, ..., 4$, θ_1 и θ_2 — тестовые баллы двух учащихся

переменной измерения. Остальные точки занимают некоторое промежуточное положение на отрезке, где лежат тестовые баллы учащихся.

Если правильно выполненные задания теста соотнести с результатами учащихся и расположить их вдоль оси переменной измерения, то можно предположить, что более трудные задания сместятся вдоль оси вправо, так как их, скорее всего, будут выполнять правильно наиболее сильные учащиеся в классе. И наоборот, более легкие задания будут смещены влево — они по силам ученикам с низким уровнем подготовки (см. рис. 5).

Из дидактических соображений на рисунке показано выполнение четырех заданий, однако все выводы, получаемые с помощью этого примера, применимы к любому числу заданий в тесте. Расположение тестового балла первого учащегося говорит о том, что он выполнил верно два самых легких задания, но не справился с третьим и четвертым заданиями. Второй учащийся имеет более высокий тестовый балл и подготовлен лучше. Он не выполнил только самое трудное — четвертое задание теста.

Ошибки измерения. Локализация места расположения результата ученика на оси переменной зависит в основном от соотношения между величиной его истинного балла и трудностью заданий теста. Если балл довольно высок, а задание довольно легкое, то у ученика все основания для успешного выполнения этого задания теста. В противном случае ученика скорее всего ждет неудача.

Конечно, наверняка предугадать ничего нельзя в силу действия различных смещающих факторов (эффект забывания, подсказки и т.д.), поэтому обычно говорят лишь о некоторой вероятности успеха или неуспеха.

Вероятностный характер наблюдаемых результатов выполнения теста обусловлен влиянием случайных и неслучайных ошибок измерения. В число последних входят те, которые появляются из-за просчетов разработчиков в процессе создания теста. К ошибкам систематического характера могут также привести нарушение тре-

Северо-Восточный федеральный университет им. М.К. Аммосова

бований к сбору статистических данных, некачественная интерпретация результатов выполнения теста и ряд других причин. К случайным факторам можно отнести настроение испытуемого, поведение экзаменатора, обстановку при тестировании в классе и многое другое — словом, все, что учесть и предвидеть при тестировании попросту невозможно.

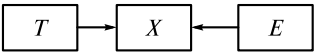

Рис. 6. Иллюстрация связи переменной измерения, истинного балла и ошибки при одномерном измерении

Одномерные измерения. Чаще всего при планировании измерений в образовании выбирают одномерные конструкты. Это упрощает процесс построения шкалы, но не всегда бывает адекватно содержанию тестов. Рис. 6 иллюстрирует случай одномерных измерений, который может быть интерпретирован следующим образом: одна латентная переменная T — истинный уровень подготовленности каждого обучаемого — приводит к возникновению одной оценки наблюдаемой переменной X — уровня подготовленности обучаемого. Помимо переменной T на оценку X оказывает влияние фактор E — ошибка измерения.

Чтобы принять гипотезу об одномерности теста, необходимо выявить связь между теоретическим конструктом и эмпирическими индикаторами, роль которых выполняют задания теста. Оценка связи требует ответа на вопрос — есть ли разница между доказательством одномерности конструкта и доказательством одномерности заданий теста?

На рис. 7 представлена измерительная модель для одномерного случая, иллюстрирующая связь между конструктом, обозначенным символом T, и четырьмя заданиями (X_1, X_2, X_3, X_4). Числа, стоящие у каждого луча, показывают меру предполагаемой корреляционной связи между конструктом и эмпирическими индикаторами — заданиями теста.

При анализе модели важно понимать, что конструкт является латентным (скрытым от возможностей непосредственного измерения) фактором, взаимодействие которого с заданиями порождает наблюдаемые результаты выполнения теста. Влияние конструкта на наблюдаемые переменные показано на рис. 7 с помощью направленных лучей.

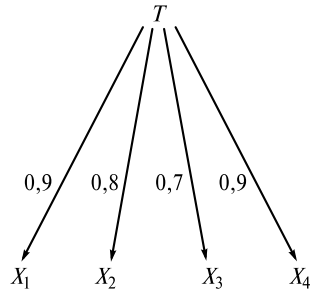

Рис. 7. Измерительная модель, иллюстрирующая связь между конструктом и заданиями теста (одномерный случай)

Поскольку каждое задание в рассмотренном гипотетическом примере измеряет только один конструкт, то справедлив вывод об одномерности заданий теста. Обратный вывод, в об-

Белорусско-Российский университет им. М.М. Шмырева

щем случае, неверен: из одномерности заданий не следует одномерности теста.

Многомерные измерения. Если конструкт включает не одну, а несколько переменных, то измерения называются многомерными. Совокупность переменных образует пространство переменных измерения, размерность которого равна их числу. Иногда при проведении многомерных измерений создают несколько субтестов, каждый из которых является одномерным и измеряет свою переменную с помощью одномерных заданий.

Примером такого подхода является полидисциплинарный тест, состоящий из набора одномерных субтестов. В другом случае в многомерных измерениях используют междисциплинарный тест, задания которого не являются одномерными. Каждое из заданий измеряет свою совокупность переменных, которые могут отличаться как по количеству, так и по содержательной трактовке конструкта.

В практике педагогических измерений существуют специальные методы анализа размерности пространства измерений. Такую группу методов предоставляет исследовательский и конфирматорный факторный анализ, применение аппарата которого основано на использовании соответствующего программного обеспечения, например статистического пакета SPSS.

4.4. Уровни измерений в образовании

Типология уровней измерения. Общая типология уровней измерения основывается на проявлении совокупности свойств, лежащей в основе построения шкал. В качестве таких свойств выделяют: идентичность, позволяющую однозначно относить объекты к одной из выделяемых совокупностей; транзитивность, способствующую ранжированию объектов в определенном порядке; метричность, обеспечивающую единую единицу измерения, и наличие абсолютного нуля.

Наиболее общая классификация, предложенная С. Стивенсом [18; 22; 60], включает четыре уровня измерений и фиксирует присущие им свойства. Согласно такой классификации различают шкалы качественные (шкала наименований, или классификаций, и порядковая шкала) и количественные (интервальная шкала и шкала отношений) шкалы. Качественные шкалы иногда называют неметрическими (концептуальными), а количественные — метрическими (материальными). Для каждого уровня измерений существуют группы допустимых преобразований и операций с различными математическими и статистическими величинами, характеризующими измеряемые признаки.

Качественные шкалы. На качественном уровне отнесение эмпирических объектов измерения к различным классам проводится по

Северо-Восточный федеральный университет им. М.К. Аммосова

признаку эквивалентности (шкала наименований, или номинальная шкала) или по признаку упорядочения внутри эквивалентных объектов одного класса (порядковая шкала). Для построения шкалы наименований и порядковой шкалы в основном применяются экспертные методы, при которых оценки на шкале считаются достоверными, если они признаны большинством экспертов.

Примером номинальной шкалы могут служить результаты зачетной сессии, когда все студенты делятся на две группы — получивших и не получивших зачет. Порядковые шкалы используются в образовании в тех случаях, когда педагогический контроль осуществляется традиционными способами без применения тестов. Например, порядковой является привычная четырехбалльная школьная шкала, которую иногда неоправданно называют пятибалльной. Каждой группе учащихся, проявляющей согласно мнению учителя сходные знания, присваивается одинаковый (один из четырех) номер места от двух до пяти.

Недостатки качественных шкал — ограниченная сфера применения и низкая точность измерения. Числа или символы, приписываемые объектам путем экспертного оценивания, субъективны и носят исключительно условный характер. Их нельзя суммировать или проводить с ними другие математические операции.

Количественные шкалы. К количественным шкалам относятся интервальная шкала и шкала отношений. Процесс их построения основывается на измерениях, поэтому представленные в них оценки характеристик объектов отличаются более высокой объективностью по сравнению с оценками в качественных шкалах и поддаются определенным математическим операциям. *Интервальная шкала* используется для упорядочения объектов, свойства которых удовлетворяют отношениям эквивалентности, порядка и аддитивности. В ней определено расстояние между объектами и предусмотрена общая для всех объектов единица измерения, а началом отсчета является условно выбранная нулевая точка. Благодаря существованию единицы измерения в интервальной шкале возможны все арифметические действия над числами, кроме операции деления в силу отсутствия абсолютного нуля. Примером интервальной шкалы в образовании, обеспечивающей корректную сравнимость результатов педагогических измерений, является шкала логитов, построение которой осуществляется на основе теории IRT [22; 60; 67; 83].

Шкала отношений описывает свойства объектов, удовлетворяющие отношениям эквивалентности, порядка, аддитивности и пропорциональности. Последнее свойство появляется благодаря существованию в этой шкале однозначного естественно определенного критерия нулевого проявления измеряемого свойства — абсолютного нуля. Другими словами, шкала отношений является интервальной шкалой с естественным, а не условным началом

Гомельский государственный университет имени Ф. Скорины

отсчета, что расширяет возможности преобразований чисел, приписанных объектам. По сравнению со всеми ранее рассмотренными шкалами эта шкала обеспечивает самый высокий уровень измерений, но реализовать ее в образовании невозможно в силу отсутствия абсолютного нуля.

4.5. Надежность и валидность результатов педагогических измерений

Общие замечания. Размерность, надежность и валидность являются взаимосвязанными свойствами, характеризующими различные аспекты качества педагогических измерений. Выявление размерности — необходимый предварительный этап работ по оцениванию надежности и валидности результатов измерений.

Надежность результатов тестирования. Надежностью (*reliability*) называется характеристика точности тестовых результатов и их устойчивости к действию случайных факторов [60]. По сложившейся традиции термин «надежность» часто, хотя и не совсем верно, используют по отношению к тесту. Однако надежность теста является необходимым, но не достаточным условием получения высокой точности измерений. В случае нарушений требований к стандартизации условий проведения тестирования, проверке и оцениванию его результатов даже с помощью очень надежного измерителя можно получить результаты со значительным ошибочным компонентом.

Концепция истинного балла. Анализ надежности основан на предположении классической теории тестов о связи между наблюдаемым баллом, истинным баллом и ошибкой измерения. Оценка истинных баллов (*true scores*) испытуемых — главная цель всех, кто создает или применяет педагогические тесты. Так как любые результаты тестирования всегда содержат в себе ошибочные компоненты, то приходится заменять истинные баллы — параметры испытуемых — их наиболее достоверными оценками, которые тем точнее, чем надежнее тест.

Концептуальная формула для коэффициента надежности. Основная аксиома классической теории тестов приводит к фундаментальному соотношению, позволяющему получить концептуальную формулу для коэффициента надежности результатов измерений (количественной характеристики надежности), которая связывает дисперсию (показатель разброса) наблюдаемых баллов S_X^2 и дисперсию ошибок измерения S_E^2 с $r_{\text{н}}$ — коэффициентом надежности теста. Эта формула имеет вид

$$r_{\text{н}} = 1 - \frac{S_E^2}{S_X^2}. \qquad (1)$$

Ее значение исключительно теоретическое, поскольку по эмпирическим результатам выполнения теста нельзя подсчитать S_E^2.

Несложный *анализ формулы* для оценивая надежности (1) позволяет сделать выводы о возможных пределах величины $r_н$. Очевидно, что дробь $\frac{S_E^2}{S_X^2}$ всегда неотрицательна, поэтому коэффициент надежности не может принимать значение больше единицы. Максимальное значение $r_н$, равное 1, получается в том случае, когда $S_E^2 = 0$, — случай, который не встречается в практике измерений. Так как величина дроби уменьшается с ростом знаменателя, то естественно предположить, что надежность увеличивается в тех случаях, когда тест обеспечивает высокий разброс тестовых баллов учеников.

Факторы, влияющие на надежность гомогенного теста. На основе постулатов классической теории тестов и различных модификаций концептуальной формулы (1) можно определить факторы, влияющие на повышение надежности теста.

1. Если при компоновке теста отбирать задания, имеющие наибольшую корреляцию с другими заданиями теста, то тест будет иметь высокую надежность и обеспечит низкую погрешность измерения. Другими словами, чем выше содержательная однородность (гомогенность) теста, тем он надежнее. Этот вывод представляет особую важность для коротких тестов (от 20 до 35 заданий). В очень длинных тестах (более 100 заданий) малые значения интеркорреляции заданий могут сочетаться с высокой надежностью теста.

2. Надежность измерений повышается с увеличением длины теста. Этот формальный вывод не всегда согласуется с реальными возможностями учеников. По мере роста длины теста повышается утомляемость и снижается мотивация к выполнению заданий, что в совокупности ведет к росту ошибки измерения. Поэтому при выборе оптимальной длины теста разработчики анализируют группу факторов, среди которых: высокая дисперсия тестовых баллов, нормальный характер их распределения, форма используемых заданий, возраст учеников и время выполнения теста, выбранное в соответствии с целями тестирования и физиологическими возможностями учащихся.

По данным Н. Гронлунда, учащиеся старших классов в среднем за 1 мин могут выполнить одно задание с выбором ответа (при числе ответов не более четырех) [88]. На задание с кратким дополняемым ответом требуется в среднем до 2 мин, а с полным свободно конструируемым ответом — до 5 мин. В целом для обеспечения достаточно высокой надежности измерений рекомендуется проводить тестирование выпускников неполной средней школы (IX класс) в течение 2—3 уроков, а выпускников средней школы (XI класс) — в течение 2—4 уроков.

Собственность Восточноукраинского национального университета им. В. Даля

Валидность результатов педагогических измерений. Валидность — это характеристика адекватности результатов измерения поставленной цели создания теста [60]. Другими словами, валидность — это характеристика того, в какой мере удается измерить именно запланированный конструкт. Поэтому оценивание валидности тесно связано с анализом размерности пространства педагогических измерений.

Оценивание валидности. Количественная оценка валидности получается путем соотнесения результатов измерения с различными внешними критериями (обычно качественного характера), независимо описывающими вне ситуации тестирования все, что собирались измерять. Высокая корреляция результатов измерений с внешними критериями свидетельствует о высокой валидности теста. Наоборот, слабая корреляция указывает на неполную адекватность теста своему предназначению и позволяет сделать вывод о низкой валидности теста. Поскольку можно выбрать достаточно много внешних критериев адекватности теста поставленным целям измерения, существуют различные виды валидности и многочисленные методы ее исследования.

Конструктная валидность. Оценка конструктной валидности связана с выявлением того, насколько хорошо измеряется концептуально выбранный латентный конструкт. При анализе конструктной валидности часто рассматривают корреляцию между результатами по новым и уже существующим тестам, валидность которых подтверждена многолетней практикой их применения. Проводят независимую экспертизу качества содержания теста, используют факторный анализ, позволяющий выстроить факторную структуру теста, анализируют внутреннюю согласованность теста методами корреляционного анализа и т.д.

Содержательная валидность. В педагогических измерениях на первый план выходит исследование содержательной валидности

Рис. 8. Модель обеспечения содержательной валидности измерения

Федеральный институт развития образования им. И.А. Зимней

теста, основанное на тщательной экспертизе. В общем случае содержательная валидность — это степень релевантности и репрезентативности отражения концептуально выделенного конструкта в содержании заданий теста. В основе работы экспертов обычно лежит анализ полноты, значимости, правильности пропорций содержания теста и его соответствия запланированным для проверки видам учебной деятельности (рис. 8).

Практическое задание и вопросы для обсуждения

1. Какие компоненты педагогических измерений вы знаете?

2. Чем результаты оценивания качества подготовленности студентов, полученные на основе педагогических измерений, отличаются от результатов традиционных экзаменов?

3. Какие виды объективности можно реализовать при использовании тестов, разработанных учителем для текущего контроля?

4. Три ученика отвечали на 6 заданий теста, ранжированных по нарастанию трудности. По результатам ответов получились профили:

первый — 1 1 1 0 0 0; второй — 1 0 1 0 1 0; третий — 0 0 0 1 1 1.

Кто, по вашему мнению, лучше усвоил содержание проверяемого курса? У кого из трех учеников будет выше истинный балл? Правомерна ли постановка последнего вопроса по отношению к результатам третьего ученика?

5. Если результаты контрольной работы ваших учеников отложить на оси, то какую шкалу по уровню измерений вы получите? Можно ли подсчитать средний балл учащихся по контрольной работе?

Северо-Восточный федеральный университет им. М.К.Аммосова

ГЛАВА

5

ПЕДАГОГИЧЕСКИЕ ТЕСТЫ, ИХ ВИДЫ И ПРЕДНАЗНАЧЕНИЕ

5.1. Нормативно-ориентированный и критериально-ориентированный подходы в педагогических измерениях

Общие подходы к интерпретации результатов педагогических измерений. При педагогических измерениях интерпретация баллов учащихся может иметь различный характер в зависимости от того, каким способом сравниваются оценки учеников. Согласно одному подходу проводится сопоставление баллов каждого учащегося с результатами определенной группы — выборки учащихся, выполнявших тот же самый тест, для определения места каждого балла по отношению к среднему результату в группе (нормативно-ориентированный подход). Согласно другому подходу результаты испытуемых интерпретируются по отношению к содержательной области, включенной в тест и снабженной определенными критериями выполнения (критериально-ориентированный подход).

Оба подхода дают информацию о подготовленности учащихся, однако она имеет различный характер. В соответствии с этими подходами к интерпретации результатов тестирования выделяют нормативно-ориентированные и критериально-ориентированные тесты.

Нормативно-ориентированный подход и нормы. Стандартизация тестов. Основная цель нормативно-ориентированного тестирования заключается в дифференциации испытуемых по результатам выполнения теста. При интерпретации результатов относительная позиция испытуемого может оцениваться по-разному, поскольку он будет выглядеть лучше на фоне более слабой, чем более сильной группы. Для корректной интерпретации результатов тестирования балл каждого учащегося необходимо сравнивать с *нормами выполнения теста.*

Нормы — это совокупность показателей, отражающая результаты выполнения теста четко определенной выборкой испытуемых — релевантной нормативной группой, репрезентативно представляющей генеральную совокупность тестируемых учащихся [1; 22; 60]. К нормам обычно относят среднее значение тестовых баллов и показатель разброса (вариативности) вокруг среднего значения всех остальных баллов, полученных представительной выборкой тестируемых учащихся (методы подсчета среднего значе-

Северо-Осетинский государственный университет имени К. Л. Хетагурова

ния и показателей вариативности приведены в главе 9). Имея нормы, можно установить положение каждого результата по отношению к среднему баллу по тесту, посмотреть, насколько результат учащегося выше или ниже среднего.

Процесс определения норм называется *стандартизацией теста*. Стандартизация всегда осуществляется на репрезентативной выборке испытуемых, формирование которой — обязательный момент при определении норм теста.

Относительность норм и выборка стандартизации. Тестовых норм, пригодных для интерпретации результатов всех учащихся по любым тестам, не существует. Область применимости любой нормы ограничивается данным тестом и конкретной совокупностью испытуемых, поэтому нормы не абсолютны и не постоянны. Они отражают результаты выборки стандартизации на момент создания теста и подлежат систематическому обновлению и перепроверке.

К нормам предъявляют следующие требования:

– нормы должны быть дифференцированными. Например, тесты для общеобразовательных и профильных школ необходимо стандартизовать на различных выборках, в результате чего получатся, скорее всего, существенно различающиеся нормы;

– нормы должны отражать реальный контингент и актуальные требования к качеству учебных достижений, вытекающие из современной ситуации в образовании;

– нормы должны быть репрезентативными, поэтому они всегда устанавливаются эмпирически в соответствии с результатами тестирования выборки стандартизации (федеральной — для ЕГЭ, муниципальной — для аттестации школ, внутришкольной — для аттестации учащихся в школе).

«Норма» — относительное понятие, тесно связанное с качеством выборки, использованной для стандартизации. Выборка должна точно отражать категорию (или несколько категорий) лиц, для которых предназначен тест, а также быть достаточно большой и сбалансированной для обеспечения столь малой стандартной погрешности, чтобы ею можно было пренебречь в процессе стандартизации теста. Таким образом, при формировании выборки стандартизации необходимо учитывать две переменные — объем и представительность, обеспечивающие в совокупности высокую точность при оценивании норм выполнения теста.

Стратификация выборки. Для равномерного представления различных групп учащихся в популяции испытуемых используют специальный процесс — стратификацию. Стратификация — расслоение выборки на страты, размеры которых должны быть пропорциональны размерам соответствующих популяций в генеральной совокупности учащихся [38]. Обычно в качестве оснований для стратификации выделяют факторы, наиболее связанные с пере-

Северо-Восточный федеральный университет им. М.К. Аммосова

менной измерения. В ЕГЭ к числу таких факторов можно отнести социальное положение родителей выпускника, регион, где расположена школа, ее принадлежность к числу сельских или городских школ и т. д.

Наличие многих факторов стратификации, необходимость анализа пропорций генеральной совокупности испытуемых, проведение апробационного тестирования для определения норм делают работу по стандартизации тестов довольно дорогостоящей и трудоемкой процедурой. Современный уровень развития тестовых технологий позволяет моделировать тесты с прогнозируемыми нормами с помощью IRT, банка калиброванных тестовых заданий и специальных программ для компьютерной генерации вариантов теста.

Информация, прилагаемая к стандартизированным тестам. К стандартизованному тесту необходимо приложить:

– нормы выполнения теста, которые определяются на выборке стандартизации;

– объем выборки стандартизации, основания для ее стратификации и временной период ее использования;

– необработанные результаты выполнения теста для выборки стандартизации.

Сопоставление норм по различным тестам возможно лишь в том случае, если есть основания для утверждения об адекватности выборок стандартизации.

Критериально-ориентированный подход в педагогических измерениях. При критериально-ориентированном подходе в педагогических измерениях результаты учащихся интерпретируются по отношению к содержательной области или требованиям, установленным к учебным достижениям. При *дихотомическом оценивании* («1» или «0») результатов выполнения отдельных заданий балл каждого учащегося подсчитывается путем перевода в проценты доли правильно выполненных заданий по отношению к общему числу заданий теста. В случае *политомических оценок* в проценты переводится отношение сырого балла учащегося, накопленного по заданиям, к максимально возможному баллу по тесту. Полученный для каждого учащегося процент сравнивается со стандартами выполнения — критериями, установленными экспертным путем и прошедшими эмпирическую валидизацию в процессе конструирования теста [1; 48; 60].

При критериально-ориентированном подходе по результатам тестирования можно:

– выявить освоенные и не освоенные знания, умения и навыки и построить индивидуальную образовательную траекторию каждого учащегося;

– ранжировать тестируемых по проценту выполнения и построить рейтинговые шкалы;

Северо-Восточный федеральный университет им. М.К. Аммосова

– разбить испытуемых на две группы с помощью одного критериального балла или на несколько групп с помощью нескольких критериальных баллов, поставив, например, школьные отметки — «два», «три», «четыре», «пять».

Недостатки критериально-ориентированного подхода. Критериально-ориентированный подход имеет недостатки, связанные с необходимостью полного охвата содержания, принимаемого за 100 %, в одном тесте. Аттестационные критериально-ориентированные тесты нередко получаются очень длинными — из 150—300 заданий, выполнить которые даже в старших классах при одноразовом предъявлении просто невозможно. Поэтому при аттестации нередко применяют адаптивное тестирование, позволяющее за счет оптимизации трудности заданий значительно сократить длину теста. Используют также сокращение содержания теста за счет минимизации целей оценивания. Для этого критериально-ориентированные тесты нередко применяют для проверки одного-двух умений или навыков, а при охвате более разнородного содержания выбирают нормативно-ориентированные тесты.

Критериально-ориентированные тесты имеют к тому же довольно ограниченную область применения. Они пригодны в тех случаях, когда можно четко определить знания, умения и навыки по конкретной области содержания и задать их верхний и нижний пределы для корректного определения критериев выполнения тестов. В более сложных и менее структурированных областях знаний, связанных с решением задач творческого уровня, определить верхний предел зачастую невозможно.

Иногда при выполнении таких заданий школьник руководствуется знаниями, но чаще все решают смекалка и догадка. Поэтому при создании тестов, предназначенных для контроля за выполнением задач творческого уровня, следует отдавать предпочтение нормативно-ориентированному подходу или стараться совмещать оба подхода в одном тесте [1].

Различия в нормативно-ориентированном и критериально-ориентированном подходах. Нормативно-ориентированные и критериально-ориентированные тесты различаются по целям создания, методике отбора содержания, характеру распределения эмпирических результатов тестирования, методам их обработки, критериям качества тестов и тестовых заданий, а главное, по интерпретации результатов испытуемых, выполнивших тест.

Содержание критериально-ориентированного теста должно быть достаточно полно. В него включают все то, что условно можно принять за 100%-ный объем, планируемый к усвоению. Содержание нормативно-ориентированного теста фрагментарно, в него включают только те разделы, которых достаточно для дифференциации учащихся по уровню учебных достижений.

Мозырский государственный университет имени И.П. Шамякина

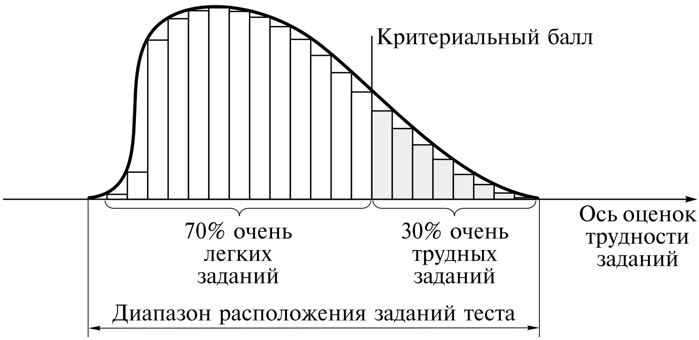

Рис. 9. Распределение заданий по трудности в нормативно-ориентированном тесте

В критериально-ориентированных тестах, используемых для аттестации, задания достаточно простые, поскольку педагоги всегда стараются спланировать процент «двоек» и ограничить число неаттестованных учеников. Например, если «двойки» не должны превышать 10 % и критерий отсева неуспевающих планируется установить на уровне 70 % (все, кто выполнил меньше 70 % заданий теста, получают «два»), то в тест необходимо включить не менее 70 % легких заданий, которые смогут выполнить 90 % тестируемых учеников (рис. 9). Нормативно-ориентированные тесты обычно намного труднее. В них включают от 50 до 70 % заданий средней трудности, т. е. тех, которые смогла выполнить верно только половина тестируемых учеников (рис. 10).

В силу того что распределения сырых баллов репрезентативной выборки испытуемых по нормативно-ориентированным и критериально-ориентированным тестам имеют, как правило, различ-

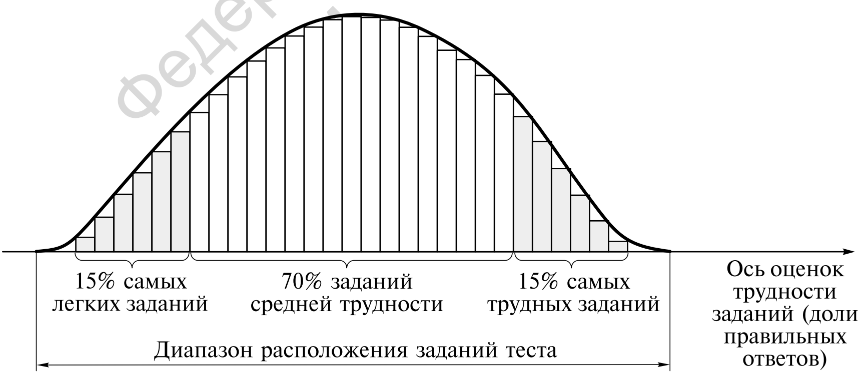

Рис. 10. Распределение заданий по трудности в критериально-ориентированном тесте

Рис. 11. Типичное распределение баллов по тестам для репрезентативной выборки учащихся

ную форму (рис. 11), приходится использовать различающиеся методы оценивания надежности и валидности результатов педагогических измерений, методики шкалирования и выравнивания.

Наиболее существенные различия между нормативно-ориентированными и критериально-ориентированными тестами представлены в табл. 1.

Для учителя наиболее информативной является ситуация, когда оба подхода взаимно дополняют друг друга. Поэтому некоторые

Таблица 1

Различия между нормативно-ориентированными и критериально-ориентированными тестами

Характеристики	Нормативно-ориентированные тесты	Критериально-ориентированные тесты
Типичное среднее число учеников, выполнивших правильно почти все задания теста	5—10 %	80—90 %
Область для сравнения результатов учащихся	Результаты других учеников	Содержательная область или совокупность видов учебной деятельности
Диапазон охвата целей проверки	Широкий, охватывает многие цели и виды учебной деятельности	Узкий, обычно охватывает несколько целей контроля
Репрезентативность охвата содержания предмета	Умеренная, фрагментарная — обычно включают не все разделы	Большая, обычно включают все то, что можно операционализировать и принять за 100 %

Северо-Восточный федеральный университет им. М.К. Аммосова

Окончание табл. 1

Характеристики	Нормативно-ориентированные тесты	Критериально-ориентированные тесты
Разброс результатов учащихся (вариативность баллов)	Высокий, поскольку основная цель тестирования — дифференциация испытуемых по уровню подготовки	Низкий, внутри результатов группы учащихся, превысивших по своим результатам критериальный балл, почти нет вариативности
Подбор заданий по трудности	Распределение оценок трудности близко к нормальному. Основная часть заданий имеет трудность 40—60 %	Распределение скошенное. Основная часть заданий имеет трудность 80—90 %

тесты разрабатываются с расчетом на то, что результаты учащихся можно соотносить как с нормами, так и с содержанием теста. Пример — контрольно-измерительные материалы (КИМ) ЕГЭ.

5.2. Задачи тестирования и виды тестов

Общая классификация задач, решаемых с помощью тестов. В соответствии с видами контроля при тестировании можно выделить:

– задачи, стоящие на входе в обучение (входной контроль);

– текущие задачи (текущий контроль);

– задачи, соответствующие концу определенного периода учебного процесса (итоговый контроль) [60; 74].

Тестирование во входном контроле. Началу обучения соответствует входное тестирование, позволяющее выявить степень владения базовыми знаниями, умениями и навыками, необходимыми для начала обучения, и определить уровень владения новым материалом до начала его изучения в классе. Последняя ситуация кажется не типичной для школы, вместе с тем достаточно вспомнить классический пример, когда в первый класс поступают хорошо читающие дети и начинают скучать на уроках.

Тесты для входного контроля, обычно называемые *претестами* (предварительными тестами), делятся на два типа. Претесты первого типа позволяют выявить готовность к усвоению новых знаний в классе. Они разрабатываются в рамках критериально-ориентированного подхода и содержат задания для проверки базовых знаний, умений и навыков, необходимых для усвоения нового материала. В основном эти претесты предназначены для наиболее слабых учеников, находящихся на границе между явно

Читать Восточный университет им. К. Жамбулова

подготовленными и явно не подготовленными к началу усвоения нового материала. По результатам выполнения претеста проводится деление тестируемых на две группы, в одну из которых попадают те, кто может двигаться дальше, а в другую — те, кто нуждается в дополнительной работе и консультациях педагога.

Претесты второго типа разрабатываются в рамках нормативно-ориентированного подхода. Они охватывают планируемые результаты предстоящего обучения и построены полностью на новом материале. По результатам выполнения претеста преподаватель принимает решение, позволяющее внести элементы индивидуализации в массовый учебный процесс. Если ученик показал некоторые предварительные знания по новому материалу, то план его обучения необходимо перестроить и начать с более высокого уровня, чтобы учебный материал имел для него действительный характер новизны. Иногда роль входного претеста выполняет итоговый тест, который предназначен для будущей оценки результатов усвоения нового материала после завершения его изучения.

На рис. 12 показаны возможные функции входного тестирования в учебном процессе.

Тестирование в текущем контроле. Для текущего контроля разрабатывают корректирующие и диагностические тесты. Корректирующие тесты, как правило, являются критериально-ориентированными: если процент ошибок учащегося превышает критериальный балл, то его знания нуждаются в коррекции. С помощью корректирующих тестов можно найти слабые места в подготовке учащихся и выявить направления индивидуальной помощи в освоении нового материала.

Северо-Восточный федеральный университет им. М.К. Аммосова

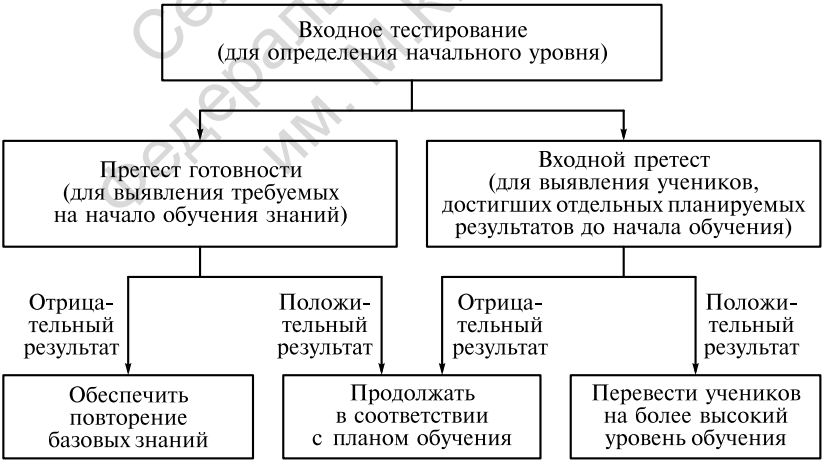

Рис. 12. Упрощенная модель функций входного тестирования в учебном процессе, соотнесенная с задачами педагога

Корректирующие тесты не следует путать со средствами текущего контроля знаний учеников, однако они в какой-то мере близки, хотя бы по целям применения. Однако между первыми и вторыми средствами есть существенные различия технологического и содержательного характера. Традиционные средства текущего контроля менее эффективны и в основном ориентированы на проверку и систематическую оценку знаний учеников по небольшим единицам учебного материала. Корректирующие тесты предназначены для выявления пробелов в знаниях по группе учебных единиц, включающих содержание нескольких тем или даже разделов. Обычно они содержат задания, расположенные по нарастанию трудности, с тем чтобы выявить первые же проблемы в усвоении учебного материала.

Если затруднения ученика при выполнении заданий носят систематический характер, то педагог может прибегнуть к помощи *диагностических тестов*. Основная цель диагностики — установление причин пробелов в знаниях учеников — достигается специальным подбором содержания заданий в тестах. Как правило, в них бывают представлены слабо варьирующие по содержанию задания, рассчитанные по форме представления на отслеживание отдельных этапов выполнения каждого задания корректирующего теста. Подробная детализация позволяет выявить причины устойчивых ошибок учеников, конкретизировать характер возникающих затруднений и получить выводы о несформированности тех или иных учебных умений.

Например, задание с выбором одного правильного ответа из корректирующего теста по математике для начальной школы может иметь следующий вид:

2 + 6 : 3 – 8 : 4 =
А. 2
Б. 3
В. 1
Г. 4

Максимальное число заданий диагностического теста определяется количеством действий при выполнении задания корректирующего теста. Например, для рассматриваемого числового выражения можно предложить четыре задания, если у педагога нет желания проверять знание учащимся порядка действий:

1) 6 : 3 =	2) 8 : 4 =	3) 2 + 6 : 3 =	4) 2 + 6 : 3 — 8 : 4 =
А. 3	А. 2	А. 5	А. 3
Б. 2	Б. 4	Б. 6	Б. 2
В. 4	В. 1	В. 4	В. 0

Подбор заданий в диагностический тест осуществляется в индивидуализированном режиме, в зависимости от тех заданий, которые выполнил неверно каждый учащийся в корректирующем тесте. Особенно эффективны процессы коррекции и диагностики

Северо-Восточный федеральный университет им. М.К. Аммосова

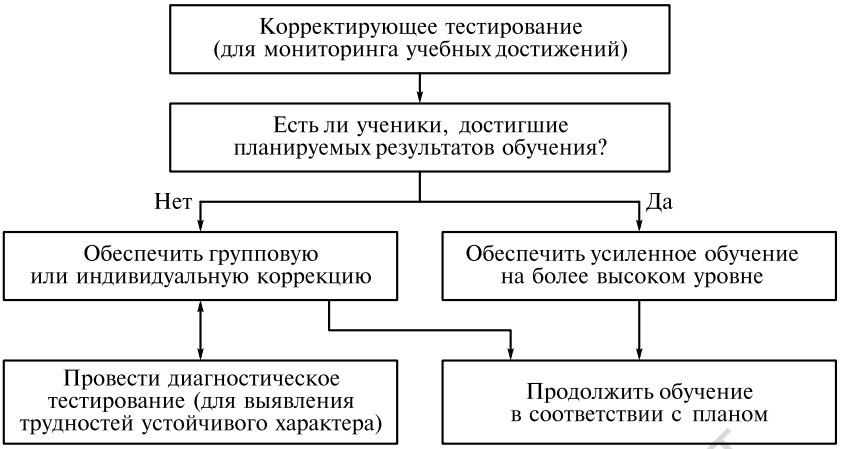

Рис. 13. Модель функций тестирования в текущем контроле

при компьютерной генерации и предъявлении тестов в сочетании с обучающими модулями по каждой единице неусвоенного учебного материала. В этом случае коррекция проводится незамедлительно, поскольку после выявления очередного пробела и установления его причины компьютер сам подбирает обучающий модуль и сразу же выдает его ученику.

Упрощенная модель функций текущего тестирования представлена на рис. 13.

Итоговое тестирование. Основная цель итогового тестирования — обеспечение объективной оценки результатов обучения, которая ориентирована на характеристику освоения содержания

Рис. 14. Модель функций итогового тестирования

Северо-Восточный федеральный университет им. М.К. Аммосова

курса (критериально-ориентированные тесты) или на дифференциацию учащихся (нормативно-ориентированные тесты). На рис. 14 приведена модель функций итогового тестирования.

Итоговые тесты обычно подвергаются стандартизации, поскольку чаще всего они применяются для принятия административных управленческих решений в образовании. Если проведение входного и текущего тестирования — функция учителя, то итоговое тестирование часто проводится внешними структурами и носит характер независимых проверок. Примером независимого итогового тестирования в России является ЕГЭ, тестирование при аттестации школ и т. д. Внутри школы итоговые тесты можно использовать при переводе учащихся из класса в класс, при отборе отстающих учеников для определения их в коррекционные классы и т. д.

5.3. Классификация видов педагогических тестов

Основные подходы к классификации тестов. В отечественной и иностранной литературе существуют различные подходы к классификации педагогических тестов, различающиеся по признакам, которые положены в основу демаркации видов. В соответствии с подходом к интерпретации данных выделяют *нормативно-ориентированные* и *критериально-ориентированные тесты*.

По размерности конструкта педагогические тесты делятся на *гомогенные* (измеряющие только одну переменную и потому однородные по содержанию) и гетерогенные (измеряющие более одной переменной — случай многомерного конструкта) *тесты*. Гетерогенные тесты бывают полидисциплинарными и междисциплинарными [52; 60]. Полидисциплинарные тесты состоят из гомогенных субтестов по отдельным дисциплинам. Результаты учеников по субтестам объединяются для подсчета итоговых баллов по всему полидисциплинарному тесту. Для выполнения заданий междисциплинарных тестов требуется применение обобщенных, межпредметных, интегративных знаний и умений. Междисциплинарные тесты всегда многомерны, их разработка требует обращения к факторным методам анализа данных, математико-статистическим методам многомерного шкалирования и т. д.

По характеру измеряемых переменных выделяют *тесты для проверки знаний*, *учебных*, *практических умений*, *навыков*, а также *компетентностные тесты*. Иногда в отдельную группу выделяют *скоростные тесты*, требующие жесткого временно́го ограничения на выполнение каждого задания и содержащие всегда избыточное число заданий, не позволяющее выполнить весь тест. В зависимости от формы предъявления различают *бланковые* и *компьютерные*, *устные* и *письменные тесты*.

Полесский государственный университет

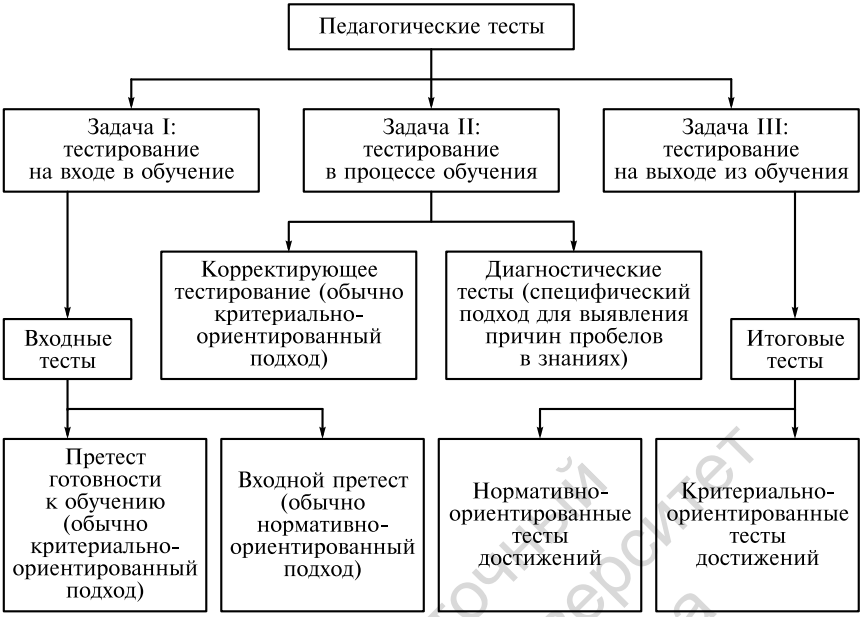

Рис. 15. Классификация педагогических тестов

Наиболее общая классификация тестов в учебном процессе позволяет разделить их на две неравные группы: *стандартизованные тесты*, обладающие нормами выполнения, и *нестандартизованные тесты*, которых значительно больше, поскольку для использования в повседневном учебном процессе их готовит каждый педагог. Нестандартизованные тесты нередко называют учительскими, или авторскими, тестами.

Классификация по видам контроля, их функциям и характеру решаемых задач. Если в качестве признака демаркации выбрать виды контроля и характер задач, решаемых преподавателем с помощью тестов, то получится классификация видов педагогических тестов, представленная на рис. 15.

Анализ классификационной таблицы позволяет выделить в качестве основополагающих четыре вида педагогических тестов, среди которых наибольшую важность по сфере использования имеют итоговые нормативно-ориентированные тесты.

Увеличение влияния тестирования на принятие управленческих решений на основе данных мониторинга и анализа качества образования во многих странах привело в XXI в. к возникновению нового вида тестов административно-управленческого предназначения (в англоязычной литературе — High-Stakes tests). Данные административно-управленческого тестирования являются важным информационным источником для анализа

Полесский государственный университет

последствий образовательных реформ и инноваций в образовании, проведения сравнительных исследований качества подготовленности выпускников различных регионов России, аттестации учебных заведений и оценки эффективности их деятельности.

5.4. Основные определения понятийного аппарата

Понятийный аппарат при разработке и использовании тестов. Необходимость создания четкого понятийного аппарата для разработки тестов не всегда понятна преподавателям-практикам. Отчасти это объясняется кажущейся простотой самих понятий, так как нередко любой набор заданий в тестовой форме в представлении учителя ассоциируется с тестом. Такие псевдотесты сплошь и рядом публикуются в специальных сборниках. Их можно использовать в текущем контроле, но не в работе аттестационных центров.

Несоответствие псевдотестов научно обоснованным критериям качества может привести к значительному ошибочному компоненту в оценках подготовленности учащихся, следствием которого будут неправильные выводы относительно эффективности работы отдельных преподавателей или педагогических коллективов. Таким образом, понятийный аппарат необходим, поскольку он служит целям отделения тестов от того, что нередко за них принимается.

Предтестовое задание. Определение предтестового задания является базовым, содержащим специфические требования, с тем чтобы отличить его от традиционного контрольного задания. Предтестовое задание — это единица контрольного материала, содержание, логическая структура и форма представления которого удовлетворяют ряду требований и обеспечивают однозначность оценок результатов выполнения благодаря стандартизованным правилам проверки [60].

В предтестовых заданиях проверяются наиболее существенные опорные элементы содержания дисциплины. В каждом предтестовом задании определяется то, что однозначно считается правильным ответом с запланированной степенью его полноты.

Требования, предъявляемые к форме предтестовых заданий, можно условно разделить на специальные, отражающие специфику формы, и общие, инвариантные относительно выбранной формы. Согласно общим требованиям задание должно иметь определенный порядковый номер, стандартную инструкции по выполнению, адекватную форме, эталон правильного ответа, стандартизованные правила по оценке результатов его выполнения и т.д. (см. гл. 7). Специальные требования к форме

Национальный технический университет имени Н.Э. Баумана

довольно многочисленны, частично они представлены в главе 7, посвященной формам предтестовых заданий.

Преимущества предтестовых заданий по сравнению с традиционными контрольными заданиями обеспечиваются предельной стандартизацией при предъявлении и оценивании результатов их выполнения, что в целом повышает объективность оценок учащихся по тесту.

Тестовое задание. Предтестовые задания должны пройти обязательную эмпирическую проверку, по результатам которой часть из них превращается в тестовые, а оставшаяся часть удаляется из первоначальной совокупности заданий теста. Предтестовое задание превращается в тестовое, если количественные оценки его характеристик удовлетворяют определенным критериям, нацеленным на эмпирическую проверку качества содержания, формы и системообразующих свойств предтестовых заданий.

Обычно требуется не менее двух-трех апробаций, по результатам которых ведется коррекция содержания, формы, трудности задания, его валидности и статистических свойств, характеризующих качество его работы вместе с остальными заданиями теста. Исследование системообразующих характеристик тестового задания проводится на основе анализа дескриптивной (описательной) статистики, а также методов корреляционного, факторного и латентно-структурного анализа. Интерпретация результатов анализа — это всегда сложная аналитическая работа, результаты которой зависят от множества условий, в том числе и от вида создаваемого теста. Статистические характеристики тестовых заданий и требования к их качеству рассмотрены в главе 9.

В длительной апробации и коррекции нуждаются в основном итоговые тесты, используемые для принятия управленческих решений в образовании. Например, при разработке учительских тестов для текущего контроля корреляционный и факторный анализ не нужны, но дескриптивная статистика, позволяющая без особых усилий отобрать валидные задания приемлемой трудности, будет также очень полезна.

Педагогический тест. В отличие от первых двух определений, инвариантных относительно целей тестирования и решаемых задач, определение педагогического теста должно быть ориентировано на конкретный вид теста. В частности итоговый нормативно-ориентированный тест — это система тестовых заданий, упорядоченных в рамках определенной стратегии предъявления и обладающих такими характеристиками, которые обеспечивают высокую дифференциацию, точность и обоснованность оценок качества учебных достижений.

Из этого определения следуют два важных вывода. Первый: нет и не может быть тестов качественных вообще, так как оценка дифференцирующего эффекта теста, точности измерений (надеж-

ности) и их адекватности поставленным целям (валидности) зависит не только от характеристик тестовых заданий, но и от особенностей тестируемого контингента учащихся. Второй: для оценки качества теста необходимы эмпирические данные тестирования, полученные на репрезентативной выборке учащихся. Работа по коррекции теста консолидирует систему тестовых заданий — постепенно нарастают внутренняя связь и целостность, интегративность системы, совершается переход от совокупности предтестовых заданий к профессионально разработанному тесту.

Итоговый критериально-ориентированный тест — это система тестовых заданий, упорядоченных в рамках определенной стратегии предъявления и обладающих такими характеристиками, которые обеспечивают валидную содержательную интерпретацию учебных достижений по отношению к установленным, статистически обоснованным критериям выполнения [1; 48]. В определении не конкретизируется базовая содержательная область, используемая при интерпретации, что позволяет применять его для различных разновидностей критериально-ориентированных тестов.

Практические задания и вопросы для обсуждения

1. Каковы функции входного тестирования? Есть ли смысл разрабатывать входные тесты в школе?

2. Каковы цели разработки корректирующих тестов? Есть ли различия между корректирующими тестами и традиционными средствами текущего контроля?

3. В рамках какого подхода, по вашему мнению, следует разрабатывать тесты для проведения выпускных экзаменов в школе?

4. Какой процесс называется стандартизацией теста? Перечислите основные факторы, влияющие на устойчивость норм теста в вашей школе.

5. Каковы отличия нормативно-ориентированных тестов от критериально-ориентированных?

6. Сформулируйте определения предтестового задания, тестового задания, педагогического теста. Сравните свой ответ с содержанием соответствующих разделов пособия.

7. Родители Миши хотят обсудить с вами его возможности для поступления по результатам ЕГЭ в университет, в котором вы работаете. Какой подход к интерпретации результатов Миши по ЕГЭ вы будете использовать? Почему?

8. Между директором школы, который вызвал к себе учителя, и учителем состоялся диалог следующего содержания.

Д и р е к т о р. Вы являетесь классным руководителем 6-го «А» класса с углубленным изучением математики. К нам из другой довольно слабой школы нашего района переводят ученика, который по математике имеет одно из первых мест в своем классе. Вот результаты тестов по математике, которые этот ученик выполнил лучше, чем 90 % учащихся его класса. Как вы думаете, он будет у вас хорошо учиться или лучше определить его в другой, нематематический класс?

Северо-Восточный федеральный университет им. М.К. Аммосова

Учитель. Я не уверена в том, что новый ученик сможет учиться в нашем классе. Вы могли бы дать мне подробную содержательную расшифровку результатов его тестирования? Какие знания и умения он получил, на каком уровне он освоил разделы программы?

Директор. Что вы имеете в виду? Я же сказал, что он отлично выполнил тест по математике.

Учитель. Я хотела узнать совсем другое. Мне бы хотелось понять, что он усвоил и чего он не знает.

Директор (раздражительно). Разве вы не умеете интерпретировать результаты тестирования? Возможно, вам следует пройти повышение квалификации по тестированию и оценке знаний учащихся.

Учитель (расстроенно). Я только что прошла курс обучения тестированию и хорошо понимаю, что означают данные, которые вы мне показываете. Я хочу знать, что посоветовать ему, чтобы он не стал получать в нашем классе «двойки».

Очевидно, что между директором и классным руководителем класса с углубленным изучением математики отсутствует взаимопонимание. Директор предоставил учителю результаты учащегося, но они, похоже, педагогу не нужны. Почему? В чем проблема? Какие результаты тестирования и какие тесты нужны, чтобы помочь новому ученику освоиться в классе? Что было бы, если аналогичный вопрос о способностях ученика задали родителям мальчика? Нуждается ли классный руководитель математического класса в дополнительном обучении тестированию или учить следует директора?

Северо-Восточный федеральный университет им. М.К.Аммосова

ГЛАВА

6

СОДЕРЖАНИЕ ПЕДАГОГИЧЕСКОГО ТЕСТА

6.1. Целеполагание при планировании содержания педагогического теста

Общие замечания. Содержание теста формируется путем отображения учебного материала в системе тестовых заданий. Для обеспечения высокой конструктной и содержательной валидности результатов педагогических измерений необходимо использовать определенную методику, включающую вопросы целеполагания, планирования и оценки качества содержания теста.

В содержании теста стараются отобразить то главное, что должны усвоить ученики, поскольку все результаты обучения проверить невозможно. Для этого прежде всего необходимо структурировать цели обучения и ввести определенную их иерархию. Таксономия целей, используемая при разработке теста, должна носить предметно-ориентированный характер, поскольку каждая дисциплина имеет свои приоритеты. При построении таксономии иногда ограничиваются простым перечислением целей обучения, как в примере, рассмотренном в следующем разделе.

Таксономии целей обучения. В настоящее время наиболее известной является *таксономия целей Б. С. Блума* (*B. S. Bloom*) [70]. Данная таксономия очень технологична и с точки зрения большинства разработчиков педагогических тестов вполне приемлема для целей тестирования. В частности в своей классификации Б. С. Блум выделяет:

– знание названий, имен, фактов;
– знание определений и понимание их смысла;
– сравнительные, сопоставительные знания;
– классификационные знания;
– знание противоположностей, противоречий, синонимичных и антонимичных объектов;
– ассоциативные знания;
– причинные знания;
– алгоритмические и процедурные знания;
– оценочные знания и т. д.

В 90-е гг. XX в. таксономия Б. С. Блума подвергалась значительной критике в связи с неполным отражением современных достижений в области психологии обучения и отсутствием связи с

Учебно-методический кабинет Гродненского государственного университета имени Янки Купалы

классификацией видов познавательной деятельности учащихся. К числу ее недостатков зарубежные исследователи отнесли излишнюю упрощенность, не позволяющую использовать современные теории процесса обучения; избыточное внимание к оценке результата обучения, а не к процессу формирования результата; отсутствие зависимости между отдельными составляющими модели [60]. В нашей стране таксономия Б. С. Блума также неоднократно критиковалась в основном в связи с тем, что в ней произошло методологически недопустимое смешение конкретных результатов обучения (знание, понимание и т.д.) с операциями, представляющими необходимое условие их достижения (анализ, синтез, оценка).

Новая концептуальная модель целей обучения, предложенная К. Бигсом и Д. Коллисом (*C. Biggs and D. Collis*), получила название *СОЛО-таксономия* (SOLO — Structure of the Observed Learning Outcomes). В ней содержится детальная классификация категорий познавательной деятельности, позволяющая планировать различные ее уровни в концептуальной модели содержания теста. Спектр уровней познавательной деятельности, представленный в СОЛО-таксономии, достаточно широк: от воспроизведения фактов и простейших алгоритмов до разнообразных интеллектуальных и практических умений, базирующихся на теории Ж. Пиаже об этапах развития познавательной деятельности. Данная таксономия имеет иерархическую структуру, поэтому ее удобно использовать как для разработки инструментария, так и при интерпретации результатов педагогических измерений [28].

В основу планирования содержания тестов может быть положен *уровневый системный подход описания достижений учащихся*, разработанный учеными отечественной педагогической школы и позволяющий сгруппировать результаты обучения в зависимости от уровней учебной деятельности. Первый уровень связан с непосредственным воспроизведением по памяти содержания изученного материала и его узнаванием. Второй уровень предполагает понимание и применение знаний в знакомой ситуации по образцу, а также выполнение действий с четко обозначенными правилами. Третий уровень включает использование знаний в измененной или незнакомой ситуации (И. Я. Лернер, В. П. Беспалько и др.) [7; 33]. Использование уровневого подхода будет более эффективно, если связать уровни усвоения учебного материала с характеристиками внешней деятельности учащегося, задав требования к ее проявлению (табл. 2).

Целеполагание на этапе планирования содержания теста. В отличие от содержания традиционных средств контроля, которое формируется в основном интуитивно на основании практического опыта, отбор содержания теста имеет четкую целевую направленность.

Учебно-методический центр УО «МГУ имени А.А. Кулешова»

Таблица 2

Требования к внешней деятельности учащегося

Уровень усвоения учебного материала	Требования к достижениям учащихся (уровню подготовки учащихся) в обобщенных терминах	Формулировки требований в терминах внешней деятельности
Воспроизведение знаний	Знать терминологию, специфические факты (даты, события, имена людей и т.д.), категории, критерии, методы, принципы, законы, теории и т.д.	Давать определение; называть; формулировать; описывать; устанавливать соответствие (между термином и определением); показывать (находить); распознавать (находить); пересказывать; перечислять (особенности); выбирать и т.д.
Понимание и применение знаний в знакомой ситуации	Понимать факты, законы, принципы, критерии, теории; понимать прочитанный текст; применять знания для объяснения, сравнения, для решения качественных и количественных задач; правильно использовать методы, алгоритмы, процедуры; строить графики, диаграммы, таблицы и др.	Объяснять; соотносить; характеризовать (приводить характеристики); сравнивать; устанавливать (различие, зависимость, причины); выделять существенные признаки; рассчитывать (определять по формулам или алгоритму); решать; составлять что-то по готовой схеме; выполнять в соответствии с правилами; демонстрировать; измерять; продолжать/заканчивать (предложение); вставлять пропущенные слова (буквы) и т.д.
Применение знаний в измененной или незнакомой ситуации	Интегрировать знания из разных разделов для решения различных проблем, анализировать, обобщать, оценивать, конструировать, планировать деятельность, эксперимент	Составлять устный или письменный ответ на проблемный вопрос; писать сочинение; проводить исследование; формулировать гипотезу (выводы); обосновывать свою точку зрения или точку зрения автора; предсказывать последствия; отличать факты от мнений (суждений), факты от гипотез, выводы от положений; анализировать информацию; находить ошибку; высказывать свое мнение, суждения о соответствии выводов и фактов; давать отзыв или рецензию; высказывать суждения о значении (роли) идей, о точности (измерений);

Северо-Восточный федеральный университет им. М.К. Аммосова

Окончание табл. 2

Уровень усвоения учебного материала	Требования к достижениям учащихся (уровню подготовки учащихся) в обобщенных терминах	Формулировки требований в терминах внешней деятельности
		высказывать суждения о качестве (точности, эффективности, экономичности) проделанной работы, о выбранном способе решения или используемых методах; выстраивать модель (изменять модель); реконструировать, составлять план эксперимента, рассказа, решения; изменять план и т. д.

При планировании содержания контроля в методической литературе рассматривают систему изучаемых объектов, виды учебной деятельности и характеристики качества усвоения учебного материала. Описание объектов изучения обычно дается с учетом глубины их освещения учителем и планируемого уровня усвоения учащимися. Группой исследователей НИИ СиМО АПН была предложена общепредметная схема, организующая множество объектов изучения в определенную структуру на основе морфологического и функционального анализа содержания предметов [32]. К важнейшим элементам системы научных знаний исследователи отнесли понятия и факты, законы, теории, идеи, знания о способах деятельности, методологические и оценочные знания и др.

В основу классификации и систематизации видов учебной деятельности можно положить структурированные умения, предложенные И. И. Кулибабой. К ним относятся:

– специальные умения, формирующиеся в процессе изучения отдельных учебных предметов;

– общие умения по организации рационального учебного труда, включающие умения пользоваться различными источниками знаний для решения познавательных задач, планировать и организовывать свою учебную деятельность, контролировать и корректировать результаты учебной деятельности, а также управлять ею в процессе учения;

– интеллектуальные умения, представляющие собой ядро учебной деятельности и объединяющие все учебные предметы.

Характеристики качества усвоения учебного материала можно рассматривать на различных уровнях. Первый уровень — планирование обучения, когда определенные представления о планируемом качестве подготовки закладываются в образовательные программы по каждому предмету. Второй уровень обычно

ассоциируется с этапом реализации образовательных программ в учебном процессе, а третий уровень связывается с оценкой качества результатов учебного процесса.

Обобщение результатов ряда отечественных научно-методических работ позволяет говорить о различии в подходах при трактовке качества результатов учебного процесса. В одних случаях категорию качества отождествляют с полнотой знаний и их глубиной. В других случаях под качеством знаний понимают степень их обобщенности или системности, конкретности и осознанности. Иногда приоритет отдают логичности изложения материала, рациональности способов и приемов решения учебных задач. В практике обучения чаще всего встречается упрощенная ситуация, когда знания обучаемого считают качественными, если он выполняет задания повышенной сложности из числа тех, что предлагаются при контроле в классе.

Возможность оценки качества подготовки как результата обучения в 60—80-е гг. XX в. в отечественной научной школе подвергалась сомнению. По мнению критиков, представление о качестве подготовки должно ассоциироваться с внутренним состоянием обучаемого, в то время как результаты обучения проявляются во внешних, наблюдаемых признаках и результатах учебного процесса. Состоятельность подобных критических суждений легко опровергается в наши дни благодаря достижениям теории педагогических измерений и современной теории тестов IRT.

Идея перехода от внешнего к внутреннему, идея интериоризации, составляет ядро теории педагогических измерений. По наблюдаемым результатам теста с той или иной степенью точности пытаются сделать вывод о внутренних устойчивых характеристиках — латентных параметрах подготовленности учащихся. При контроле совершается обратный ход, поскольку на деле именно множество этих параметров испытуемых в процессе взаимодействия с множеством заданий порождает наблюдаемые результаты выполнения теста. Использование математических моделей измерения теории IRT, соединяющих оба множества и устанавливающих между ними функциональную связь, позволяет решить проблему оценки качества подготовки испытуемых и перейти к оценкам параметров учащихся путем специальной обработки результатов педагогических измерений.

Операционализация и конкретизация планируемых результатов обучения. Корректное планирование содержания теста затрудняют излишняя общность, расплывчатость, многообразие и неопределенность формулировок образовательных целей. Поэтому для создания тестов необходима предварительная конкретизация и операционализация планируемых результатов обучения.

Процесс операционализации заключается в придании форме представления целей обучения тех характеристик, которые позво-

Учреждение образования «Гомельский государственный университет имени Ф. Скорины»

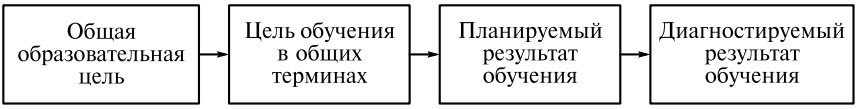

Рис. 16. Этапы операционализации результатов обучения

ляют однозначно отобразить их в содержании тестов [60]. Идея операционализации очень близка положениям М. В. Кларина и зарубежных авторов Т. Кубисзина и Г. Борича (*T. Kubiszyn*, *G. Borich*), которые используют иной, довольно удачный термин «конкретизация» [27].

Конкретизация, по мнению М. В. Кларина, должна начинаться с описания воздействия обучения на ученика, прояснения характера воздействия и детализации его результатов. Поэтому процесс конкретизации требует характеристики образовательных условий, выявления внутренних параметров учащихся, их способности к усвоению нового учебного материала и описания результатов образовательного процесса. Повышению конкретизации описания результатов учебной деятельности, по мнению М. В. Кларина, способствует использование ряда глаголов, непосредственно характеризующих действия ученика.

Для конкретизации учебных целей исследователь предлагает применять следующие глаголы: «анализировать», «вычислять», «высказывать», «демонстрировать», «знать», «интерпретировать», «использовать», «оценивать», «понимать», «преобразовывать», «применять», «создавать» и т. п.

Для конкретизации целей творческого типа — «варьировать», «видоизменять», «модифицировать», «перегруппировать», «перестроить», «предсказать», «поставить вопрос», «синтезировать», «систематизировать» и т. п.

Для обозначения целей в области развития устной и письменной речи — «выделить», «выразить в словесной форме», «записать», «обозначить», «подчеркнуть» (в смысле выделить), «продекламировать», «произнести», «прочитать», «разделить на составные части», «рассказать» и т. п.

Процесс операционализации характеризуется рядом этапов, которые схематично изображены на рис. 16.

После выполнения всех этапов планируемые результаты обучения представляются в виде совокупности контрольных заданий теста. На этой стадии операционализация позволяет структурировать, расчленять, а иногда, наоборот, укрупнять, уточнять и детализировать содержание дисциплины для его отображения в содержании теста. Именно этот этап обеспечивает переход от теоретического уровня анализа целей обучения к эмпирическому и позволяет концептуально правильно подойти к выделению эмпирических индикаторов — заданий теста.

Электронный архив библиотеки Витебского государственного университета имени П. М. Машерова

6.2. Планирование содержания теста

Модели планирования. В процессе планирования содержания теста становится понятным, что далеко не весь набор целей обучения по разным причинам можно отобразить в тесте. Конечно, чем глубже и полнее отображение, тем выше содержательная валидность теста, тем больше оснований доверять тестовым баллам учащихся. Однако на практике приходится думать не только о требованиях тестовой технологии, но и о реальных возможностях школьников определенного возраста, которые в процессе выполнения теста не должны испытывать излишнего напряжения и усталости. В этой связи набор целей обучения необходимо структурировать и выделить самые важные цели, которые должны обязательно попасть в тест.

Структуризация целей различного уровня иерархии позволяет уточнить определенные предметные области, разделы и темы, содержание которых необходимо отразить в тесте. Знание предметных областей выражается правилами измерения с перечислением элементов содержания в совокупности с желаемыми, планируемыми при обучении уровнями владения этими элементами. Таким путем общие цели обучения конкретизируются. В результате возникает прагматическое определение знания учебной дисциплины, когда испытуемый должен правильно выполнить определенные задания теста.

Особые сложности в оценке учебных достижений вызывает выявление уровня владения интеллектуальными и практическими умениями, которые, по своей сути, слабоалгоритмизируемы, сложны, неоднозначны при проверке и требуют, как правило, создания концептуальных моделей, альтернативных существующим. Из апробированных на практике в настоящее время наибольшее применение находит трехмерная модель планирования теста, включающая содержание, технологические требования и планируемый уровень познавательной деятельности, необходимый для выполнения заданий теста [35].

Первая составляющая модели — содержание — обеспечивает содержательную валидность инструментария, его соответствие учебным программам. Под второй составляющей модели — технологическими требованиями — в основном понимается используемый тип заданий. Необходимость введения второй составляющей появилась сравнительно недавно в связи с расширением применяемых форм заданий теста. Помимо традиционных заданий с выбором ответа используются задания со свободным ответом как в краткой, так и в развернутой форме, экспериментальные задания и др. Третьей составляющей модели является уровень осуществляемой познавательной деятельности, на оценку которого направлено измерение.

Учебно-Методический Центр Информационных Технологий

Спецификация теста. После определения целей тестирования и их конкретизации разрабатывается спецификация теста. Для этого делается примерная раскладка процентного соотношения содержания разделов и определяется необходимое число заданий по каждому разделу дисциплины, исходя из его важности и числа часов, отведенных на него в программе.

Раскладку начинают с подсчета планируемого исходного числа заданий в тесте, которое в процессе работы над тестом будет неоднократно меняться в сторону увеличения или уменьшения. Обычно предельное число заданий по математике или физике для учащихся старших классов не превышает 20 — 30, а по гуманитарным наукам нередко доходит до 60 — 80. Оптимальное время тестирования зависит от многих факторов: целей тестирования, возраста учащихся, подхода к разработке теста, специфики контролируемого содержания и т. д. Как правило, в старших классах это время составляет от 1,5 до 2 часов. В ЕГЭ, например, на выполнение теста по математике в 2006 г. отводилось 4 астрономических часа, а на тест по русскому языку — 3.

В спецификации теста обязательно фиксируется структура, содержание проверки и процентное соотношение заданий в тесте. Иногда спецификацию делают в развернутой форме, содержащей дополнительные сведения о тесте. Спецификация в развернутой форме включает:

1) цель создания теста, описание подхода к его созданию и возможных сфер применения теста;

2) перечень нормативных документов (базисных программ, требований к уровню подготовки выпускников и др.) и перечень учебников, используемых при планировании содержания теста;

3) описание общей структуры теста, включающее перечень субтестов (если они есть) с указанием подходов к их разработке;

4) число заданий различной формы с указанием ключей ответов и оценочных правил для заданий со свободно конструируемыми ответами;

5) число параллельных вариантов теста;

6) вес каждого задания, рекомендуемый автором теста, в общей оценке результатов выполнения теста;

7) рекомендуемое время выполнения теста, в том числе на каждый субтест, среднее время выполнения одного задания с учетом специфики формы;

8) соотношение заданий по различным разделам и видам учебной деятельности;

9) рекомендации по контингенту учащихся для апробации теста;

10) характеристику полноты охвата требований ГОС (в процентах) и перечень требований ГОС, не вошедших в тест;

Северо-Восточный федеральный университет имени М.К. Аммосова

11) рекомендуемую автором стратегию расположения заданий в тесте.

Один из наиболее распространенных подходов к созданию краткой спецификации основан на сопряжении системы знаний и умений с процентным соотношением заданий по различным разделам или содержательным линиям проверяемой дисциплины в тесте (пункт 8 развернутой спецификации). Гипотетический пример реализации подобного сопряжения без привязки к какому-либо предмету приведен в табл. 3. В нее включен перечень знаний и умений по четырем содержательным разделам.

Подобного рода перечень знаний и умений, ориентированный на конкретную дисциплину, составляется всегда при планировании содержания теста. Естественно, что в приведенный перечень знаний и умений необходимо ввести некоторые пропорции сообразно тем акцентам, которые делаются преподавателем в учебном процессе.

Для заполнения таблицы было выбрано 60 заданий в качестве первоначальной длины теста. Доля каждого из разделов в процентах указана в табл. 3. Конечно, при заполнении таблицы в распределении заданий удается лишь приближенно отобразить пропорции разделов. К тому же числа вписываются не во все ячейки (например, ячейка на пересечении второго столбца и пятой строки), поскольку некоторые умения могут оказаться несовместимыми с содержанием отдельных разделов. Однако даже в таком приближенном видении общей раскладки заданий есть огромная польза

Таблица 3

Содержательный план теста

№ п/п	Планируемые к проверке знания и умения	Содержательные линии (разделы) дисциплины				Суммарное число заданий
		I (20 %)	II (10 %)	III (30 %)	IV (40 %)	
1	*A* (10 %)	1	1	2	2	6
2	*B* (20 %)	2	1	4	5	12
3	*C* (30 %)	4	2	5	7	18
4	*D* (30 %)	4	2	5	7	18
5	*E* (10 %)	1	—	2	3	6
Итого заданий		12	6	18	24	60

Примечание. *A* — знание понятий, определений, терминов; *B* — знание законов и формул; *C* — умение применять законы и формулы для решения задач; *D* — умение интерпретировать результаты на графиках и схемах; *E* — умение проводить оценочные суждения.

Северо-Восточный федеральный университет им. М.К. Аммосова

для разработки теста. Спецификация может также содержать процентное соотношение заданий, планируемое сообразно пропорциям разделов и видам предполагаемой деятельности испытуемого в процессе выполнения теста.

Естественно, что в процессе работы над тестом первоначальная раскладка заданий будет претерпевать всяческие изменения. Это объясняется тем, что не все задания окажутся удачными и уместными в той мере, как это считается на этапе планирования теста. Поэтому после коррекции теста необходима доработка спецификации для приведения ее в соответствие с окончательными пропорциями содержания теста.

Общие принципы отбора содержания теста. Общие принципы отбора содержания теста способствуют обеспечению высокой содержательной валидности теста [60]. Первый принцип — *принцип репрезентативности* — регламентирует процедуру отбора содержания таким образом, чтобы обеспечить оптимальную полноту и правильность пропорций содержания теста.

Второй принцип — *принцип значимости* — предписывает включать в тест наиболее значимые элементы содержания, относящиеся к опорным темам курса. Выделение опорных элементов требует структурирования содержания предмета перед его отбором в тест.

Третий принцип — *принцип системности* — предполагает подбор упорядоченных содержательных элементов, связанных между собой определенной иерархией и общей структурой знаний. При соблюдении этого принципа тест можно использовать не только для выявления уровня учебных достижений, но и для оценки качества структуры знаний учеников.

После планирования содержания теста и его отбора начинается наиболее ответственный этап создания предтестовых заданий, по окончанию которого проводится экспертное оценивание содержательной валидности теста.

Учреждение образования «Гомельский государственный университет имени Ф. Скорины»

6.3. Экспертиза качества содержания теста

Методика экспертизы. Оценка качества содержания теста должна проводиться по определенной методике независимыми экспертами, не участвовавшими в разработке теста. Как правило, число экспертов составляет не менее трех человек по каждому тесту. К экспертизе привлекаются наиболее опытные учителя, имеющие большой стаж работы с теми учащимися, для которых в конечном итоге предназначен тест.

Методика экспертизы качества содержания теста обычно включает три раздела, которые выстраиваются сообразно трем направлениям работы экспертов. Перед началом работы каждый эксперт

должен ознакомиться со спецификацией рецензируемого теста, пояснениями по его структуре, целями создания теста.

Направления работы экспертов. Работа экспертов по первому направлению заключается в анализе качества содержания отдельных заданий теста. При работе по этому направлению эксперт сам выполняет весь тест, сопоставляя полученные правильные ответы с ключом ответов автора (для заданий с выбором ответа) и анализируя оценочные категории, представленные автором (к заданиям со свободно конструируемым ответом). Особое внимание следует обратить на возможную неоднозначность, когда на месте планируемого единственного ответа могут возникнуть дополнительные, частично правильные ответы.

В процессе экспертизы определяется также уровень базовости содержания каждого задания, который, в определенной степени, является субъективной оценкой его трудности. Дифференциация заданий по уровням позволяет разделить их на три группы: базовые, повышенной трудности и наиболее трудные. В том случае, когда задание проверяет степень достижения требований ГОС на минимальном уровне, достаточном для выставления оценки «удовлетворительно», оно считается соответствующим первой группе базовых. Ко второй группе относятся задания, правильное выполнение которых позволяет выставить ученику хорошие и отличные оценки. И, наконец, в третью группу включаются задания, рассчитанные на проверку творческих аспектов подготовки школьников, выполнение которых позволяет судить не только об уровне, но и о качестве подготовки.

Эксперт оценивает уровень значимости содержания каждого задания теста и ожидаемое время его выполнения учащимся со средним уровнем подготовленности, а также находит логически некорректные задания, нуждающиеся в коррекции. Особенно важно в процессе экспертизы выявить случаи отсутствия четкого логического выделения одного предмета измерения. Для достижения логической четкости в каждом задании следует спрашивать о чем-либо одном. Попытки проверить сразу несколько аспектов подготовленности и ввести несколько величин в ответы, как правило, отрицательно сказываются на качестве заданий и всего теста.

При экспертизе также следует охарактеризовать качество формулировок заданий теста, отметить (в случае необходимости) лексическую избыточность формулировок, охарактеризовать качество представления графической информации и других компонентов заданий.

Второе направление работы эксперта связано с анализом качества содержания всего теста. При выполнении работ по этому направлению эксперт должен оценить степень отображения требований ГОС в тесте, правильность пропорций содержания теста и их соответствие спецификации теста. Простой оцен-

Электронный архив библиотеки МГУ имени А.А. Кулешова

ки полноты охвата требований недостаточно, необходима также уверенность в том, что задания теста затрагивает все важные аспекты предметной области и в правильной пропорции. Зачастую при разработке теста возможно смещение пропорций, так как тест легко перенасытить теми разделами содержания, по которым легче составить задания. Например, в рамках гуманитарного цикла дисциплин легко разрабатывать задания на выявление фактологических знаний, и потому эти задания нередко преобладают в тестах.

Третье направление работы эксперта рассчитано на подготовку обобщающих выводов и рекомендаций по улучшению содержания теста. В третьем разделе рецензии эксперт высказывает свое общее впечатление о содержании теста. Здесь должны быть представлены все сомнения эксперта и его рекомендации авторам по коррекции теста. Возможна оценка соотношения заданий, проверяющих знание теории предмета и его практики. Желательно особо отметить задания, предназначенные для проверки системы понятий, а также задания интегрального характера, позволяющие оценить умения учащихся обобщать знания по различным разделам предмета, задания с межпредметными компонентами и т.д.

Практические задания и вопросы для обсуждения

1. Перечислите этапы процедуры планирования содержания теста.
2. Что, по вашему мнению, является первичным: выбор формы заданий, определение длины теста или разработка спецификации теста?
3. Существует ли различие в подходах к планированию содержания нормативно-ориентированных и критериально-ориентированных тестов?
4. Многие авторы полагают, что для правильного отбора содержания теста вполне достаточно педагогического опыта, а разработка спецификации является излишней потерей времени при создании теста. Что вы думаете по этому поводу?
5. Какой принцип ориентирует разработчика на достижение полноты и значимости содержания теста?

Северо-Восточный федеральный университет им. М.К. Аммосова

ГЛАВА
7
ФОРМЫ ПРЕДТЕСТОВЫХ ЗАДАНИЙ

7.1. Классификация предтестовых заданий и общие требования к ним

Классификация предтестовых заданий. В соответствии с наиболее часто встречающейся в отечественной и зарубежной литературе классификацией предтестовых заданий выделяют:

1) задания с выбором, в которых учащиеся выбирают правильный ответ из данного набора ответов;

2) задания с конструируемым ответом, требующие при выполнении от ученика самостоятельного получения ответов;

3) задания на установление соответствия, выполнение которых связано с выявлением соответствия между элементами двух множеств;

4) задания на установление правильной последовательности, в которых от учащегося требуется указать порядок элементов, действий или процессов, перечисленных педагогом [26; 60; 72; 78].

Названные четыре формы тестовых заданий являются основными и наиболее распространенными, но абсолютизировать их нет оснований. Часто специфика содержания контролируемого предмета требует использования новых форм, более адекватных целям разработки тестов. Обычно такие инновации строятся на основе сочетания отдельных элементов основных форм.

Общие требования к предтестовым заданиям и процедурам их применения. Вне зависимости от формы предтестовые задания должны удовлетворять следующим общим требованиям:

– каждое предтестовое задание имеет свой порядковый номер, который может изменяться после статистической оценки трудности задания и выбора стратегии предъявления заданий теста;

– каждое предтестовое задание имеет эталон правильного ответа (эталон оценивания для заданий со свободно конструируемым ответом);

– в предтестовом задании все элементы располагаются на четко определенных местах, фиксированных в рамках выбранной формы;

– для предтестовых заданий разрабатывается стандартная инструкция по выполнению, которая не меняется в рамках каждой формы и предваряет формулировку заданий в тесте;

Полесский государственный университет им. А.Кулешова

– для каждого задания разрабатывается правило выставления дихотомической или политомической оценки, общее для всех заданий одной формы и сопровождающееся инструкцией по проверке и подсчету сырых (первичных) баллов по тесту.

Процесс тестовых измерений предельно стандартизируется, если:

– ни одному ученику не дается никаких преимуществ перед другими;

– заранее разработанная система подсчета баллов применяется ко всем ответам учеников без исключения;

– в тест включены задания одной формы либо разных форм с оптимальными весовыми коэффициентами, значения которых получены статистическим путем;

– тестирование различных групп испытуемых проводится в одинаковое время, в сходных условиях;

– группа тестируемых выравнена по мотивации;

– все испытуемые выполняют одни и те же задания.

Последнее условие не исключает возможности списывания, подсказки и других нарушений, поэтому обычно стараются создать несколько параллельных по содержанию и трудности вариантов одного теста. В целом выбор формы заданий и числа вариантов теста зависит от содержания контролируемого курса, целей контроля и требуемого уровня надежности измерений.

Северо-Восточный федеральный университет им. М.К. Аммосова

7.2. Предтестовые задания с выбором одного или нескольких правильных ответов

Основные элементы предтестовых заданий с выбором ответов. В предтестовых заданиях с выбором (закрытых заданиях) можно выделить основную часть, содержащую постановку проблемы, и готовые ответы, сформулированные преподавателем. Среди ответов правильным чаще всего бывает только один, хотя не исключаются и другие варианты с выбором нескольких правильных (в том числе и в разной степени) ответов.

Неправильные, но правдоподобные ответы называются *дистракторами* [60; 78]. Если в задании два ответа, один из которых является дистрактором, то вероятность случайного выбора правильного ответа путем угадывания равна 50 %. Число дистракторов подбирается таким образом, чтобы задание не было слишком громоздким. Вместе с тем стараются не допустить слишком большой вероятности угадывания правильного ответа. Поэтому чаще всего в заданиях бывает три или четыре дистрактора и один правильный ответ.

Задания с двумя и тремя ответами обычно используют для экспресс-диагностики, например, в автоматизированных конт-

рольно-обучающих программах для входа в обучающий модуль, при адаптивном тестировании или для самоконтроля, когда испытуемому необходимо оперативно выявить пробелы в собственных знаниях. Однако из-за высокой вероятности угадывания задания с двумя объектами не включают в итоговые тесты.

Правдоподобность дистракторов, их валидность. Если дистракторы сформулированы некорректно и не привлекательны даже для самых слабых испытуемых группы, они перестают выполнять свою функцию. На деле получается задание не с запланированным, а с меньшим числом ответов. В худшем случае, если в задании не работают все дистракторы, большинство учащихся выполнит задание верно, выбрав единственный правдоподобный правильный ответ. В идеале каждый дистрактор должен в равной мере привлекать всех испытуемых, выбирающих неправильный ответ.

Мера привлекательности дистракторов оценивается после первой апробации теста на репрезентативной выборке учащихся с помощью подсчета долей учеников, выбравших каждый из дистракторов в качестве правильного ответа. Например, если задание с четырьмя дистракторами и одним правильным ответом выполняли 25 учеников и 12 из них ответили неверно, то каждый из четырех дистракторов в качестве правильного ответа должны выбрать 3 ученика. Соответственно доля учащихся для каждого дистрактора будет равна 0,25. В этом случае все неправильные ответы к заданию равновероятно правдоподобны и, следовательно, сформулированы удачно. Конечно, точное равенство долей практически недостижимо, но тем не менее, создавая задания, к нему нужно стремиться.

Углубленный анализ частоты выбора каждого дистрактора учащимися с различным уровнем подготовленности позволяет сделать вывод о валидности неправильных ответов. Если дистрактор чаще привлекает слабых учащихся, выполнивших верно незначительное число заданий в тесте, то он считается валидным. Если же дистрактор кажется привлекательным в основном сильным ученикам, его валидность невысока, и задание подлежит переработке. В целом можно сказать, что тестовое задание считается «хорошо работающим», если знающие ученики выполняют его правильно, а незнающие выбирают любой из дистракторов с равной вероятностью.

Инструкции к предтестовым заданиям с выбором. Если тестирование проводится с помощью бланков, то задания с выбором одного правильного ответа сопровождаются следующей инструкцией: «Обведите номер (букву) правильного ответа». Задания с несколькими правильными ответами обычно используются в текущем контроле для проверки классификационных и фактуальных знаний, хотя встречаются случаи, когда специфика содержания дисциплины вынуждает включать их в итоговые тесты. Такие

Электронный архив библиотеки Северо-Восточного федерального университета имени М.К. Аммосова

задания сопровождаются специальной инструкцией, подчеркивающей необходимость выбора всех правильных ответов. Когда дистракторов намного меньше правильных ответов, их легко угадать. В подобной ситуации можно включить в число ответов только один неправильный, а учеников попросить выбрать один ошибочный ответ, если это не противоречит дидактическим целям контроля и допускается содержанием предмета.

Иногда, по замыслу автора, при разработке задания закладываются несколько правильных ответов, среди которых есть более и менее предпочтительные. В этом случае задание сопровождается следующий инструкцией: «Обведите номер наиболее правильного ответа». При компьютерной выдаче заданий инструкция может иметь такой вид: «Для ответа нажмите клавишу с номером (буквой) правильного ответа».

Обычно если все задания имеют одну форму, то инструкция приводится в начале теста. Если же в тест включены задания разных форм, они сопровождаются различными инструкциями. Легко представить, насколько осложнит выполнение теста чередование инструкций на выбор правильного и неправильного ответов. Невнимательные ученики могут не заметить изменения инструкции и выполнить часть заданий неправильно, даже в том случае, если они наверняка знают правильный ответ. Поэтому рекомендуется менять инструкцию как можно реже — ровно столько раз, сколько требует стратегия предъявления заданий теста.

Преимущества предтестовых заданий с выбором ответов. Задания с выбором имеют ряд преимуществ, связанных с быстротой их выполнения, простотой подсчета итоговых баллов учеников, возможностью автоматизации процедур проверки ответов учащихся и вытекающей отсюда минимизацией субъективного фактора при оценивании результатов выполнения теста. С их помощью можно более полно охватить содержание проверяемой учебной дисциплины и, следовательно, повысить содержательную валидность теста. Несомненным достоинством формы заданий с выбором является ее универсальность; она годится практически для любого предмета.

Недостатки предтестовых заданий с выбором ответов. К числу недостатков заданий с выбором следует отнести эффект угадывания, характерный для слабоподготовленных учеников при ответах на наиболее трудные задания теста. Хотя возможность угадывания действительно существует, тестологи с помощью различных методов научились избегать подобных ситуаций. Для этого вводятся специальные инструкции, ориентирующие испытуемых на пропуск незнакомого задания вместо ответа путем догадки. При подсчете баллов слабых учеников, полученных по наиболее трудным заданиям теста, добавляются специальные весовые коэффициенты, близкие к нулю, или увеличивается число заданий теста. Иногда

Федеральное государственное образовательное учреждение высшего профессионального образования

применяется специальная формула для коррекции индивидуальных баллов с поправкой на догадку [74]. Последний метод и поясняющая его формула приводятся в конце этой главы.

Отраженное в названии задания действие выбора вызывает необоснованную критику со стороны противников педагогических тестов, которые ошибочно отождествляют действия по выбору ответа с уровнем деятельности, необходимой для выполнении заданий теста. В заданиях, разработанных опытными авторами и не предназначенных для проверки фактологического материала, ученику для получения правильного ответа нередко приходится применять цепочку знаний и умений на продуктивном уровне деятельности. Ряд критиков подчеркивает отрицательную роль неправильных ответов, способствующих, по их мнению, запоминанию ошибочной информации при выполнении теста. С этой точкой зрения можно не соглашаться в том случае, если задания с выбором используются при итоговом контроле. Но в текущем контроле, выполняющем корректирующие и обучающие функции, лучше отдавать предпочтение заданиям с конструируемыми ответами.

Реальные, а не надуманные сложности возникают при использовании заданий с выбором для проверки умений продуктивного уровня, связанных с применением знаний учащимися в незнакомой ситуации, творческими аспектами подготовки, а также в случаях, когда требуется преобразование условий поставленной перед учащимся задачи. В таких ситуациях задания с выбором готовых ответов использовать чаще всего невозможно. В условиях же массового аттестационного тестирования, когда необходимо привлечь эффективные компьютеризованные технологии для подсчета баллов учеников и получить высокую объективность результатов педагогического измерения, достоинства заданий с выбором явно перевешивают недостатки. Поэтому эта форма нередко доминирует при разработке итоговых тестов.

Разработка предтестовых заданий с выбором ответов. Задания с выбором ответа должны удовлетворять ряду требований, выполнение которых позволяет повысить качество теста. Они подробно изложены в работах отечественных и зарубежных авторов [54; 67]. С ними обязательно следует ознакомиться тем, кто собирается создавать задания теста.

Общая рекомендация — не следует разрабатывать задания накануне тестирования. Готовить задания лучше заранее, с тем чтобы оставить достаточно времени на их переработку. Содержание каждого задания необходимо соотносить с измеряемой переменной и стараться обеспечивать планируемый уровень трудности.

После окончания работы над заданиями до привлечения экспертов следует самостоятельно провести анализ их соответствия спецификации теста, оценить значимость их содержания и ясность формулировок. При разработке заданий необходимо обеспе-

чить их относительную независимость, исключающую цепочную логику выполнения, когда ответ из одного задания служит условием выполнения другого задания теста. Не стоит использовать в тесте задания, приведенные в учебниках, в противном случае будет проверяться память учащихся, а не освоение умений. Тесты учебных достижений не должны содержать заданий-ловушек, присутствующих в психологических тестах.

Примеры предтестовых заданий с двумя и тремя ответами. В заданиях с двумя ответами проще всего подбирать дистракторы посредством отрицания того, что является верным. Не рекомендуется использовать вместо дистракторов слова «да», «нет», поскольку в противном случае будет довольно трудно сформулировать утверждения, на которые можно дать однозначный ответ. Например:

Задание 1

Если вычитаемое увеличили на 12 единиц, а разность также увеличилась на 15 единиц, то уменьшаемое

1) увеличилось

2) уменьшилось.

Задания с тремя ответами обычно используют в экспресс-диагностике в тех случаях, когда в силу специфики содержания недостаточно двух ответов. Иногда они появляются вследствие удаления «неработающих» дистракторов. В целом такие задания неудачны, поскольку они недостаточно кратки и в них высока вероятность угадывания правильного ответа. Например:

Задание 2

Импульс, поступающий по блуждающему нерву

1) учащает работу сердца

2) замедляет работу сердца

3) не влияет на деятельность сердца.

Предтестовые задания с четырьмя и пятью ответами. В большинстве тестов встречаются задания с четырьмя — пятью ответами, из которых один верный. При умелой разработке они могут быть достаточно краткими, и в них невысока вероятность угадывания правильного ответа (0,25 при четырех ответах и 0,20 при пяти). Например:

Задание 3

Кадеты считали главным методом решения основных проблем России

1) революционное восстание масс

2) политический террор

3) тактику давления на правительство через представительные органы, парламент

4) всеобщую политическую стачку.

Задание 4

Древние люди не могли охотиться на

Электронный архив Полоцкого государственного университета им. Ф. Скорины

1) мамонта
2) морскую корову
3) эластомерия
4) иностранцевия
5) глиптодонта.

Наиболее удачными можно считать задания, выполнение которых помимо традиционного длинного пути предполагает возможность довольно быстрого (3—4 с) ответа. Разумеется, быстрое решение по силам только учащемуся с четкой структурой знаний и твердыми навыками по проверяемому разделу. Слабо подготовленные ученики пойдут по традиционному пути и истратят на задание не 3—4 с, а положенные 1—2 мин.

Фасетные задания в тесте. Даже в условиях хорошо организованного процесса тестирования при наличии единственного варианта теста велика вероятность списывания, подсказок и других нежелательных моментов. Поэтому обычно разрабатывают 5—8 параллельных вариантов теста, для которых можно использовать фасетные задания. Под *фасетом* понимается форма, обеспечивающая представление нескольких вариантов одного и того же элемента содержания теста [60; 72].

Каждый испытуемый получает из фасета только один вариант задания. При этом все испытуемые группы выполняют однотипные задания, но с разными элементами фасета и, соответственно, с разными ответами. Таким образом, решаются одновременно две задачи: устраняется возможность списывания и обеспечивается параллельность вариантов тестов, предлагаемых различным ученикам. В приведенном далее примере содержатся два задания, одно из которых предназначено для выбора архитектурных комплексов окрестностей Москвы, а другое — Санкт-Петербурга:

Задание 5
К дворцовым комплексам окрестностей { Москвы / Санкт-Петербурга } относятся
1) Павловск, Ораниенбаум
2) Архангельское, Царицыно
3) Петергоф, Гатчина
4) Царское село, Стрельня.

Особенно легко и удобно создавать фасетные задания по естественному циклу дисциплин путем введения параметров в задания теста. Например, выбирая различные значения параметров a, b, c — коэффициентов квадратного уравнения, обеспечивающих неотрицательность его дискриминанта, — можно получить множество однотипных заданий.

Предтестовые задания с выбором нескольких правильных ответов. Задания с несколькими правильными ответами обычно стараются не включать в итоговые тесты, результаты которых ис-

Объединенный Федеральный университет им. М.А. Шолохова

пользуются для административно-управленческих решений в образовании. Появление частично правильных ответов учащихся, возникающих при выборе не всех запланированных верных ответов, приводит к снижению объективности и сопоставимости оценок, получаемых учениками по тесту.

В текущем контроле такие задания, наоборот, желательны, поскольку ученик должен не только найти правильные ответы, но и сам определить их число, что значительно разнообразит и усложняет задачу.

Предтестовые задания на выбор неправильного ответа. Ориентация учащихся на выбор неправильного ответа часто вызывает негативную реакцию у многих педагогов. Особенно неуместны задания на выбор неправильного ответа в тестах по русскому языку или по истории. Недопустимо, например, когда задание нацеливает ученика на неправильное написание слов либо на неверную оценку исторических событий. Однако если нужно проверить знание учеником определенных правил по технике безопасности, например во время проведения химических опытов, то выбор неправильного ответа становится просто находкой.

В случае когда бóльшая часть заданий в тесте ориентирована на выбор правильного ответа, заданий с противоположной инструкцией в тесте должно быть не более двух-трех.

Оценка результатов выполнения заданий, первичный, или сырой, балл. При подсчете результатов выполнения заданий с выбором одного правильного ответа обычно предпочитают дихотомическую оценку. За правильное выполнение задания испытуемый получает «1», а за неправильный ответ или пропуск — «0». Суммирование всех единиц позволяет вычислить индивидуальный (первичный, или сырой) балл испытуемого, который в случае дихотомической оценки равен количеству правильно выполненных заданий в тесте.

Если правильный ответ не один, то чаще всего используется политомическая оценка, которая пропорциональна числу правильно выбранных ответов.

Коррекция на догадку первичных тестовых баллов. Из-за эффекта угадывания ответов в заданиях с выбором сырые баллы стараются скорректировать путем ввода поправки на догадку. Формула коррекции баллов, полученных в результате выполнения заданий с κ-ответами, из которых только один верный, имеет следующий вид:

$$X_i' = X_i - \frac{W_i}{k-1}, \qquad (2)$$

где i — номер любого испытуемого группы; X_i' — скорректированный балл i-го испытуемого; X_i — тестовый балл до коррекции; W_i — число невыполненных (неправильно выполненных, пропу-

щенных, недостигнутых) заданий теста, а $X_i + W_i = N$, где N — число заданий в тесте.

Если в заданиях только один дистрактор и один верный ответ, то $k - 1 = 1$, поэтому коррекция баллов осуществляется довольно просто. Для каждого испытуемого вычисляется разность между числом правильно выполненных и невыполненных им заданий теста. Например, если в тесте из 60 заданий испытуемый выполнил правильно 50, а неправильно — 10, то скорректированный балл будет равен 50 – 10 = 40. Для более слабого ученика, выполнившего правильно всего 30 заданий из 60, балл после коррекции станет равен 30 – 30 = 0. Таким образом, балл сильного ученика уменьшился в результате коррекции весьма незначительно, всего на 10 единиц. Иначе обстоит дело с баллом учащегося, который выполнил правильно всего половину заданий теста. После коррекции он получит 0 баллов, так как в половине заданий с двумя ответами он вполне мог угадать правильный ответ.

Формула коррекции имеет определенные недостатки, снижающие точность тестовых измерений. Это связано с тем, что в основу ее построения положен ряд искусственных предположений, нередко не согласующихся с реальной процедурой выполнения теста. В частности далеко не в полной мере выполняется предположение о том, что все неправильные ответы являются следствием случайного угадывания. Столь же условно другое предположение об одинаковой вероятности выбора каждого ответа задания теста.

7.3. Предтестовые задания с конструируемым ответом

Общая характеристика. В заданиях с конструируемым ответом (заданиях на дополнение, открытых заданиях) готовые ответы не даются, их должен придумать или получить сам ученик. Задания с конструируемым ответом бывают двух видов. Первый предполагает получение учащимся строго регламентированных по содержанию и форме представления правильных ответов. Второй — задания со свободно конструируемыми ответами, в которых учащиеся составляют развернутые ответы, произвольные по длине и форме представления и содержащие полное решения задачи с пояснениями, микросочинения (эссе) и т.д. [60; 72].

Предтестовые задания с конструируемым регламентированным ответом. В заданиях первого вида заранее определяется то, что однозначно считается правильным ответом, и задается степень полноты его представления. Обычно ответ бывает достаточно кратким — в виде слова, числа, формулы, символа и т.д. Регламентированная краткость ответов накладывает определенные ограничения на сферу применения, поэтому задания первого вида в основном используются для оценки довольно узкого круга учебных

Полесский государственный университет

умений. Обычно с их помощью проверяются умения воспроизводить и применять знания в знакомой ситуации, а также выявляются уровень понимания изученного фактологического материала, знание понятийного аппарата и т. д.

Для разработки заданий с конструируемым регламентированным ответом необходимо мысленно сформулировать вопрос, затем записать четкий и краткий ответ, в котором на месте ключевого слова, символа или числа ставится прочерк. В силу однозначности правильного ответа проверка результатов выполнения заданий с конструируемым регламентированным ответом носит довольно объективный характер, ее осуществляют в компьютерной форме с последующей перепроверкой всех неправильных ответов учащихся экспертным путем. Ответы на задания приводятся на месте прочерка или заносятся учащимся в специальный бланк. Например:

Задание 6
Решите уравнение $\sqrt{2x+37} = x+1$ и занесите ответ в бланк.

Задание 7
Процесс, для которого теплоемкость постоянна, называется _____.

Задания с конструируемым регламентированным ответом малотехнологичны. В них нередко появляются частично правильные и правильные в разной степени ответы. Вписывая ответ на место прочерка, ученик может выбрать синонимы пропущенного запланированного разработчиком слова или изменить порядок следования элементов в пропущенной формуле, что значительно затрудняет автоматизированную проверку и оценку результатов.

Предтестовые задания со свободно конструируемым ответом. Задания второго вида не имеют ограничений по содержанию и форме представления ответов. За отпущенное время на специальных бланках для ответов ученик может писать что угодно и как угодно. Несомненно, такие условия выполнения во многом близки к традиционным письменным работам, поэтому задания со свободно конструируемым ответом воспринимаются положительно абсолютным большинством педагогов. Они интересны и разнообразны в содержательном плане. С их помощью можно выявить способы решения учебных задач, вычленить этапы мыслительного процесса и подвести итоги отдельных этапов, что особенно важно для анализа типичных ошибок учеников.

Разработка заданий со свободно конструируемым ответом может показаться неоправданно легкой. На самом деле сформулировать задание просто, а вот предложить эталон оптимального ответа вместе со стандартизованными правилами оценки результатов его выполнения достаточно сложно. Так, формулировка задания по истории с развернутым ответом достаточно кратка. Например:

Гомельский государственный университет имени Ф. Скорины

Задание 8

Назовите основные задачи, которые решались во внешней политике России в XVII в. (укажите не менее двух задач). Приведите примеры войн, походов и экспедиций XVII в., предпринимавшихся для решения этих задач (не менее трех примеров).

Но для того чтобы задание попало в тест, его автору необходимо стандартизовать процедуру проверки, а это объемная работа, вызывающая подчас много нареканий из-за неоднозначности результатов ее выполнения. В естественных науках предложить эталон выполнения вместе с оценочными критериями гораздо легче. Например:

Задание 9

При каких значениях x соответственные значения функций $f(x) = \log_2 x$ и $g(x) = \log_2(3 - x)$ будут отличаться меньше, чем на 1?

Далее идет решение, в соответствии с которым разрабатываются оценочные критерии, представленные в следующей таблице.

Баллы	Критерии оценки выполнения задания 9
2	Приведена верная последовательность шагов решения: 1) составление неравенства, содержащего модуль; 2) решение неравенства. Все преобразования и вычисления проведены правильно, получен верный ответ
1	Приведена верная последовательность шагов решения. При решении неравенства в шаге 2 допущена описка и/или негрубая вычислительная ошибка, не влияющая на правильность дальнейшего хода решения. В результате этой описки и/или ошибки может быть получен неверный ответ
0	Все случаи решения, не соответствующие указанным выше критериям выставления оценок в 1 или 2 балла

Федеральный учебно-методический университет им. М. Ломоносова

Проверка заданий с развернутыми ответами проводится экспертами в соответствии со стандартизированными инструкциями, содержащими эталон оптимального ответа с описывающими его характеристиками и признаками качества, как в приведенном примере. К эталону должны прилагаться оценочные категории для выставления политомической оценки, нуждающиеся в апробации и статистическом обосновании качества, поскольку среди них могут быть как не «работающие», так и снижающие дифференцирующий эффект теста.

Оценивание результатов выполнения предтестовых заданий со свободно конструируемыми ответами. Задания типа эссе можно оценивать в соответствии:

– с простыми схемами оценивания, когда при выборе критериев ориентируются на содержание ответов учащихся;

– с усложненными схемами оценивания, учитывающими при экспертизе содержание ответов, характеристики качества представления текста, его полноту и стиль или любые другие факторы, кажущиеся важными разработчику задания;

– с рейтинговым методом, предполагающим накопительную оценку, которая получается путем сложения отдельных оценок в соответствии с общим впечатлением экспертов от полного ответа на задание.

При любой схеме оценивания задания со свободно конструируемыми ответами нуждаются в политомической оценке, что иногда неоправданно завышает их общий вес в балле по тесту. Для того чтобы избежать такой ситуации и уменьшить влияние субъективного компонента, число критериев оценивания обычно стараются ограничить, например от «0» до «3» или от «0» до «4».

В целом задания с развернутыми ответами требуют значительных затрат преподавательского труда при проверке, так как экспертам приходится анализировать множество в разной степени правильных ответов и сравнивать их с эталоном. При этом не принимаются во внимание полнота, внешнее оформление ответов, орфографические ошибки и то, что не входит в критерии для выставления политомической оценки, хотя сейчас для проверки существуют специальные программы ПК [66]. Обычно в силу низкой технологичности такие задания занимают не более 10—15 % от всех заданий теста. Правда, в последнее время в связи с тенденцией к проверке творческих аспектов подготовленности учащихся число заданий с развернутыми ответами может составлять 50 % от общей длины теста.

Инструкции для предтестовых заданий с конструируемыми ответами. Для заданий с кратким регламентированным ответом, сформулированных в виде незаконченных утверждений и предъявляемых без специальных бланков для ответа, обычно используют инструкцию, состоящую из одного слова: «Дополните».

В тех случаях, когда для ответов к заданиям с кратким регламентированным ответом предлагаются специальные бланки, инструкция может иметь такой вид: «Ответы к заданиям запишите в бланке ответов справа от соответствующих номеров заданий. Каждую букву пишите в отдельной клеточке в соответствии с приведенными образцами на бланке ответов».

Инструкция для заданий со свободно конструируемым ответом обычно имеет произвольную форму. Главное — сказать достаточно, чтобы в максимальной степени облегчить и стандартизовать работу экспертов при проверке результатов тестирования, снизить влияние субъективных факторов и повысить надежность педагогических измерений.

Электронный архив библиотеки Белорусско-Российского университета

7.4. Предтестовые задания на установление соответствия

Общая характеристика. Задания на соответствие имеют специфический вид: под инструкцией располагаются элементы двух множеств, соответствие между которыми предлагается установить учащемуся [60; 72; 74]; слева обычно приводятся элементы задающего множества, содержащего постановку проблемы; справа — элементы, подлежащие выбору.

Соответствие между элементами двух столбцов может быть взаимно однозначным, когда каждому элементу слева соответствует только один элемент справа. Если число элементов в двух столбцах одинаковое, то для последнего элемента задающего множества выбора не произойдет, поэтому в множество для выбора стараются включить несколько дистракторов. Встречаются случаи, определяемые спецификой содержания предмета, когда для нескольких элементов левого столбца выбираются одни и те же элементы справа. Например, задание 10 построено удачно, а задание 11 — неудачно, поскольку оно включает шесть заданий с выбором одного ответа из двух.

Задание 10

Установите соответствие между датами и внешнеполитическими событиями. К каждому из 4 элементов (1, 2, 3, 4) подбирается один соответствующий элемент (а, б, в, г, д).

Даты	События
1) 1922 г.	а) подписание Рапалльского договора с Германией
2) 1924 г.	б) заключение договора о ненападении с Германией
3) 1934 г.	в) заключение Брестского мира с Германией
4) 1939 г.	г) вступление в Лигу Наций
	д) «полоса дипломатического признания» СССР

Ответы: 1____, 2____, 3____, 4____.

Задание 11

Установите соответствие между признаком животных и классом, для которого этот признак характерен.

Признак	Класс
1) оплодотворение внутреннее	а) земноводные
2) оплодотворение у большинства видов наружное	б) пресмыкающиеся
3) непрямое развитие	
4) размножение и развитие происходит на суше	
5) тонкая кожа, покрытая слизью	
6) яйца с большим запасом питательных веществ	

1	2	3	4	5	6

Северо-Восточный федеральный университет им. М.К. Аммосова

Задания на установление соответствия по алгоритму выполнения близки к заданиям с выбором ответа, поскольку ученик выбирает из числа ответов, предложенных преподавателем, правильный ответ. За рубежом задания на соответствие не выделяют в отдельный вид, а считают одной из разновидностей заданий с выбором ответов. Как и в заданиях с выбором ответов, наибольшие трудности при разработке связаны с подбором правдоподобных избыточных элементов в правом множестве. Мера правдоподобности каждого дистрактора устанавливается эмпирически. В итоговом контроле задания на соответствие малоэффективны в силу их громоздкости, не позволяющей охватить большой объем содержания.

Инструкция к заданиям на соответствие. К заданиям на соответствие прилагается стандартная инструкция, состоящая из двух слов: «Установите соответствие». Иногда используют развернутую инструкцию, особенно в тех случаях, когда есть отдельный бланк ответов. Например, инструкция может иметь вид: «Буквы, соответствующие заданным элементам, сначала запишите в таблицу, приведенную в тексте задания, а затем перенесите в бланк».

Оценивание результатов выполнения заданий на соответствие. Результаты выполнения заданий на соответствие оцениваются либо дихотомической, либо политомической оценкой. При дихотомическом оценивании за все правильно установленные соответствия в задании теста ставится «1». Если хотя бы одно соответствие неверно, то за частично правильно выполненное задание на соответствие учащийся получает «0».

При политомическом оценивании за каждое правильное соответствие ставится «1». В этом случае при проверке заданий на соответствие используется политомическая оценка, и общее количество баллов за задание равно числу правильно установленных соответствий.

Северо-Восточный федеральный университет им. М.К. Аммосова

7.5. Задания на установление правильной последовательности

Общая характеристика. Тестовые задания четвертой формы предназначены для оценки уровня владения последовательностью действий, процессов и т. п. Элементы, связанные с определенной задачей, приводятся в заданиях в произвольном порядке, а ученик должен установить правильный порядок предложенных элементов и указать его заданным способом в специально отведенном для этого месте [60; 72; 81].

Стандартная инструкция к заданиям четвертой формы имеет следующий вид: «Установите правильную последовательность». Иногда инструкцию включают в текст задания. Например:

Задание 12

Расположите имена русских полководцев в хронологической последовательности их деятельности. Запишите буквы, которыми обозначены имена, в правильной последовательности в приведенную в тексте задания таблицу, а затем перенесите их в бланк.

А) Дмитрий Пожарский
Б) Алексей Ермолов
В) Михаил Скобелев
Г) Алексей Орлов

Задание 13

Установите правильную последовательность механизма выдоха, расставив номера в отведенных местах :

☐ — спадение легких

☐ — торможение центра дыхательных мышц в спинном мозге

☐ — возбуждение центра выдоха в продолговатом мозге

☐ — расслабление диафрагмы и вспомогательных мышц

☐ — уменьшение грудной полости

Во многих случаях задания на установление правильной последовательности крайне нетехнологичны или неприменимы в силу специфики содержания предмета. Они громоздки и нередко допускают неоднозначную последовательность ответов.

Северо-Восточный федеральный университет им. М.К. Аммосова

7.6. Сравнительная характеристика форм предтестовых заданий

В процессе разработки теста у автора всегда возникает вопрос: «Остановиться на какой-нибудь одной форме заданий или совместить различные формы в одном тесте?» Выбор автора во многом должен определяться спецификой содержания учебной дисциплины, целями создания и применения теста. Немало в данном случае зависит от технологии проверки, сбора и обработки эмпирических данных, от технического и материального обеспечения процесса применения теста.

Организовать компьютеризованный сбор и анализ результатов выполнения теста легко, если тест состоит лишь из заданий с выбором ответов. Результаты выполнения заданий с конструируемыми ответами требуют ручной обработки и привлечения экспертов, а следовательно — дополнительных материальных затрат и времени на проверку. Обилие форм в тесте затрудняет работу

ученика и значительно усложняет статистическую обработку эмпирических результатов выполнения теста.

К сожалению, требование моноформности не всегда выполнимо, поскольку не все знания и умения ученика можно проверить с помощью моноформного теста. В связи с этим часто приходится идти на совмещение форм, что при прочих равных условиях всегда негативно отражается на точности измерений, обеспечиваемой тестом.

Выбор оптимальной формы предтестовых заданий обычно бывает связан со спецификой содержания теста. При этом приходится учитывать достоинства и недостатки каждой из форм (табл. 4) и принимать определенное компромиссное решение в процессе такого выбора.

Таблица 4

Достоинства и недостатки различных форм предтестовых заданий

Форма предтестовых заданий	Достоинства	Недостатки
Задания с двумя ответами	Благодаря краткости позволяют охватить большой объем материала, легко разрабатываются (только один дистрактор), результаты выполнения обрабатываются быстро с высокой объективностью	Стимулируют механическое запоминание, поощряют угадывание, требуют увеличения количества заданий и, соответственно, времени тестирования для компенсации эффекта угадывания
Задания с выбором из четырех-пяти ответов	Годятся для самых различных предметов, в силу краткости формулировок в тесте можно охватить большой объем содержания, обеспечивают возможность автоматизированной проверки и высокую объективность оценок учащихся, позволяют провести развернутый статистический анализ своих характеристик, скорректировать их и значительно повысить надежность педагогических измерений	Требуют значительной работы авторов при подборе дистракторов, не годятся для проверки продуктивного уровня деятельности и когнитивных умений
Задания с конструируемыми регламентированными ответами	Просты в разработке, исключено угадывание, частично годятся для автоматизированной проверки	Проверяют в основном знание фактологического материала или понятийного аппарата, иногда (в гуманитарных предметах)

Сибирский государственный федеральный университет

Окончание табл. 4

Форма предтестовых заданий	Достоинства	Недостатки
		приводят к неоднозначным правильным и частично правильным ответам
Задания со свободно конструируемыми ответами	Позволяют оценивать сложные учебные достижения, в том числе творческий уровень деятельности, легко формулируются, исключают угадывание	Требуют длительной дорогостоящей процедуры проверки, значительного времени выполнения, не позволяют охватить значительный объем содержания предмета, снижают надежность педагогических измерений
Задания на соответствие	Просты в разработке, идеально подходят для оценивания ассоциативных знаний и проведения текущего контроля, уменьшают эффект угадывания	В основном используются лишь для проверки репродуктивного уровня деятельности и алгоритмических умений, громоздки по форме представления

Северо-Восточный федеральный университет им. М.К.Аммосова

Практические задания

1. Обведите номер правильного ответа.

Тестовые задания с двумя ответами эффективнее использовать в контроле:

1) текущем;
2) тематическом;
3) итоговом;
4) входном.

2. Обведите номер правильного ответа.

Вероятность угадывания номера правильного ответа в задании с пятью дистракторами равна:

1) 0,2;
2) 0,166666;
3) 0,5;
4) 0,666666.

ГЛАВА

8

КОМПЬЮТЕРНОЕ ТЕСТИРОВАНИЕ В ОБРАЗОВАНИИ

8.1. Специфика компьютерного тестирования и его формы

Общие представления о компьютерном тестировании. С начала XXI в. в образовании при проведении тестирования стали широко применяться компьютеры. В педагогических инновациях появилось отдельное направление — компьютерное тестирование, при котором предъявление тестов, оценивание результатов учащихся и выдача им результатов осуществляется с помощью ПК.

Этап генерации тестов технологически может протекать по-разному, в том числе путем ввода в компьютер бланковых тестов. На сегодняшний день по компьютерному тестированию имеются многочисленные публикации, разработаны программно-инструментальные средства для генерации и предъявления тестов [2; 31; 39; 76; 79].

Когда необходимо обращаться к компьютерному тестированию. Хотя компьютерное тестирование значительно облегчает работу учителя при предъявлении и оценивании результатов выполнения тестов, его распространение во многом не более чем дань моде, все негативные последствия которого до сих пор не выявлены в полной мере. Выбор компьютерного формата экзамена должен основываться на более важных и обоснованных предпосылках, чем просто увлечение инновациями, поскольку он порождает множество проблем и ставит учащихся в неравные условия. Обращаться к компьютерному тестированию следует в тех случаях, когда есть настоятельная потребность в отказе от традиционных бланковых тестов.

Например, компьютерное тестирование необходимо при проведении ЕГЭ в труднодоступных районах России. Сбор выпускников школ отдаленных районов в обозначенное время проведения ЕГЭ становится настолько сложным и дорогостоящим мероприятием, что обойтись без компьютерного тестирования и современных средств коммуникации просто невозможно. Компьютерное тестирование целесообразно также применять при проведении экзаменов для детей с ограниченными возможностями, имеющих серьезные нарушения зрения или слуха. С помощью ПК можно использовать большие по размерам шрифты, аудиозаписи, до-

Северо-Восточный федеральный университет им. М.К. Аммосова

полнительные устройства для ввода данных тестирования и другие приспособления, компенсирующие на экзаменах потенциальное отставание детей с ограниченными возможностями.

Формы осуществления компьютерного тестирования. Компьютерное тестирование может проводиться в различных формах, различающихся по технологии объединения заданий в тест (рис. 17). Часть из них пока не получили специального названия в литературе по тестовой проблематике.

Первая форма — самая простая. Готовый тест, стандартизованный или предназначенный для текущего контроля, вводится в специальную оболочку, функции которой могут различаться по степени полноты. Обычно при итоговом тестировании оболочка позволяет предъявлять задания на экране, оценивать результаты их выполнения, формировать матрицу результатов тестирования, обрабатывать ее и шкалировать первичные баллы испытуемых путем перевода в одну из стандартных шкал для выдачи каждому испытуемому тестового балла и протокола его оценок по заданиям теста.

Вторая форма компьютерного тестирования предполагает автоматизированную генерацию вариантов теста, осуществляемую с помощью инструментальных средств. Варианты создаются перед экзаменом или непосредственно во время его проведения из банка калиброванных тестовых заданий с устойчивыми статистическими характеристиками. Калибровка достигается благодаря длительной предварительной работе по формированию банка, параметры заданий которого получают на репрезентативной выборке учащихся, как правило, на протяжении 3—4 лет с помощью бланковых тестов. Содержательная валидность и параллельность вариантов обеспечиваются за счет строго регламентированного отбора заданий каждого варианта в соответствии со спецификацией теста.

Третья форма — компьютерное адаптивное тестирование — базируется на специальных адаптивных тестах. В основе идей

Рис. 17. Формы компьютерного тестирования

Федеральный институт развития образования

адаптивности лежат соображения о том, что учащемуся бесполезно давать задания теста, которые он выполнит наверняка правильно без малейших затруднений или гарантированно не справится с ними в силу высокой трудности. Поэтому предлагается оптимизировать трудность заданий, адаптируя ее к уровню подготовленности каждого испытуемого, и сократить за счет исключения части заданий длину теста.

Достоинства и недостатки компьютерного тестирования. Компьютерное тестирование имеет определенные преимущества по сравнению с традиционным бланковым тестированием, которые проявляются особенно заметно при массовых проверках, например при проведении национальных экзаменов типа ЕГЭ. Предъявление вариантов теста на компьютере позволяет сэкономить средства, расходуемые обычно на печать и транспортировку бланковых тестов.

Благодаря компьютерному тестированию можно повысить информационную безопасность и предотвратить рассекречивание теста за счет высокой скорости передачи информации и специальной защиты электронных файлов. Упрощается также процедура подсчета результирующих баллов в тех случаях, когда тест содержит только задания с выбором ответов.

Другие преимущества компьютерного тестирования проявляются в текущем контроле, при самоконтроле и самоподготовке учащихся; благодаря компьютеру можно незамедлительно выдать тестовый балл и принять неотложные меры по коррекции усвоения нового материала на основе анализа протоколов по результатам выполнения корректирующих и диагностических тестов. Возможности педагогического контроля при компьютерном тестировании значительно увеличиваются за счет расширения спектра измеряемых умений и навыков в инновационных типах тестовых заданий, использующих многообразные возможности компьютера при включении аудио- и видеофайлов, интерактивности, динамической постановки проблем с помощью мультимедийных средств и др.

Благодаря компьютерному тестированию повышаются информационные возможности процесса контроля, появляется возможность сбора дополнительных данных о динамике прохождения теста отдельными учащимися и для осуществления дифференциации пропущенных и не достигнутых заданий теста.

Помимо неоспоримых достоинств компьютерное тестирование имеет ряд недостатков, которые представлены на рис. 18.

Типичные психологические и эмоциональные реакции учащихся на компьютерное тестирование. Обычно психологические и эмоциональные реакции учащихся на компьютерное тестирование носят позитивный характер. Учащимся нравится незамедлительная выдача тестовых баллов, протокола тестирования с результатами

Северо-Восточный федеральный университет им. М.К. Аммосова

Рис. 18. Проблемы, возникающие при компьютерном тестировании

по каждому заданию, а также сам инновационный характер контроля в том случае, когда привлекаются современные гипермедийные технологии для выдачи теста. Динамическое мультимедийное сопровождение заданий на компьютере, объединенное программными средствами для представления в интерактивном режиме, по мнению учащихся, обеспечивает более точную оценку знаний и умений, сильнее мотивирует к выполнению заданий по сравнению с бланковыми тестами. Удобно также то, что вместо заполнения специальных форм для ответов можно просто выбрать ответ мышью. Если тестирование проходит в адаптивном режиме, то сокращаются время проведения экзамена и длина теста.

Негативные реакции обычно вызывают различные ограничения, которые иногда накладываются при выдаче заданий в компьютерном тестировании. Например, фиксируется либо порядок предъявления заданий, либо максимально возможное время выполнения каждого задания, после истечения которого независимо от желания испытуемого появляется следующее задание теста. В адаптивном тестировании учащиеся бывают недовольны тем, что они не имеют возможности пропустить очередное задание, просмотреть весь тест до начала работы над ним и изменить ответы на предыдущие задания. Иногда школьники возражают против компьютерного тестирования из-за трудностей, которые возникают при выполнении и записи математических вычислений и т.д.

Воздействие на выполнение теста предшествующего уровня компьютерного опыта. Результаты зарубежных исследований показали, что опыт работы на компьютерах, имеющийся у школьников, во многих случаях значительно влияет на валидность результатов выполнения теста. Если в тест включены задания без инноваций с выбором ответов, то влияние опыта работы с компьютером на результаты тестирования незначительно, поскольку от учащихся в таких заданиях не требуется никаких сложных действий при выполнении теста. При предъявлении на экране инновационных типов заданий, широко использующих средства компьютерной графики и другие новшества, влияние предшествующего ком-

пьютерного опыта на тестовый балл становится очень значительным. Таким образом, при компьютерном тестировании необходимо учитывать уровень компьютерного опыта учащихся, для которых предназначается тест.

Для снижения влияния опыта работы с компьютером на тестовые баллы рекомендуется включать в оболочки для компьютерного тестирования специальные инструкции и тренировочные упражнения для каждой инновационной формы заданий. Необходимо также предварительно ознакомить учащихся с интерфейсом, провести репетиционное тестирование и выделить в самостоятельные группы учащихся, не имеющих достаточного опыта работы с ПК, для того чтобы дополнительно обучить их или дать им бланковый тест.

Влияние интерфейса пользователя на результаты компьютерного тестирования. Интерфейс пользователя включает доступные учащемуся функции и возможности движения по заданиям теста, элементы размещения информации на экране, а также общий визуальный стиль представления информации. Хороший интерфейс пользователя должен обладать ясностью и корректностью логической последовательности взаимодействия с экзаменуемым, отражая общие принципы дизайна графической информации. Чем более продуман интерфейс, тем меньше внимания учащийся на него обращает, сосредоточивая все свои усилия на выполнении заданий теста.

Северо-Восточный федеральный университет им. М.К. Аммосова

8.2. Инновационные формы тестовых заданий при компьютерном тестировании

Цели разработки инновационных заданий в компьютерном тестировании. Инновационные задания, использующие возможности компьютерного тестирования, на сегодняшний день являются наиболее перспективным направлением развития автоматизации педагогических измерений. Основной причиной этого является большой потенциал инновационных заданий для повышения информативности педагогических измерений и увеличения содержательной валидности тестов.

Основная цель разработки инновационных заданий для компьютерного тестирования состоит в оценивании тех когнитивных умений, функциональной грамотности и коммуникативных умений, которые остаются не выявленными при традиционном контроле или использовании бланковых тестов.

Предметом оценивания при инновациях может быть уровень аналитико-синтетической деятельности обучаемого, скорость обобщения новой информации, гибкость мыслительного процесса и многие другие показатели умственной деятельности, сформиро-

вавшиеся в процессе обучения и не поддающиеся оцениванию с помощью обычных тестов.

Возможности инновационных заданий в компьютерном тестировании. В использовании инновационных заданий можно выделить два аспекта: дидактический и психолого-педагогический. Первый предполагает развернутую содержательную интерпретацию результатов тестирования в контексте освоенных на момент предъявления теста когнитивных, учебных и общеучебных умений, а второй позволяет оценить уровень развития мыслительных процессов у учащегося и выявить особенности усвоения им новых знаний. Большинство инновационных заданий, разработанных к настоящему времени, обеспечивают совершенствование измерений в обоих направлениях. Таким образом, инновационные задания позволяют расширить возможности самого педагогического измерения за счет получения результатов в новых, недоступных ранее направлениях оценивания качества подготовленности учащихся. Например, для оценивания уровня сформированности функциональной грамотности экзаменующимся можно предложить отрывок текста, в котором есть ошибки, а затем попросить идентифицировать их и исправить путем перепечатывания разделов текста.

Инновационные задания способствуют сокращению влияния случайного угадывания за счет увеличения числа возможных ответов без нарастания громоздкости заданий теста. Например, при оценивании понимания прочитанного текста можно попросить учащегося выбрать ключевое предложение в тексте и указать на него щелчком мыши. Таким образом, каждое предложение в текстовом отрывке становится опцией для выбора вместо 4—5 ответов в традиционных заданиях с готовыми ответами. Для совершенствования формы заданий используют сложный рисунок, динамические элементы, включая изображения, мультипликацию или видео; тем самым сокращается время чтения условия. Расширение возможностей тестирования происходит при включении звука, что позволяет вести диалог с учащимся, оценивать фонетические особенности его произношения при тестировании по иностранному языку, проверять правильность интерпретации различных звуков.

Основные направления инноваций при разработке заданий. Инновации при разработке заданий для компьютерного тестирования охватывают пять связанных между собой направлений. К ним относятся: форма задания, действия испытуемого при ответе, уровень использования мультимедийных технологий, уровень интерактивности и методика подсчета баллов.

Нововведения в форме задания включают визуальный и звуковой информационные ряды или их сочетание. Визуальная информация может носить реалистический (фото, кино) и синтезированный (рисунок, анимация) характер. Тип информации в сочетании с тестовой формой определяет формат ответа, выбираемо-

Учреждение образования «Гомельский государственный университет имени Ф. Скорины»

го или создаваемого экзаменующимся. При использовании фотографий или рисунков информация, содержащаяся в тестовых заданиях, носит статический характер. Кино, отражающее реальный мир, и анимация вносят динамику в выполнение теста.

Действия учащегося при ответе на задания зависят от тех инновационных средств, которые включены в тест. При включении в задания звуковой информации, предполагающей голосовой ответ учащегося, для ответа используются клавиатура, мышь или микрофон. Значительное место при ответах отводится интерактивным процессам. Интерактивный режим работы учащихся при компьютерном тестировании означает поочередную выдачу аудиовизуальной информации, при которой каждое новое высказывание со стороны учащегося или компьютера строится с учетом предыдущей информации с той и другой стороны. При организации интерактивного режима в компьютерном тестировании используется в основном экранное меню, в котором учащийся для ответа на тестовые задания выбирает, создает или перемещает объекты — компоненты ответа. Реже в интерактивном режиме применяют голосовой ввод ответа.

В целом уровень интерактивности, обеспеченный в компьютерном тестировании, характеризует степень, в которой определенная форма задания реагирует или отвечает на ввод информации со стороны экзаменующегося. Этот уровень варьируется от простейшего случая, когда совершается один шаг, до сложных, многошаговых заданий с разветвлением после каждого очередного ответа ученика.

Сравнительная характеристика инновационных форм заданий при компьютерном тестировании для различных целей совершенствования педагогического измерения приведена в табл. 5.

Проблемы, возникающие при использовании заданий повышенной трудности в компьютерном тестировании. Задания повышенной трудности всегда требуют больше времени для ответов вне зависимости от того, предъявляются ли они с помощью компьютерного моделирования виртуальной реальности, имеют ли форму лабораторной работы, эссе или используют мультимедийные технологии. Из-за временны́х затрат число сложных заданий должно быть незначительно — не более 10—15 %, в отдельных случаях — 20—25 %. Многообразие звуковых и зрительных образов в компьютерном тестировании приводит к возникновению у школьников усталости, поэтому при включении в тест даже небольшого количества трудных инновационных заданий приходится значительно уменьшать длину теста, что негативно сказывается на содержательной валидности, надежности и информационной безопасности педагогического измерения.

Несмотря на преимущества инновационных форм заданий, предъявляемых с помощью компьютера, к ним нужно относиться

Таблица 5

Сравнительная характеристика инновационных форм заданий при компьютерном тестировании

Цель совершенствования педагогического измерения	Характеристика формы ответа	Основные направления инноваций	Характеристика трудности задания
Снизить эффект угадывания	Ответ числовой (или текстовый), конструируемый учащимся, ввод с клавиатуры или голосовой через микрофон	Использование формы задания с конструируемым ответом	Обычно высокая
Повысить содержательную валидность	Ответ выбирается мышью на графическом изображении, используется обычное меню или гипертекст	Использование аудиовизуального ряда. Включение мультимедиа без интерактивности	Низкая или средняя
Обеспечить повышение конструктной и содержательной валидности	Ответ выбирается мышью на графическом изображении, запрашивается дополнительная информация, используется гипертекст	Использование мультимедиа для моделирования естественной окружающей среды и действий пользователя в ней. Представление объектов с помощью анимации вне режима интерактивности	Средняя или высокая
Расширить возможность измерения интеллектуальных умений, когнитивных навыков	Ответ осуществляется перемещением объектов на экране и конструируется учащимся, используется клавиатура, левая и правая кнопки мыши. Возможен интерактив	Использование формы задания с конструируемым ответом и интерактивом простейшего уровня	Средняя или высокая
Очеспечить возможность оценивания творческих и практических умений	При конструировании ответа учащимся обязательно используется двухступенчатый или многоступенчатый разветвляющийся интерак-	Использование формы задания с конструируемым ответом и интерактивом сложного уровня	Средняя или высокая

Северо-Восточный федеральный университет им. М.К. Аммосова

Окончание табл. 5

Цель совершенствования педагогического измерения	Характеристика формы ответа	Основные направления инноваций	Характеристика трудности задания
	тивный переход к различным этапам выполнения задания		
Обеспечить повышение конструктной и содержательной валидности; расширить охват содержания; реализовать возможность измерения коммуникативных и интеллектуальных умений, когнитивных навыков	Ответ моделируется учащимся пошагово с использованием многоступенчатого разветвляющегося интерактивного перехода к различным этапам выполнения задания и виртуальной реальности	Действия испытуемого при ответе	Высокая

Северо-Восточный федеральный университет им. М.К. Аммосова

с осторожностью, тщательно анализировать их адекватность целям измерения и уместность в тесте. Обычно инновационные задания высокой трудности выделяют в отдельный блок и помещают в конце теста. Их выполнение не должно отнимать времени у наиболее слабых учащихся, которые, скорее всего, не дойдут до конца теста.

Подсчет баллов учащихся. Если в компьютерном тестировании не используются мультимедийные и интерактивные технологии, то подсчет первичных баллов учащихся проводится традиционно путем суммирования оценок по отдельным заданиям. Привлечение мультимедийных технологий приводит к многомерности результатов выполнения теста, поскольку оценивание целого спектра творческих, коммуникативных, общепредметных и других умений с помощью инновационных форм заданий всегда связано с несколькими переменными измерения. Появление интерактивности еще больше усложняет процедуру подсчета баллов учащихся, она становится зависимой от ответа экзаменующегося на каждом шаге выполнения заданий теста и требует политомических оценок.

Проверка результатов выполнения заданий с конструируемым регламентированным ответом осуществляется путем сравнения ответа экзаменующегося с эталоном, хранящимся в памяти ком-

пьютера, и включает различные синонимы правильного ответа с приемлемыми орфографическими ошибками.

Намного сложнее автоматизированный подсчет баллов в заданиях со свободно конструируемым ответом (типа эссе) в гуманитарных дисциплинах. На сегодняшний день зарубежными тестологами разработаны специальные программы для автоматизированной проверки эссе. Критерии оценивания в этих программах довольно разнообразны: от рассмотрения поверхностных характеристик эссе типа длины и степени полноты ответа до сложных случаев анализа с использованием достижений компьютерной лингвистики. Обычно все эти различные автоматизированные программы подсчета баллов требуют участия экспертов только на момент начала работы, когда квалифицированным педагогам необходимо «обучить» компьютерную программу оцениванию любых развернутых ответов.

8.3. Тесты фиксированной длины, компьютерная генерация параллельных вариантов теста

Основные компоненты процесса автоматизированной компоновки теста для компьютерного предъявления. Процесс автоматизированной компоновки теста в том случае, когда он происходит заранее и не в адаптивном режиме, включает сборку (генерацию) параллельных вариантов, выбор правила подсчета баллов тестируемых учащихся и коррекцию вариантов для выполнения требований теории педагогических измерений.

Неизбежные различия по трудности вариантов, возникающие вследствие существования ошибок измерения, устраняются после тестирования путем выравнивания шкал, получаемых при подсчете тестовых баллов по отдельным вариантам теста. К числу сопутствующих вопросов, решение которых также необходимо при автоматизированной компоновке теста, относится работа по наполнению банка тестовых заданий и оцениванию информационной безопасности тестирования.

Компьютерная генерация параллельных вариантов теста фиксированной длины. Автоматизированная сборка теста с фиксированным числом заданий предполагает наличие установленной длины теста, его спецификации и банка калиброванных заданий. В работоспособный банк, поддерживающий генерацию многовариантного теста, должны входить фреймы заданий различной трудности по каждому содержательному элементу с устойчивыми оценками параметров. С помощью специального программно-инструментального обеспечения получается аналог традиционного бланкового теста, готовый к предъявлению спустя несколько минут от

Гомельский областной институт переподготовки и повышения квалификации

начала генерации и обеспечивающий высокое качество педагогических измерений.

Метод автоматизированной компоновки теста для компьютерного предъявления в режиме offline (без использования локальных компьютерных сетей или Интернета) или в режиме online (с использованием локальных компьютерных сетей или Интернета) называют *автоматизированным тестовым дизайном*. Целью дизайна является формирование вариантов теста, удовлетворяющих целому ряду условий, к которым относятся: число заданий, структура содержания, частота выбора заданий в варианты, а также ряд требований, обеспечивающих генерацию параллельных вариантов теста.

Технология компоновки вариантов должна поддерживать систематический контроль за частотой включения каждого задания из банка в тест. Количество одинаковых заданий в параллельных вариантах, используемых для выравнивания шкал по вариантам, не должно превышать 15—20 %. Для контроля частоты включения задания в варианты в качестве ограничения вводится максимально возможный процент выбора каждого задания из банка. При его достижении задание перестает использоваться в дальнейших процедурах генерации теста.

Обычно многочисленные параллельные или квазипараллельные варианты теста создаются в режиме offline для последующего предъявления в режиме online, в том числе при интерактивном взаимодействии с обучающимися [19]. Для расширения коммуникативных возможностей компьютерного контроля в real time рекомендуется использование адаптивного тестирования, обеспечивающего пошаговую оптимизацию подбора трудности заданий при генерации адаптивного теста (см. раздел 8.4).

8.4. Компьютерное адаптивное тестирование

Адаптивное тестирование и его возможности. Появление адаптивного тестирования было вызвано стремлением к повышению эффективности педагогических измерений, которая, как правило, связывалась с уменьшением числа заданий, времени, стоимости тестирования, а также с повышением точности оценок учащихся. В основе адаптивного подхода лежит индивидуализация процедуры отбора заданий теста, которая за счет оптимизации трудности заданий применительно к уровню подготовленности обучаемых обеспечивает генерацию эффективных тестов [59; 62; 71].

Оптимизация трудности заданий обычно проводится пошагово. Если учащийся выполняет задание верно, то затем ему дается более трудное задание. При неправильном выполнении задания

Северо-Восточный федеральный университет им. М.К. Аммосова

совершается отход назад к более легким заданиям банка. При невыполнении трех заданий подряд процесс останавливается и специальными методами (чаще всего с помощью теории IRT) определяется балл учащегося за выполненные задания по сформированному специально для него адаптивному тесту. Таким образом, в компьютерном адаптивном предъявлении число тестовых заданий и их трудность индивидуально подбираются для каждого экзаменующегося на основании его ответов, а индивидуальная совокупность заданий образует адаптивный тест. Адаптивные тесты в группе испытуемых состоят в основном из разных заданий и различаются по количеству и трудности заданий тем сильнее, чем больше разброс среди испытуемых тестируемой группы по подготовленности.

Получить одновременный прирост эффективности измерений по всем критериям невозможно, поэтому обычно при организации адаптивного тестирования на первый план выходит один, в лучшем случае, два критерия. Например, в одних случаях при экспресс-диагностике в адаптивном режиме наибольшее внимание уделяется минимизации времени испытания и количеству предъявляемых заданий, а вопросы точности оценок отходят на второй план. В других случаях приоритетной может быть точность измерения и тестирование каждого испытуемого продолжается до тех пор, пока не достигается запланированная минимальная ошибка измерения.

На длине адаптивного теста существенно сказывается качество структуры знаний учащихся. Обычно испытуемые с четкой структурой знаний выполняют задания нарастающей трудности, уточняя с каждым очередным верно выполненным заданием оценку подготовленности. Они выполняют небольшое число заданий адаптивного теста и быстро доходят до порога своей компетентности. Учащиеся с нечеткой структурой знаний, у которых чередуются верные и неверные ответы, получают колеблющиеся по трудности задания. Процесс тестирования затягивается, поскольку при скачкообразном изменении трудности заданий не происходит пошагового нарастания точности измерения и число заданий, адаптированных по трудности, нередко оказывается даже бóльшим, чем в обычном, традиционном тесте.

Преимущества адаптивного тестирования. К числу важных преимуществ компьютеризованного адаптивного тестирования можно отнести:

– высокую эффективность;

– высокий уровень секретности;

– индивидуализацию темпа выполнения теста;

– высокий уровень мотивации к тестированию у наиболее слабых обучающихся за счет исключения из процесса предъявления излишне трудных заданий;

Северо-Восточный федеральный университет им. М.К. Аммосова

– сообщение результата в интервальной шкале тестовых баллов каждому испытуемому незамедлительно, сразу после окончания его работы над индивидуально подобранным набором заданий в адаптивном тесте.

Стратегии адаптивного тестирования. Стратегии предъявления тестовых заданий в адаптивном тестировании можно разделить на двухшаговые и многошаговые, сообразно которым используется различная технология формирования адаптивных тестов. Двухшаговая стратегия предполагает наличие двух этапов. На первом этапе всем испытуемым выдается одинаковый входной тест, цель которого — осуществление предварительной дифференциации учащихся вдоль оси переменной измерения. По результатам дифференциации на втором этапе организуется адаптивный режим и строятся адаптивные тесты.

В результате развития теории IRT, обеспечивающей единую интервальную шкалу для оценок параметров испытуемых и трудности заданий теста, появилась возможность по-новому осуществить оптимизацию процедуры отбора заданий для моделирования эффективных адаптивных тестов. Стали развиваться многошаговые стратегии адаптивного тестирования, в рамках которых в процессе выполнения наборов заданий каждый испытуемый движется по своей индивидуальной траектории.

Многошаговые стратегии адаптивного тестирования подразделяются на *фиксировано-ветвящиеся* и *варьирующе-ветвящиеся* в зависимости от того, как конструируются многошаговые адаптивные тесты. Если один и тот же набор заданий с их фиксированным расположением на оси трудности используется для всех испытуемых, но каждый учащийся движется по набору заданий индивидуальным путем в зависимости от результатов выполнения очередного задания, то стратегия адаптивного тестирования является *фиксировано-ветвящейся*.

Задания по трудности в наборе заданий обычно располагают на равном расстоянии друг от друга или выбирают убывающий шаг сообразно нарастанию трудности, что позволяет подстроить темп тестирования под испытуемого, поскольку по мере выполнения заданий у него нарастает утомление и снижается мотивация к выполнению заданий теста.

Варьирующе-ветвящаяся стратегия адаптивного тестирования предполагает отбор заданий непосредственно из банка по определенным алгоритмам, которые прогнозируют оптимальную трудность последующего задания по результатам выполнения испытуемым предыдущего задания адаптивного теста. Таким образом, шаг за шагом из отдельных заданий получается адаптивный тест. В нем варьирует не только трудность, но и шаг, определяемый разностью трудностей двух соседних заданий адаптивного теста. Отличительной особенностью варьирующей ветвящейся стратегии адап-

Электронный архив библиотеки МГПУ имени А.А. Кулешова

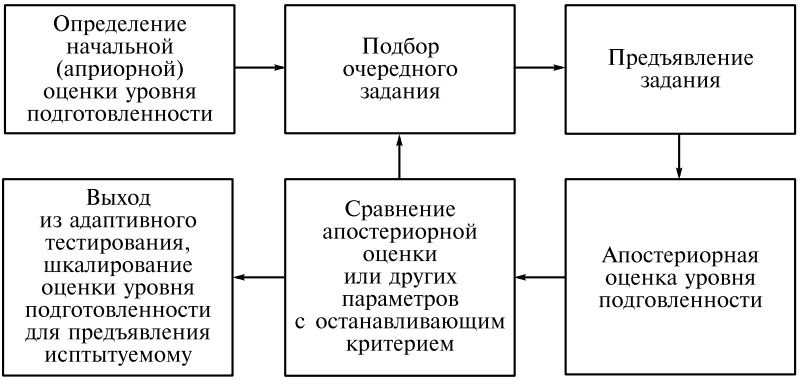

Рис. 19. Алгоритм варьирующего многошагового тестирования

тивного тестирования является пошаговая переоценка уровня подготовленности испытуемого, предпринимаемая после каждого выполнения очередного задания теста.

Алгоритм, реализующий варьирующую стратегию адаптивного тестирования, носит циклический характер и имеет вид, представленный на рис. 19.

Вход и выход из адаптивного тестирования. Выбор начальных оценок для входа в адаптивное тестирование осуществляется поразному, в зависимости от вида стратегии и имеющихся технологических возможностей при генерации адаптивных тестов. Один из методов определения начальных оценок основан на выдаче испытуемым перед началом адаптивного тестирования входного претеста. В претест обычно включают 5—10 заданий из различных разделов содержания, охватывающих по трудности весь диапазон предполагаемого расположения тестируемой выборки учащихся на оси переменной измерения. Иногда входное тестирование заменяют процессом самоадаптации, в котором испытуемому предлагают набор заданий возрастающей трудности. Он выполняет задание, отражающее уровень его знаний и умений.

Для выхода из режима тестирования либо вводят ограничения по времени или по числу заданий, либо задаются планируемой точностью измерений. Ориентация на точность при организации адаптивных циклов порождает многообразие индивидуальных траекторий испытуемых, которые можно визуализировать в виде ломаных линий. Вершины ломаной линии соответствуют отдельным заданиям адаптивного теста, длина звена определяется варьирующим шагом, размер которого равен разности оценок параметра трудности двух смежных заданий адаптивного теста. Очевидно, что чем меньше длина ломаной, тем лучше структура знаний учащегося и эффективнее подобраны по трудности задания адаптивного теста (рис. 20).

Гомельский государственный университет имени Ф. Скорины

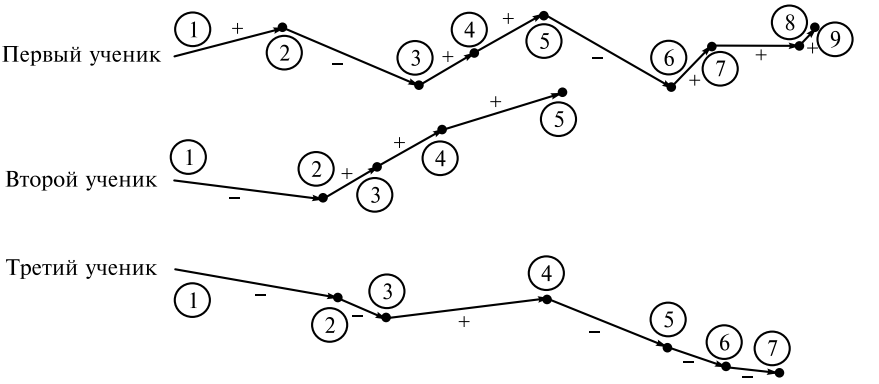

Рис. 20. Визуализация индивидуальных траекторий испытуемых: в кружках — номера заданий

На рис. 20 изображены траектории адаптивного тестирования трех учащихся, начинавших свой вход в адаптивный режим по результатам выполнения претеста. Чем выше расположена вершина ломаной, тем труднее было первое задание адаптивного теста. На момент входа в претесте самый высокий результат показал первый учащийся, поэтому он начинает адаптивное тестирование с более трудного задания. Для удобства обсуждения результатов визуализации на рисунке приводятся непересекающиеся траектории. Над ломаными ставится «плюс» в тех случаях, когда испытуемый выполнил задание верно, или «минус», если испытуемый выполнил задание неверно. В качестве критерия окончания тестирования выбрано простое правило: тестирование прекращается, если учащимся подряд выполнены верно или неверно три задания адаптивного теста.

Несмотря на высокий начальный результат, первый учащийся, по-видимому, обладает плохо структурированными знаниями, что следует из чередования верных и неверных ответов. Тестирование первого ученика прекращается, если ему удается справиться с идущими подряд тремя заданиями адаптивного теста. Траектория ответов второго учащегося намного короче благодаря хорошо структурированным знаниям. После неудачи при выполнении первого задания он все делает верно и поэтому быстро заканчивает адаптивный тест. Третий учащийся самый слабый. Он начинает тестирование с наиболее легкого задания, с которым не справляется. Второе, более легкое задание он также выполняет неверно. Наконец, после трех следующих подряд неправильных ответов он выходит из адаптивного теста.

Представленный рисунок является идеализацией, иллюстрирующей реальные ситуации варьирующих многошаговых стратегий генерации адаптивных тестов, в которых после выполнения

Северо-Восточный федеральный университет им. М.К. Аммосова

каждого задания осуществляется пересчет текущей оценки уровня подготовленности для выбора очередного задания адаптивного теста.

Надежность, валидность и длина теста при адаптивном тестировании. Так же как и при традиционном тестировании, отбор заданий в адаптивные тесты осуществляется в соответствии со спецификацией теста. Оптимизируя трудность, можно лишь уменьшить число предъявляемых заданий по каждому разделу и сохранить при этом для каждого испытуемого содержательный план теста. Таким образом, адаптивное тестирование вне зависимости от стратегии предъявления заданий и их числа должно обеспечивать высокую содержательную валидность каждого генерируемого адаптивного теста.

Надежность в адаптивном тестировании зависит от совокупности факторов. К ним относятся: число заданий, наличие систематического контроля за частотой выбора заданий банка при генерации адаптивного теста. На надежность также влияют характеристики банка тестовых заданий, связанные с качеством измерений (устойчивость и диапазон вариации оценок трудности) и качеством входного (стартового) контроля.

Адаптивный алгоритм организуется так, что после каждого очередного предъявления задания проверяется разность между полученной и запланированной точностью измерений. По достижению запланированной точности алгоритм подбора заданий приостанавливается, достигается ожидаемая надежность адаптивного теста.

Северо-Восточный федеральный университет им. М.К. Аммосова

8.5. Online-тестирование, его применение в дистанционном обучении

Уровни интерактивности. В самом простом понимании интерактивного режима обучения учащийся имеет возможность получать (читать, смотреть, слушать) только ту информацию, которую он выбирает для усвоения с использованием компьютера. Усложнение возможностей и технологии осуществления интерактивного режима приводит к моделированию окружающего мира и поведения объектов в нем, позволяя имитировать реальность.

Конечно, на сегодняшний день, в силу многих причин, в обучении используются далеко не все возможности интерактивного режима. В частности, по мнению А. Г. Шмелева, являющегося крупнейшим специалистом в России по применению интерактивных технологий в образовательном и психологическом тестировании (система «Телетестинг»), в современном Интернете преобладают неинтерактивные формы преподнесения образовательной информации [62].

Простейший интерактивный режим в локальной сети и в Интернете. В соответствии с классификацией компьютерных сетей на локальные и глобальные простейший интерактивный режим организуется в пределах одной комнаты или учебного заведения либо с использованием Интернета. Как правило, интерактивность основывается на асинхронной коммуникационной связи, когда реакция педагога на результаты тестирования носит отсроченный характер из-за времени, которое необходимо на проверку теста в автоматизированном режиме и подсчет баллов учащихся по результатам его выполнения.

В первом случае, когда в локальную сеть объединено несколько десятков или сотен компьютеров, специальная программа-реализатор — инструментальная оболочка — обеспечивает выдачу заданий online-теста для всей группы тестируемых, обычно в индивидуальном временно́м режиме. На экране каждого компьютера из локальной сети появляется задание одного из параллельных вариантов теста. При обеспечении режима информационной безопасности для всей группы учащихся может использоваться только один вариант теста.

Выполнение online-теста с использованием Интернета не имеет принципиальных отличий от случая применения локальной сети при простейшем уровне интерактивности без адаптивного режима, когда все учащиеся выполняют одинаковые варианты теста. Задания в подавляющем большинстве требуют от учащихся выбора одного или нескольких правильных ответов с помощью таких известных диалоговых объектов, как «селекторные кнопки» (radiobuttons). Подсчет тестовых баллов производится путем сличения ответов учащихся с ключом и сводится, чаще всего, к простому суммированию. Передача итогового балла по тесту может быть осуществлена с помощью электронной почты.

Время, затраченное на предъявление результата тестирования, определяется длительностью пересылки (обычно от нескольких секунд до нескольких часов) и тем временным промежутком, который пройдет до момента, когда учащийся прочтет пришедшую ему почту. В отдельных случаях, когда учащемуся требуется документальное подтверждение баллов, результаты тестирования могут быть доставлены offline с помощью записи на носитель информации. Таким образом, низкий уровень интерактивности вполне пригоден для итогового тестирования вне адаптивного режима, когда учащийся должен работать без помощи педагога, а получение результатов может носить отсроченный по времени характер.

Средний уровень интерактивности в online-тестировании. В текущем контроле при дистанционном обучении обычно реализуется средний уровень интерактивности. В соответствии с возможностями синхронного обмена информацией в реальном времени с помощью интернет-пейджеров учащемуся обеспечиваются помощь

Полесский государственный университет

и консультации педагога при выполнении заданий корректирующего и диагностического тестов.

При среднем уровне интерактивности большое разнообразие приобретают формы тестовых заданий. В частности у школьника появляется возможность редактирования текста, представленного в задании, с помощью введения новых предложений или замены одной части текста на другую. В заданиях на установление правильной последовательности сразу после выбора испытуемым некоторого порядка элементов компьютер отображает новую последовательность на экране и т.д. Если установлению синхронной связи не мешают временны́е пояса, интерактив незамедлительно обеспечивает эффект «педагог рядом», благодаря которому при выполнении заданий текущего контроля ученик получает помощь, оценку или подсказку педагога.

Высокий уровень интерактивности в online-тестировании. Высокий уровень интерактивности обеспечивается в тех случаях, когда при взаимодействии с педагогом используются звук и видеоизображение, что требует значительных финансовых затрат, но без труда позволяет идентифицировать личность учащегося, выполняющего тест в дистанционном контроле.

С педагогической точки зрения высокому уровню интерактивности отвечает адаптивное тестирование, включающее разветвленные технологии оптимизации трудности заданий в зависимости от ответов учащегося на каждое предыдущее задание адаптивного теста.

Северо-Кавказский федеральный университет им. М.К. Аммосова

Вопросы для обсуждения

1. Всегда ли компьютерное тестирование облегчает работу учителя? В каких случаях необходимо обращаться к методам компьютерной выдачи тестов?

2. Какие формы проведения компьютерного тестирования вы знаете? Существуют ли компьютерные тесты или любой набор предтестовых заданий, сгенерированный случайным образом, выдается за компьютерный тест?

3. Каково влияние опыта работы учащихся с ПК на результаты выполнения заданий при компьютерном тестировании? Какие исследования по данному направлению вы знаете?

4. В чем проявляются ограниченные возможности компьютерного тестирования?

5. Какие инновации при выдаче заданий на компьютере встречались вам в процессе обучения? Облегчали ли они обучение или, наоборот, мешали усваивать новый материал?

ГЛАВА

9

КЛАССИЧЕСКАЯ ТЕОРИЯ И МЕТОДИКИ КОНСТРУИРОВАНИЯ ТЕСТОВ

9.1. Основные этапы конструирования теста

Перечень этапов и их очередность. Процесс создания теста, его научного обоснования, переработки и улучшения можно разбить на ряд этапов, представленных ниже.

1. Определение цели тестирования, выбор вида теста и подхода к его созданию.
2. Концептуальный выбор конструкта (переменной измерения).
3. Анализ содержания учебной дисциплины и планирование содержания теста, априорный выбор длины теста и времени его выполнения, разработка спецификации теста.
4. Определение структуры теста, форм заданий и стратегии их расположения в тесте.
5. Создание предтестовых заданий.
6. Отбор заданий в тест и их ранжирование согласно выбранной стратегии предъявления на основании априорных авторских оценок трудности заданий.
7. Экспертиза формы предтестовых заданий и содержания теста.
8. Коррекция заданий и теста по результатам экспертизы.
9. Разработка методики апробационного тестирования, инструкций для учеников и преподавателей, проводящих апробацию теста.
10. Формирование репрезентативной выборки апробации.
11. Проведение апробационного тестирования.
12. Проверка результатов выполнения теста (автоматизированная или ручная), подготовка эмпирических данных тестирования к виду, удобному для обработки и проведения анализа.
13. Статистическая обработка результатов выполнения теста (автоматизированная с помощью специального программного обеспечения).
14. Анализ и интерпретация результатов обработки в целях улучшения качества теста. Проверка соответствия характеристик теста научно обоснованным критериям качества.
15. Коррекция содержания и формы заданий на основании данных предыдущего этапа. Чистка теста и добавление новых заданий для оптимизации диапазона значений параметра трудности и улуч-

Гомельский государственный университет им. Ф. Скорины

шения системообразующих свойств заданий теста. Оптимизация длины теста и времени его выполнения на основании статистических оценок характеристик теста. Оптимизация порядка расположения заданий в тесте.

16. Повторение этапа апробации для выполнения очередных шагов по повышению качества теста.

17. Интерпретация данных обработки, установление норм теста и создание шкалы для оценки результатов испытуемых.

Апробация, анализ и коррекция теста. Апробация теста неоднократно повторяется. Обычно на разработку стандартизованного теста уходит не менее 3—4 лет, поскольку для апробации важно не только сформировать репрезентативную выборку учащихся, но и выбрать подходящее время в учебном процессе.

При разработке теста возникает своеобразный цикл, так как после его чистки создателю приходится возвращаться к этапу апробации и анализа эмпирических данных тестирования, причем, как правило, не один раз.

Тщательная коррекция теста необходима особенно в тех случаях, когда тест должен быть стандартизован, а его результаты планируется использовать для принятия административно-управленческих решений в образовании.

Северо-Восточный федеральный университет им. М.К. Аммосова

9.2. Классическая (традиционная) теория тестов

Основное предположение классической теории тестов. Предположение о существовании истинного балла (true score) является основополагающим в классической теории тестов.

Нередко в одномерных измерениях истинный балл называют параметром учащегося, при этом предполагается, что каждому ученику можно поставить в соответствие единственное на момент измерения значение параметра, не зависящее от применяемого теста. В целом истинный балл — это идеализированная константа испытуемого в гипотетической генеральной совокупности заданий бесконечного теста.

Постулаты классической теории тестов. Помимо предположения о существовании истинного балла в классической теории тестов выделяют несколько постулатов, позволяющих построить математико-статистический аппарат для разработки научно обоснованных тестов и оценки качества результатов педагогических измерений [60; 81]. Эти постулаты связаны с предположениями:

– о равенстве ковариаций результатов тестирования по параллельным формам;

– о приближении средних значений ошибок измерения истинных баллов к нулю при числе тестирований, стремящемся к бесконечности;

– о инвариантности истинных баллов относительно различных параллельных форм теста;

– о континуальном (непрерывном) распределении истинных баллов в генеральной совокупности учащихся;

– о нормальном законе распределения наблюдаемых баллов, истинных баллов и ошибок измерения.

9.3. Математико-статистический анализ качества тестов и тестовых заданий на основе классической теории тестов

Матрица тестовых результатов. Если за каждый правильный ответ на задание испытуемому давать один балл, а за неправильный ответ или пропуск задания — нуль баллов, то профиль ответов учащегося будет иметь вид последовательности из единиц и нулей. Поскольку каждая единица или нуль появляются в результате взаимодействия испытуемого с заданием, то наиболее адекватной формой представления наблюдаемых результатов выполнения теста будет служить матрица, т.е. прямоугольная таблица, сводящая воедино профили ответов учащихся (строки из оценок учащегося по всем заданиям теста) и профили заданий теста (столбцы из оценок всех учащихся по каждому заданию теста).

Интегрирование данных тестирования в форме матрицы удобно для обработки и отражает взаимодействие множеств испытуемых и заданий, происходящее при выполнении теста (рис. 21).

При геометрической интерпретации этого взаимодействия по горизонтальной оси откладываются оценки параметра трудности заданий теста, по вертикальной — оценки подготовленности тестируемых учащихся. Взаимодействие между i-м испытуемым и j-м заданием порождает наблюдаемый ответ x_{ij}, который при дихотомической оценке принимает одно из двух значений (см. табл. 6).

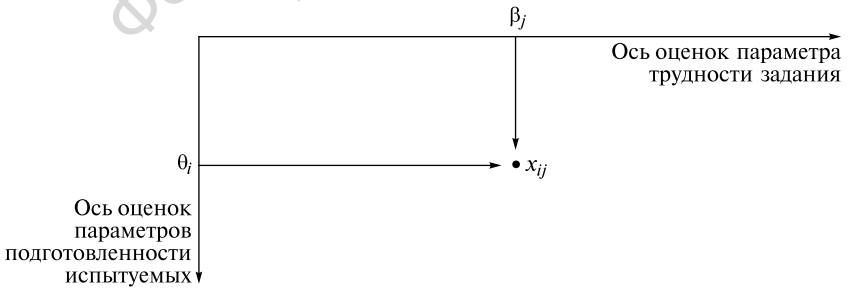

Рис. 21. Геометрическая интерпретация взаимодействия множеств испытуемых и заданий теста

Гомельский государственный университет имени Ф. Скорины

Таблица 6

Правило дихотомического оценивания ответа

Ответ x_{ij}	Оценка
Ответ i-го ученика на j-е задание правильный	1
Ответ i-го ученика на j-е задание неправильный	0

Таблица 7

Матрица наблюдаемых результатов выполнения теста

Испытуемые	Задания $1 \ldots j \ldots n$	Индивидуальный балл (X_i)
1 ... i ... N	$x_{ij} = \begin{cases} 1 \\ 0 \end{cases}$	$X_i = \sum_{j=1}^{n} x_{ij}$
Количество правильных ответов на задания (R)	$R_j = \sum_{i=1}^{N} x_{ij}$	$\sum_{i=1}^{N} X_i = \sum_{j=1}^{n} R_j$

Общий вид матрицы наблюдаемых результатов выполнения N учащимися n заданий теста при дихотомических оценках по заданиям приведен в табл. 7.

Справа в матрице, в вертикальном столбце, содержатся индивидуальные баллы учеников X_i ($i = 1, 2, \ldots, N$), которые получаются суммированием единиц по горизонтали в каждом профиле ответов учащегося. Сложение единиц в столбцах по профилям отве-

Таблица 8

Матрица результатов тестирования

Номер испытуемого i	Номер заданий j									
	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	0	0	0	0
2	1	1	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	1	0	0
4	1	1	0	1	1	1	1	1	1	1
5	1	0	1	0	1	1	0	0	0	0
6	1	1	1	0	0	0	0	1	0	0
7	1	1	1	1	0	1	0	0	0	0
8	1	1	1	1	0	0	0	0	0	0
9	1	1	1	1	1	1	1	1	1	0
10	1	1	1	1	1	0	1	0	0	0

Северо-Восточный федеральный университет им. М.К. Аммосова

тов на n заданий теста позволяет получить числа R_j ($j = 1, 2, \ldots, n$), соответствующие количеству правильных ответов на задания теста.

После занесения результатов выполнения теста в матрицу начинается этап математико-статистической обработки, который включает ряд шагов. Из дидактических соображений для иллюстрации методов обработки выбрана небольшая матрица, когда 10 учеников отвечали всего на 10 заданий теста (табл. 8). Однако все формулы и подсчеты, обсуждаемые в разделе, могут быть распространены на любые выборки испытуемых и применимы к тестам любой длины.

Первый шаг математико-статистической обработки эмпирических данных тестирования. На первом шаге обработки данных тестирования подсчитываются индивидуальные баллы и число правильных ответов на каждое задание теста. Для подсчета индивидуального балла суммируются все единицы, полученные учащимся за правильно выполненные задания теста. Например, четвертый испытуемый выполнил правильно 9 заданий, поэтому его индивидуальный балл равен 9. Для удобства полученные индивидуальные баллы X_i ($i = 1, 2, \ldots, 10$) приводятся в последнем столбце матрицы результатов (табл. 9).

Число правильных ответов на каждое задание R_j также получается суммированием единиц, но уже расположенных по столб-

Таблица 9

Матрица результатов с индивидуальными баллами испытуемых и количеством правильных ответов на задания теста

Номер испытуемого i	Номер заданий j										Индивидуальные баллы (множество X_i)
	1	2	3	4	5	6	7	8	9	10	
1	1	1	1	1	1	1	0	0	0	0	6
2	1	1	0	0	0	0	0	0	0	0	2
3	0	0	0	0	0	0	0	1	0	0	1
4	1	1	0	1	1	1	1	1	1	1	9
5	1	0	1	0	1	1	0	0	0	0	4
6	1	1	1	0	0	0	0	1	0	0	4
7	1	1	1	1	0	1	0	0	0	0	5
8	1	1	1	1	0	0	0	0	0	0	4
9	1	1	1	1	1	1	1	1	1	0	9
10	1	1	1	1	1	0	1	0	0	0	6
Число правильных ответов (множество R_j)	9	8	7	6	5	5	3	4	2	1	50

Северо-Восточный федеральный университет им. М.К. Аммосова

цам, и размещается в матрице результатов в последней строке под номером соответствующего задания теста.

Второй шаг математико-статистической обработки эмпирических данных тестирования. На втором шаге обработки данных осуществляется упорядочение матрицы результатов тестирования. Для этого производится перестановка столбцов, числа R_i располагаются в порядке убывания. Затем меняются местами строки матрицы так, чтобы верхняя строка соответствовала обучаемому с минимальным индивидуальным баллом. Значения X_i располагаются сверху вниз в порядке возрастания. Упорядоченная матрица данных тестирования приведена в табл. 10.

Третий шаг математико-статистической обработки эмпирических данных тестирования. На данном этапе производится графическая интерпретация распределений эмпирических данных, которые можно представить в виде полигона, гистограммы или сглаженной кривой (процентилей, огивы). Для графической интерпретации результатов учащихся необходимо их предварительное упорядочение в виде несгруппированного ряда произвольной формы (табл. 11), ранжированного ряда (табл. 12), частотного распределения или распределения сгруппированных частот [1; 18; 59].

В табл. 11 содержатся индивидуальные баллы испытуемых, взятые из последнего столбца матрицы эмпирических результатов выполнения теста (см. табл. 9). В табл. 12 эти баллы располагаются в

Таблица 10

Упорядочная матрица данных тестирования

Номера испытуемых i	Номера заданий j										X_i
	1	2	3	4	5	6	7	8	9	10	
3	0	0	0	0	0	0	1	0	0	0	1
2	1	1	0	0	0	0	0	0	0	0	2
5	1	0	1	0	1	1	0	0	0	0	4
6	1	1	1	0	0	0	1	0	0	0	4
8	1	1	1	1	0	0	0	0	0	0	4
7	1	1	1	1	0	1	0	0	0	0	5
1	1	1	1	1	1	1	0	0	0	0	6
10	1	1	1	1	1	0	0	1	0	0	6
9	1	1	1	1	1	1	1	1	1	0	9
4	1	1	0	1	1	1	1	1	1	1	9
R_j	9	8	7	6	5	5	4	3	2	1	50

Северо-Восточный федеральный университет имени М.К. Аммосова

Таблица 11

Несгруппированный ряд

Номер	1	2	3	4	5	6	7	8	9	10
Балл	6	2	1	9	4	4	5	4	9	6

Таблица 12

Ранжированный ряд

Номер	3	2	5	6	8	7	1	10	4	9
Балл	1	2	4	4	4	5	6	6	9	9
Ранг	1	2	3	3	3	4	5	5	6	6

порядке возрастания слева — направо, а также приводятся места (ранги) испытуемых, соответствующие их индивидуальным баллам.

Данные таблицы удобны для подведения итогов тестирования в работе педагога, поскольку в классе распределения сырых баллов вполне достаточно для сообщения тестовых результатов ученикам. Например, балл 6 обеспечивает первому испытуемому ранг 5 в группе из 10 учеников. Аналогичным образом можно интерпретировать любую оценку ученика в терминах рангов. Очевидно, что равным баллам приписываются равные ранги.

Если группа учащихся велика, то для определения рангов используют классификацию оценок по распределению частот или строят сгруппированное частотное распределение. По ряду частотного распределения можно получить графическое представление результатов тестирования в виде полигона частот и гистограммы — последовательности столбцов, каждый из которых опи-

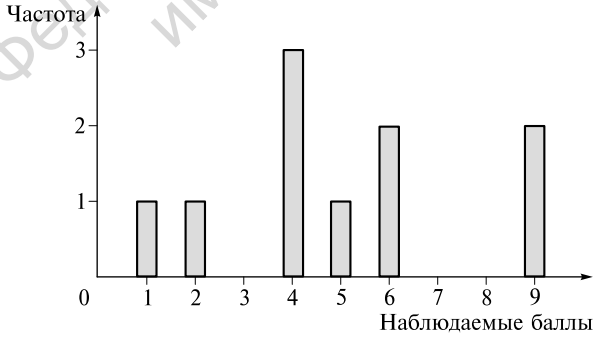

Рис. 22. Столбиковая гистограмма для распределения баллов в матрице, представленной в табл. 9

Оренбургский федеральный университет им. И.С. Аммосова

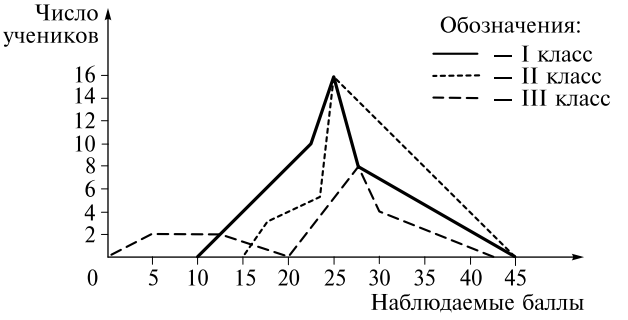

Рис. 23. Сравнение результатов тестирования

рается на единичный (разрядный) интервал и высота которых пропорциональна частоте наблюдаемых баллов [18; 59].

Например, матрице, представленной в табл. 10, соответствует гистограмма, приведенная на рис. 22. Середина столбца совмещается с серединой интервала разряда, длина которого равна одному баллу.

Для сравнения двух или более распределений обычно используют полигоны частот, так как при наложении гистограмм получается довольно запутанная картина.

Например, с помощью полигонов можно сравнить результаты выполнения теста учащимися различных классов, имеющих одинаковое количество учеников (рис. 23). На рисунке отчетливо видно значительное сходство в результатах тестирования у первых двух классов, имеющих довольно похожие полигоны распределения оценок.

Четвертый шаг математико-статистической обработки эмпирических данных тестирования. На данном этапе обработки данных оцениваются меры центральной тенденции в распределении результатов тестирования, предназначенные для выявления той точки, вокруг которой в основном группируются все результаты выполнения теста [1; 18; 59]. При анализе результатов тестирования можно использовать разные способы определения такой центральной точки. Наиболее простой из них основан на выявлении *моды распределения*.

Мода — это такое значение, которое встречается наиболее часто среди результатов выполнения теста. Например, для данных матрицы, представленной в табл. 10, модой является балл «4», потому что он встречается чаще (три раза) любого другого значения балла. Распределение может иметь одну или несколько мод. В случае существования двух мод распределение называется бимодальным. Если все значения баллов учеников встречаются одинаково часто, принято считать, что моды у распределения нет.

Мозырский государственный педагогический университет им. И.П. Шамякина

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

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}. \qquad (3)$$

Среднее арифметическое индивидуальных баллов испытуемых для рассмотренного выше примера матрицы (см. табл. 10) равно

$$\bar{X} = \frac{6 + 2 + 1 + 9 + 4 + 4 + 5 + 4 + 9 + 6}{10} = 5.$$

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. С другими свойствами среднего выборочного можно познакомиться в учебнике по статистике.

Меры центральной тенденции полезны при оценке качества теста в том случае, когда есть результаты апробации теста на репрезентативной выборке учеников. Обычно считают, что хороший нормативно-ориентированный тест обеспечивает нормальное распределение индивидуальных баллов репрезентативной выборки учеников, когда среднее значение баллов находится в центре распределения, а остальные значения концентрируются вокруг среднего по нормальному закону, т.е. примерно 70 % значений в центре, а остальные сходят на нет к краям распределения, как показано на рис. 24.

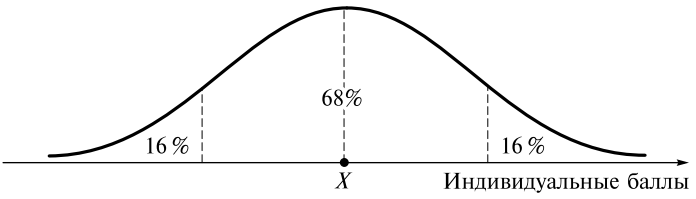

Рис. 24. Нормальная кривая распределения индивидуальных баллов

Современный гуманитарный университет
филиал в г. Минске

Нормальная кривая — изобретение математиков — в сглаженном, идеальном виде описывает реальный полигон частот. На практике никогда не была и не будет получена совокупность данных, распределенных точно по нормальному закону, просто иногда полезно, допуская определенную ошибку, утверждать, что распределение эмпирических данных близко к нормальной кривой. Нормальное распределение унимодально и симметрично, т. е. половина результатов, расположенная ниже моды, в точности совпадает с другой половиной, расположенной выше, а мода и среднее значение равны.

Если тест обеспечивает распределение баллов, близкое к нормальному, то это означает, что с его помощью можно определить устойчивое среднее, которое принимается в качестве одной из репрезентативных норм выполнения теста. Обратный вывод неверен: устойчивость тестовых норм вовсе не предполагает обязательного нормального распределения эмпирических результатов выполнения теста. Таким образом, правильно сконструированный нормативно-ориентированный тест на репрезентативной выборке учеников должен обеспечивать близкое к симметричному распределение индивидуальных баллов, когда мода и среднее значение примерно равны, а остальные результаты расположены вокруг среднего по нормальному закону.

Пятый шаг математико-статистической обработки эмпирических данных тестирования. На данном этапе определяются описательные характеристики, служащие мерами изменчивости в распределении данных по тесту [1; 18; 59]. Введение мер изменчивости связано с необходимостью выявления дополнительных оснований для сравнения различных распределений по тестам. Если распределения имеют одинаковые средние, то, оценивая и анализируя меры изменчивости, можно выявить существенные отличия в качестве тестов.

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

Размах измеряет на шкале расстояние, в пределах которого изменяются все значения показателя в распределении. Например, для распределения индивидуальных баллов, представленных в табл. 10, размах равен 9 – 1 = 8. Вариационный размах легко вычисляется, но при характеристике распределения баллов по тесту используется крайне редко. В о-п е р в ы х, размах является весьма приближенным показателем, так как не зависит от степени изменчивости промежуточных значений, расположенных между

Электронный архив библиотеки МГУ имени А.А. Кулешова

крайними значениями в распределении баллов по тесту. В о-в т о р ы х, крайние значения индивидуальных баллов, как правило, ненадежны, поскольку содержат в себе значительную ошибку измерения. В этой связи более удачной мерой изменчивости считается *дисперсия*.

Подсчет дисперсии основан на вычислении отклонений $X_1 - \bar{X}$ (i = 1, 2, ..., N) каждого значения показателя от среднего арифметического в распределении. Для ученика с индивидуальным баллом выше среднего значение разности $X_1 - \bar{X}$ будет положительно, а для тех, у кого результат ниже $\bar{X}$, отклонение $X_1 - \bar{X}$ будет меньше нуля.

Если просуммировать все отклонения, взятые со своим знаком, то для симметричных распределений сумма будет равна нулю. Чтобы отрицательные и положительные слагаемые не уничтожали друг друга, каждое отклонение возводят в квадрат, а затем находят сумму квадратов отклонений. Эта сумма будет большой, если результаты тестирования отличаются существенной неоднородностью, и малой — в случае близких результатов испытуемых по тесту. Для матрицы, представленной в табл. 9, сумма квадратов отклонений будет равна

$$\sum_{i=1}^{10}\left(X_i-\bar{X}\right)=(-4)^2+(-3)^2+(-1)^2+(-1)^2+(-1)^2+0^2+1^2+1^2+4^2+4^2=62.$$

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

$$S_x^2=\frac{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2}{N-1}. \qquad (4)$$

Для рассматриваемого примера $S_x^2=\frac{62}{10-1}=\frac{62}{9}\approx 6,89.$

Кроме дисперсии для характеристики меры изменчивости распределения удобно использовать еще один показатель вариации, который называется стандартным отклонением и вычисляется путем извлечения квадратного корня из дисперсии:

$$S_x=\sqrt{S_x^2}. \qquad (5)$$

Для рассматриваемого примера данных тестирования $S_x \approx$ $\approx \sqrt{6,89} \approx 2,62$. Свойства дисперсии и стандартного отклонения

Северо-Восточный федеральный университет им. М.К. Аммосова

рассматриваются подробно в учебниках по статистике. Заинтересованному читателю можно порекомендовать, например, книгу Дж. Гласс и Дж. Стенли «Статистические методы в педагогике и психологии» [18].

Дисперсия играет важную роль в оценке качества тестов. Низкая дисперсия указывает на плохое качество нормативно-ориентированного теста, поскольку не обеспечивается высокий дифференцирующий эффект. Излишне высокая дисперсия, характерная для случая, когда все учащиеся различаются 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. $\bar{X} \pm 1S_x$); 95% площади под кривой расположено в пределах двух S_x ($\bar{X} \pm 2S_x$); 99,7% площади под кривой — в пределах трех S_x ($\bar{X} \pm 3S_x$). Из всех нормальных кривых наиболее удобна единичная, площадь под которой равна 1. Для нее среднее значение равно нулю ($\bar{z} = 0$), а стандартное отклонение единице ($\sigma_z = 1$).

При использовании теста необходимо помнить о том, что получаемое распределение индивидуальных баллов учащихся является следствием подбора трудности заданий теста, как показано на рис. 25.

Для верхнего распределения слева характерно смещение в сторону легких заданий, поэтому бóльшая часть учащихся выполнит почти все задания теста и получит высокие индивидуальные баллы, показанные на правом верхнем рисунке. Средние графики отражают тенденцию к приоритетному подбору самых трудных заданий при разработке теста и вытекающий отсюда всплеск у начала горизонтальной оси там, где располагаются низкие индивидуальные баллы. Тест, представленный на нижнем графике слева, обладает сбалансированной трудностью, что автоматически приводит к нормальности распределения индивидуальных баллов репрезентативной выборки учеников.

Это позволяет считать полученное распределение устойчивым по отношению к генеральной совокупности, а также помогает определить репрезентативные нормы выполнения теста.

Последующие шаги обработки данных предназначаются для оценивания мер симметрии и островершинности кривых распределений [1; 18; 60; 63] и выполняются обычно при разработке

Северо-Осетинский государственный университет имени К.Л. Хетагурова

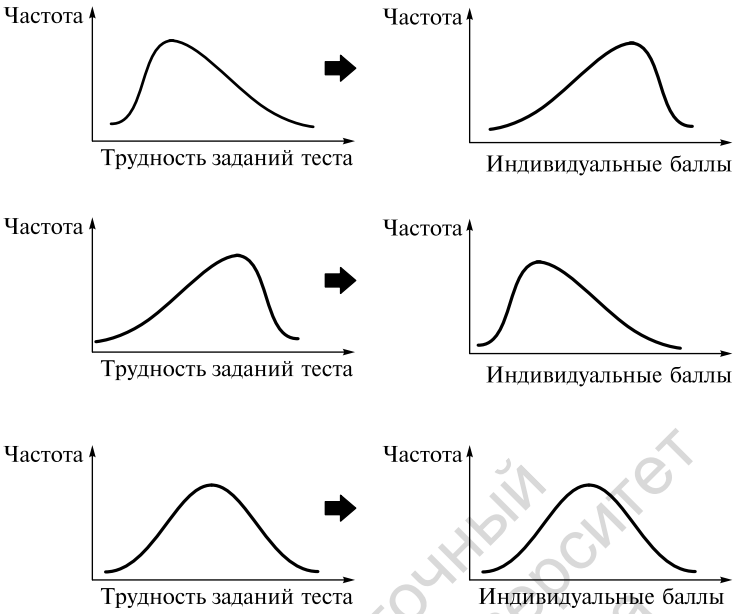

Рис. 25. Связь распределения индивидуальных баллов и трудности заданий теста

Северо-Восточный федеральный университет им. М.К. Аммосова

тестов административно-управленческого предназначения не «руками», а с помощью специальных статистических пакетов для ПК.

9.4. Показатели связи между заданиями теста

Корреляция результатов учащихся по заданиям. Для итогового контроля полезно вычислять показатели связи между результатами учеников по отдельным заданиям теста. При этом важно понять, существует ли тенденция, когда одни и те же ученики добиваются успеха в какой-либо паре заданий теста, или состав учеников, 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 [18; 60; 63]. Для ее оценивания в общем случае применяют коэффициент корреляции Пирсона r_{xy}, значения которого меняются в интервале от –1 до +1.

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\left(\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2\right)\left(\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2\right)}}, \quad (6)$$

где $X_1, \ldots, X_N$ — первый набор данных со средним значением $\bar{X}$, а $Y_1, \ldots, Y_N$ — второй набор данных со средним значением $\bar{Y}$.

При исследовании связи между наборами данных необходимо правильно выбрать вид и форму показателя, зависящие от шкал, в которых представлены данные [18]. В частности, для оценки связи между результатами выполнения учащимися двух заданий теста коэффициент корреляции Пирсона r_{xy} необходимо преобразовать, поскольку результаты выполнения заданий представляются в дихотомической шкале (столбцы из нулей и единиц в матрице данных по тесту). Преобразованный коэффициент Пирсона для дихотомических данных называется коэффициентом «фи» и вычисляется по формуле

$$\varphi_{jl} = \frac{p_{jl} - p_j p_l}{\sqrt{(p_j q_j)(p_l q_l)}}, \qquad (7)$$

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

Анализ значений коэффициента корреляции φ позволяет выявить неудачные задания теста, которые отрицательно коррелируют с большинством остальных заданий и, следовательно, измеряют нечто иное, чем та переменная, для которой предназначался тест. Отрицательные значения коэффициента корреляции указывают на определенный просчет разработчиков в содержании заданий, которые рекомендуется удалить из теста. Наиболее распространенная причина появления отрицательной корреляции — отсутствие предметной чистоты содержания — встречается при разработке самых разных тестов довольно часто.

Предметная чистота — скорее идеализируемое, чем реальное требование к содержанию любого теста. Так, в любом тесте по физике встречаются задания с большим количеством математических преобразований, в тесте по биологии — задания, требующие серьезных знаний по химии, в тесте по истории — задания, рассчитанные на выявление культурологических знаний, и т. п. Поэтому можно лишь стремиться к тому, чтобы при выполнении каждого задания доминировали знания по проверяемому предмету.

Для тематических тестов характерна высокая корреляция между заданиями, так как они в большинстве случаев имеют слабо варьирующее исходное содержание, что вполне объясняется назначением теста. Однако в итоговых тестах по возможности стараются избегать высокой корреляции между заданиями, поскольку

Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М.А. Шолохова»

вряд ли имеет смысл включать в итоговый тест несколько заданий, оценивающих одинаковые содержательные элементы. В итоговых тестах обычно стремятся к невысокой положительной корреляции, когда значения коэффициента варьируют в интервале (0; 0,3) и каждое задание вносит свой вклад в общее содержание теста.

Бисериальный коэффициент корреляции. 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 по заданиям теста (интервальная или квазиинтервальная шкала). С помощью подсчета значений бисериального коэффициента корреляции оценивается валидность, иногда называемая *показателем дифференцирующей способности* (*дискриминативности*) заданий теста.

Объяснение, на котором основан вывод формулы для подсчета бисериального коэффициента корреляции, приводится в ряде исследований [18; 60; 73]. Его вычисление требует использования специальных таблиц для нахождения ординат стандартной нормальной кривой и определенной математической подготовки. Поэтому нередко используют другой коэффициент корреляции, называемый точечным бисериальным коэффициентом — r_{pbis}. Основанием для подобной замены служит близость значений этих коэффициентов: первый незначительно превышает второй, если они подсчитаны для одних и тех же наборов данных из распределений. Однако формула для r_{pbis} намного проще, поэтому именно ему часто отдают предпочтение в практической работе.

Анализ значений коэффициента бисериальной корреляции, подсчитанного для оценки связи результатов по каждому заданию с суммой баллов по тесту, позволяет выявить задания с низкой валидностью, с помощью которых трудно отделить хорошо подготовленных учащихся от слабо подготовленных учащихся тестируемой группы. Значения, близкие к нулю, указывают на низкую дифференцирующую способность заданий теста. Если коэффициент бисериальной корреляции принимает отрицательные значения, задание следует удалить из теста, так как при выполнении такого теста слабые ученики выполняют его верно, а сильные выбирают неверный ответ либо пропускают задание.

9.5. Оценка характеристик заданий теста

Оценка трудности заданий по классической теории тестов. Оценка трудности тестовых заданий в классической теории тестов осуществляется по формуле

$$p_j = \frac{R_j}{N}, \qquad (8)$$

где p_j — доля правильных ответов на j-е задание; R_j — количество учеников, выполнивших j-е задание верно; N — число учеников в тестируемой группе; j — номер задания теста (j = 1, 2, ..., n). Трудность задания нередко выражают в процентах. Для этого оценку, полученную по формуле (8), умножают на 100 %.

Долю правильных ответов на задание p_j правильнее было бы назвать легкостью задания, в то время как трудность 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 нормативно-ориентированных тестах. В хорошо сбалансированном по трудности тесте всегда есть несколько самых легких заданий со значениями $p \to 0$ и несколько самых трудных со значением $p \to 1$. Остальные задания по значениям p занимают промежуточное положение между этими крайними ситуациями и имеют в основном трудность 60 — 70 %.

Дополнительный аргумент в пользу преимущественного включения заданий средней трудности с p = 0,5 связан с подсчетом дисперсии по каждому заданию теста, которая для дихотомического набора данных будет равна $\sigma_j = p_j q_j$ (j = 1, 2, ..., n). Так как произведение $p_j q_j$ достигает максимального значения (0,5 · 0,5 = 0,25) при $p_j = 0{,}5 = q_j$, в рамках нормативно-ориентированного подхода наиболее удачными считаются задания средней трудности $p = q = 0{,}5$, обеспечивающие максимальный вклад в общую дисперсию теста.

В пользу преимущественного выбора заданий средней трудности также говорит подсчет ошибки измерения, которая уменьшается по мере продвижения к центру, где расположены задания средней трудности, и увеличивается на концах распределения.

Связь оценок трудности и валидности заданий. Интересна взаимосвязь показателей трудности и валидности (дискриминативности) заданий теста. Задания с высокой дискриминативностью обычно имеют среднюю трудность, поскольку именно для них характерен в первую очередь высокий дифференцирующий эффект. Однако обратное заключение неверно. Задания с p = 0,5 могут иметь как высокий, так и низкий дифференцирующий эффект.

При подсчете статистик по тесту всегда проводится проверка значимости полученных оценок дисперсии, асимметрии и т. д. Для этого к данным, собранным по тесту, необходимо добавить информацию о принимаемом уровне риска допущения ошибки в статистическом выводе. Наиболее приемлемым для педагогических измерений является уровень в 5 %, который допускает ошибку в 5 случаях из 100. После выбора степени риска проверка зна-

чимости проводится одним из описанных в литературе методов [18; 74].

Гомогенность (содержательная однородность) задания. При конструировании теста необходимо иметь четкое представление о содержании заданий, которые предполагается включить в окончательную версию теста. При одномерных измерениях содержание заданий должно отвечать свойству гомогенности, указывающему на степень его однородности с точки зрения оцениваемого параметра подготовленности ученика. Таким образом, гомогенность (однородность) — это характеристика задания, отражающая степень соответствия его содержания измеряемому свойству ученика. Степень гомогенности содержания обычно оценивают с помощью факторного и корреляционного анализа.

Практическое задание

В приведенной ниже таблице содержатся ответы 30 испытуемых на одно задание теста. Всех испытуемых можно разбить на две подгруппы, в одной из которых — 15 испытуемых с высоким уровнем подготовленности (сильная группа), а в другой — 15 испытуемых с низким уровнем подготовленности (слабая группа). По данным таблицы вычислите:

а) среднее значение тестовых баллов по сильной и по слабой группе, сравните их;

б) дисперсию баллов по всей группе (30 испытуемых);

в) долю правильных ответов на задание по сильной группе;

г) долю правильных ответов на задание по слабой группе;

д) корреляцию между ответами испытуемых на задание и суммой баллов по тесту для сильной группы;

е) корреляцию между ответами испытуемых на задание и суммой баллов по тесту для слабой группы.

Сибирский Федеральный университет им. М.К. Аммосова

Группа с низкой подготовленностью			Группа с высокой подготовленностью		
Испытуемый	Ответы на задание	Индивидуальный балл по тесту	Испытуемый	Ответы на задание	Индивидуальный балл по тесту
1	0	8	8	1	33
2	0	12	9	0	28
3	0	6	10	1	29
4	0	12	11	1	30
5	0	8	12	1	29
6	0	8	13	0	28
7	0	8	14	1	33

Окончание таблицы

Группа с низкой подготовленностью			Группа с высокой подготовленностью		
Испытуемый	Ответы на задание	Индивидуальный балл по тесту	Испытуемый	Ответы на задание	Индивидуальный балл по тесту
8	0	11	8	1	32
9	1	13	9	1	32
10	0	4	10	1	33
11	1	14	11	0	34
12	1	13	12	1	35
13	1	10	13	1	34
14	1	9	14	1	38
15	0	8	15	1	37

Дайте интерпретацию полученных результатов. Наблюдается ли инвариантность вычисленных характеристик задания относительно уровня подготовленности выборки? Как можно объяснить наличие (отсутствие) инвариантности?

Северо-Восточный федеральный университет им. М.К.Аммосова

ГЛАВА

10

СОВРЕМЕННАЯ ТЕОРИЯ КОНСТРУИРОВАНИЯ ТЕСТОВ

10.1. Основные положения современной теории

Item Response Theory (IRT) и область ее применения. В 80-х гг. XX в. в педагогических измерениях получили широкое развитие методы современной теории тестов Item Response Theory, сокращенно IRT [60; 65; 75; 83]. Прямой перевод на русский язык исторически сложившегося названия этой теории ничего не говорит о ее сути, поэтому в русскоязычной литературе широко используется название «теория IRT». В целом IRT предназначена для оценивания латентных параметров испытуемых и заданий тестов на основе математико-статистических моделей измерения и является частью более общей теории латентно-структурного анализа (LSA), хотя каждое из этих направлений имеет свои характерные особенности и свою сферу применения. В частности область использования LSA — социально-психологические исследования, в то время как IRT применяется в основном для конструирования и интерпретации результатов выполнения педагогических тестов.

IRT намного эффективнее традиционной теории тестов, поскольку обеспечивает более высокие точность, уровень измерений и качество тестов. Это осуществляется благодаря математико-статистическому аппарату теории, требующему привлечения дорогостоящих программных продуктов, тщательной стратификации выборок испытуемых при разработке тестов и значительным трудозатратам на согласование данных измерения с требованиями математических моделей измерения. Эти трудности разработчикам необходимо учитывать при выборе основополагающей теории конструирования тестов.

Латентные параметры и их связь с наблюдаемыми результатами тестирования. Построение теории IRT основано на предположении о существовании функциональной связи между латентными параметрами испытуемых и наблюдаемыми результатами выполнения теста. Первопричиной являются латентные параметры испытуемых, взаимодействие которых с заданиями в процессе тестирования порождает наблюдаемые результаты выполнения теста. На практике всегда ставится обратная задача: по ответам испыту-

Свободный репозиторий Полесского государственного университета им. М.А. Шолохова

емых на задания теста оценить значения латентного параметра θ_i ($i = 1, 2, ..., N$), определяющие уровень подготовки N испытуемых, и латентного параметра β_j ($j = 1, 2, ..., n$), равные оценкам трудности n заданий теста.

Для решения этой задачи датский математик Г. Раш предложил математическую модель связи между латентными параметрами и наблюдаемыми результатами тестирования, содержащую соотношение между латентными параметрами θ и β в виде разности $\theta - \beta$ при условии, что параметры θ и β оцениваются в одной и той же шкале. В качестве такой единой шкалы Г. Раш ввел интервальную шкалу логитов.

Если рассматривать значение параметра θ_i как положение i-го испытуемого на шкале логитов, а значение β_j — как положение j-го задания на той же шкале, то разность параметров получает интересную геометрическую интерпретацию. Абсолютная величина разности $|\theta_i - \beta_j|$ — это расстояние, на котором находится испытуемый с уровнем подготовки θ_i от задания с трудностью β_j. Если эта разность велика по модулю и отрицательна, то задание бесполезно для измерения уровня подготовленности i-го ученика, поскольку он наверняка не сможет выполнить такое трудное задание верно. Большие положительные значения этой разности тоже не представляют интереса ни для процесса контроля, ни для обучения i-го испытуемого, поскольку они говорят о том, что задания такой трудности давно освоены учащимся и он справится с ними успешно при выполнении теста. С точки зрения подхода, предлагаемого в IRT, такие задания неэффективны для оценивания данного значения θ. Наименьшую ошибку измерения обеспечивают задания, трудность которых приблизительно равна уровню подготовленности испытуемого, т. е. задания, подобранные по критерию $\theta \approx \beta$.

10.2. Математические модели современной теории тестов

Условная вероятность правильного выполнения обучаемыми заданий теста как функция одной переменной. В качестве математической модели взаимосвязи между значениями латентных переменных θ, β и наблюдаемыми результатами выполнения теста в IRT выбрана условная вероятность правильного выполнения обучаемыми заданий теста. В частности можно рассматривать условную вероятность P_i правильного выполнения i-м испытуемым с уровнем подготовки θ_i различных по трудности заданий теста, считая θ_i параметром i-го ученика, а β — независимой переменной.

$$P_i\left\{x_{ij} = 1 \middle| \theta_i\right\} = f\left(\theta_i - \beta\right), \quad i = 1, 2, ..., N. \qquad (9)$$

Северо-Восточный федеральный университет имени М.К. Аммосова

Аналогично вводится P_j для обозначения вероятности правильного выполнения j-го задания трудностью β_j различными испытуемыми группы. Здесь θ — независимая переменная, а β_j — параметр, определяющий трудность j-го задания теста:

$$P_j\{x_{ij} = 1|\beta_j\} = \varphi(\theta - \beta_j),\ j = 1, 2, \ldots, n, \qquad (10)$$

где $x_{ij} = \begin{cases} 1, \text{ если ответ } i\text{-го испытуемого на } j\text{-е задание верный;} \\ 0, \text{ если ответ } i\text{-го испытуемого на } j\text{-е задание не верный;} \end{cases}$

f и φ — символы функциональной зависимости; N — число испытуемых; n — количество заданий в тесте.

Если подставить в функцию $P_j(\theta)$ значение переменной $\theta = \theta_i$ или в функцию $P_i(\beta)$ значение $\beta = \beta_j$, то получится выражение для вероятности P_{ij}, значения которой можно охарактеризовать следующим образом:

$P_{ij} \to 1$, когда $\theta_i - \beta_j$ намного больше 0;

$P_{ij} \to 0$, когда $\theta_i - \beta_j$ намного меньше нуля;

$P_{ij} = 1/2$ при $\theta_i = \beta_j$.

Геометрическая интерпретация связи между разностью латентных параметров и вероятностью правильного ответа на задания теста. Связь между значениями разности $\theta_i - \beta_j$ и вероятностью правильного ответа i-го испытуемого на j-е задание теста показана на рис. 26.

В теории IRT график функции P_j получил название характеристической кривой j-го задания (ICC), а график функции P_i — индивидуальной кривой i-го испытуемого (PCC).

При выборе вида функций P_j и P_i учитываются обстоятельства как эмпирического, так и математического характера. В предположении нормального распределения значений латентных переменных θ и β предлагаются две такие функции. Одна из них, обычно обозначаемая символом $\Psi(x)$, относится к семейству логистических кривых, другая — $\Phi(x)$ — является интегральной функцией нормированного нормального распределения. Поскольку для од-

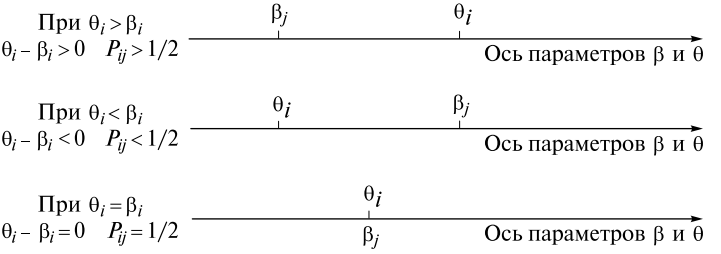

Рис. 26. Соотношение между значениями разности $\theta_i - \beta_j$ и вероятностью правильного ответа

Филиал Белорусско-Российского университета в г. Могилеве

них и тех же значений x ординаты точек графиков функций $\Phi(x)$ и $\Psi(1{,}7x)$ отличаются друг от друга незначительно, то в том, что их две, нет ни ошибки, ни противоречия. Наиболее убедительный аргумент в пользу логистической функции связан не с качеством измерений, а с относительной простотой ее аналитического задания, облегчающей оценивание параметров θ и β. Поэтому в практических приложениях предпочтение обычно отдают функции $\Psi(1{,}7x)$.

Классы логистических функций. Число параметров, входящих в аналитическое задание функций, является основанием для подразделения семейств логических функций на классы. Среди логистических функций различают:

1) однопараметрическую модель Г. Раша —

$$P_j(\theta) = \frac{e^{1,7(\theta-\beta_j)}}{1+e^{1,7(\theta-\beta_j)}} \qquad (11)$$

и

$$P_j(\beta) = \frac{e^{1,7(\theta_i-\beta)}}{1+e^{1,7(\theta_i-\beta)}}, \qquad (12)$$

где θ и β — независимые переменные для первой и второй функций соответственно;

2) двухпараметрическую модель А. Бирнбаума —

$$P_j(\theta) = \frac{e^{1,7a_j(\theta-\beta_j)}}{1+e^{1,7a_j(\theta-\beta_j)}} \qquad (13)$$

и

$$P_i(\beta) = \frac{e^{1,7a_i(\theta_i-\beta)}}{1+e^{1,7a_i(\theta_i-\beta)}}. \qquad (14)$$

Кроме прежних обозначений в формулах (13) и (14) появляются параметры a_j и a_i. Параметр a_j был введен А. Бирнбаумом для характеристики дифференцирующей способности задания при измерении различных значений θ. Параметр a_i указывает на меру структурированности знаний i-го ученика;

3) трехпараметрическую модель А. Бирнбаума

$$P_j\{x_{ij}=1|\beta_j\} = c_j + (1-c_j)\frac{e^{1,7a_j(\theta-\beta_j)}}{1+e^{1,7a_j(\theta-\beta_j)}}, \qquad (15)$$

где c_j является третьим параметром модели, характеризующим вероятность правильного ответа на j-е задание в том случае, если этот ответ угадан, а не основан на знаниях.

Северо-Восточный федеральный университет им. М.К. Аммосова

В каждой из представленных моделей параметры θ и β выражаются как шкалированные показатели единой для всех моделей шкалы логитов. Введение единой шкалы для элементов двух различных множеств позволяет подобрать оптимальные значения β, дающие возможность измерить искомое θ с минимальной ошибкой измерения. Перевод значений θ и β в общую шкалу логитов с помощью специальных преобразований был предложен Б. Д. Райтом и М. Г. Стоуном [83]. Подробно он рассмотрен в книге М. Б. Челышковой «Теория и практика конструирования педагогических тестов» [60].

Однопараметрическая модель Г. Раша. Аналитическое задание однопараметрической модели Г. Раша представлено формулами (16) и (12). В первом случае вероятность правильного выполнения *j*-го задания теста является возрастающей функцией от переменной θ. Это свойство функции согласуется с практическим опытом педагога. Естественно ожидать, что чем больше уровень подготовки испытуемого, тем больше вероятность правильного выполнения им *j*-го задания теста.

На рис. 27 изображена характеристическая кривая *j*-го задания теста, показывающая взаимосвязь между значениями независимой переменной θ и величиной P_j. Точке перегиба характеристической кривой соответствует значение $\theta = \beta_j$, а P_j в этой точке равно 0,5.

Свойство инвариантности оценок параметра испытуемых от трудности заданий теста. Модель Раша обладает интересным свойством, позволяющим на репрезентативной выборке испытуемых реализовать идею инвариантности оценок параметров θ и β, которая не характерна для двух и трех параметрических моделей. Не останавливаясь на математическом доказательстве, можно привести несложную геометрическую интерпретацию свойства инвариантности (см. рис. 28).

Пусть испытуемый с уровнем подготовки θ_1 ответит на задание *j* с вероятностью P_1. Увеличение трудности *j*-го задания теста на константу *c* ($c > 0$) вызовет смещение характеристической кри-

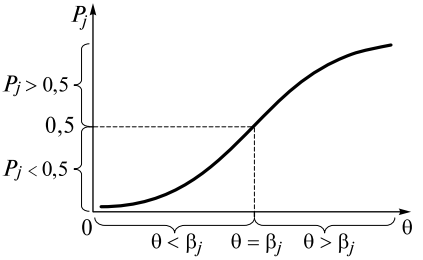

Рис. 27. Характеристическая кривая *j*-го задания теста

ФГБОУ ВО «Восточный университет им. М. Ауэзова»

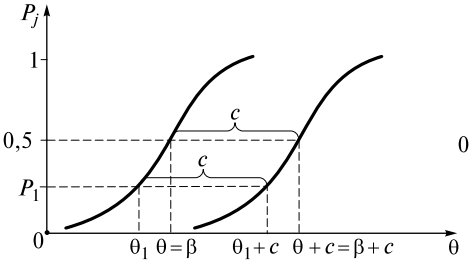

Рис. 28. Иллюстрация инвариантности оценок уровня подготовки испытуемых от трудности заданий теста

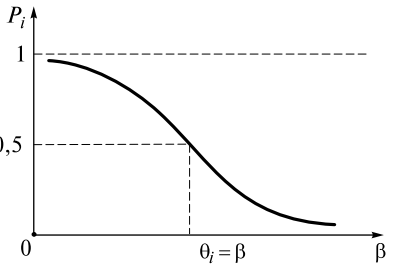

Рис. 29. Индивидуальная кривая *i*-го испытуемого

вой вправо. С прежней вероятностью на это более трудное задание будет отвечать испытуемый с уровнем подготовки $\theta_1 + c$. Так как $\theta - \beta_j = (\theta + c) - (\beta_j + c)$, то значение вероятности правильного ответа P_1 не изменится, что дает основание для вывода об относительной инвариантности уровня подготовки испытуемых от трудности заданий теста.

Вероятность правильного выполнения *i*-м испытуемым различных по трудности заданий P_i является убывающей функцией переменной β. Это означает, что с ростом трудности заданий значения вероятности $P_i(\beta)$ будут уменьшаться. График функции $P_i(\beta)$ представлен на рис. 29.

В точке перегиба кривой, соответствующей значению независимой переменной $\theta_i = \beta$, функция $P_i(\beta)$ принимает значение $P_i = 0{,}5$. В процессе обучения по мере накопления знаний индивидуальная кривая испытуемого смещается вправо.

Поскольку вдоль кривой откладываются доли правильных ответов на задания, которые не зависят от характера распределения

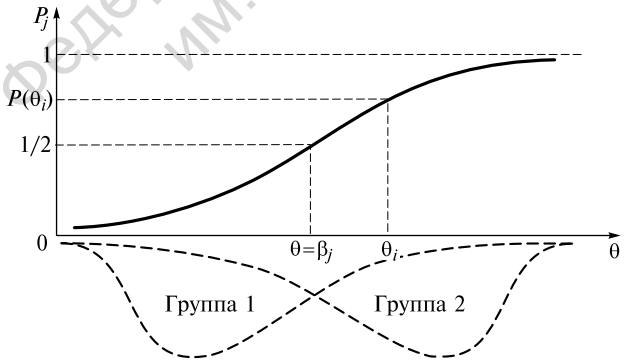

Рис. 30. Иллюстрация инвариантности формы характеристической кривой задания от уровня подготовленности тестируемой выборки

Сибирский государственный федеральный университет им. М.Ф. Решетнева

группы тестируемых учеников, форма характеристической кривой задания и ее положение при построении кривой на выборках в первой слабой и во второй сильной группах получатся одними и теми же (рис. 30). Конечно, практика свидетельствует о том, что эффект инвариантности наблюдается далеко не всегда, а только в тех случаях, когда реальная статистика — доли правильных ответов учащихся на задания — лежит достаточно близко к теоретической кривой. Причем чем ближе подходят точки распределения долей к кривой — графику функции P_j, тем ярче проявляется инвариантность.

10.3. Оценивание параметров подготовленности учащихся и трудности заданий теста в IRT

Начальные оценки параметра испытуемых и параметра трудности заданий теста. Начальные оценки параметра подготовки учащихся в логитах находят по формуле

$$\theta_i^0 = \ln \frac{p_i}{q_i},$$

где θ_i^0 — уровень подготовленности i-го ученика; p_i и q_i — доли правильных и неправильных ответов соответственно, подсчитанные по матрице наблюдаемых результатов выполнения теста; ln — символ натурального логарифма.

Начальные оценки параметра трудности заданий β получают по формуле

$$\beta_j^0 = \ln \frac{q_j}{p_j},$$

где p_j и q_j — доли правильных и неправильных ответов на j-е задание теста, соответственно; ln — символ натурального логарифма. Как следует из формул для подсчета, оценки параметров θ и β могут меняться в интервале $(-\infty, +\infty)$, но практически при $\theta_i - \beta_j < -5$ значения P_{ij} близки к нулю. Аналогичная пограничная ситуация наблюдается, когда $\theta_i - \beta_j > 5$. В этом случае значения вероятности P_{ij} будут почти равны 1. Поэтому значения разностей параметров, выходящие за указанные пределы, не рассматриваются в практике педагогических измерений.

Свести оценки логитов подготовленности испытуемых и логитов трудности заданий в единую шкалу позволяют специальные преобразования, выполняемые после завершения подсчетов начальных оценок θ и β. Вслед за преобразованиями оценки каждого из параметров выражаются в интервальной шкале с одним значением среднего и стандартного отклонения.

Северо-Восточный федеральный университет им. М.К.Аммосова

Метод наибольшего правдоподобия. Хотя теория IRT обеспечивает инвариантность оценок параметров θ и β, на практике в силу действия различных случайных факторов свойство инвариантности не выполняется в полной мере. Если объем выборки испытуемых достаточно велик, то можно ставить вопрос о вычислении устойчивых значений параметров θ и β, которые будут наиболее эффективными оценками и могут быть приняты в качестве объективных значений этих параметров.

Существуют различные методы вычисления эффективных оценок параметров распределения. Одним из них является метод наибольшего правдоподобия, предложенный Р. Фишером [63] и реализуемый итерационными процедурами с помощью специальных программных продуктов на ПК. Применение метода наибольшего правдоподобия требует введения предположения о локальной независимости заданий теста, которое означает, что при данном значении θ вероятность правильного ответа на конкретное задание теста не зависит от результатов выполнения остальных его заданий.

Для вычисления эффективных оценок параметров составляется вероятностная модель выполнения заданий теста группой испытуемых, которая называется функцией правдоподобия. Значения параметра θ, при которых функция правдоподобия достигает максимума, принимается в качестве объективных оценок параметра подготовленности испытуемых. Таким же образом вычисляются оценки наибольшего правдоподобия для параметра трудности заданий теста. Согласно теории оценки наибольшего правдоподобия являются наиболее эффективными и могут быть приняты за истинные значения латентных переменных θ и β.

Северо-Кавказский федеральный университет им. М.К. Аметова

10.4. Информационные функции тестовых заданий и теста

Понятие «информационная функция». В отличие от классической теории тестов, не позволяющей прогнозировать надежность измерений, в IRT можно априорно получать дифференцированные оценки точности, обеспечиваемой *j*-м заданием теста в различных точках оси θ. Эти оценки основаны на подсчете значений информационной функции, введенной А. Бирнбаумом. По одному из определений, предложенных этим исследователем, количество информации, обеспеченное *j*-м заданием теста в данной точке θ, — это величина, обратно пропорциональная стандартной ошибке измерения данного значения θ с помощью задания *j*.

Соответствие количества информации, получаемой при оценивании параметра θ с помощью задания *j*, и различных точек оси θ отражается с помощью специальной функции, получившей название информационной. Значения этой функции являются своеобразной характеристикой эффективности *j*-го задания в каждой

точке оси латентной переменной θ. Чем больше количество информации, тем лучше, образно говоря, работает задание на рассматриваемом интервале оси θ.

Информационная кривая теста. Благодаря свойству аддитивности информация, полученная при измерении данного θ с помощью всего теста, складывается из отдельных значений ординат информационных функций, построенных для каждого задания теста. На рис. 31 приведены графики трех информационных функций, одна из которых (кривая *1*) имеет две точки максимума θ_1 и θ_2, что недопустимо в правильно сконструированном тесте.

Кривая *2* на том же рисунке принадлежит менее информативному тесту, проигрывающему в точности измерений тесту, представленному кривой *1*, при оценке подготовленности учащихся в окрестности точки θ_2. Однако у кривой *2* есть явное преимущество по сравнению с кривой *1*, поскольку она имеет один четко выраженный максимум, что позволяет отдать ей предпочтение при сравнительном анализе качества первого и второго тестов. Пологая кривая *3* отображает неудачный тест, который является малоинформативным на всем протяжении оси θ.

Моделирование теста. В тех случаях, когда есть банк калиброванных заданий, тест можно моделировать с помощью информационных функций, построенных на том участке оси θ, где по предварительным данным (экспресс-диагностика или опыт предварительного контроля учащихся) будут в основном расположены оценки подготовленности испытуемых. За счет специального подбора заданий на основе графика целевой информационной функции теста появляется возможность оптимизировать подбор трудности теста и минимизировать стандартную ошибку измерения на нужном интервале оси θ.

В целом процесс моделирования теста, представленный на рис. 32, включает следующие этапы:

– построение целевой информационной кривой теста, обеспечивающей заданную стандартную ошибку измерения в нужном интервале оси θ;

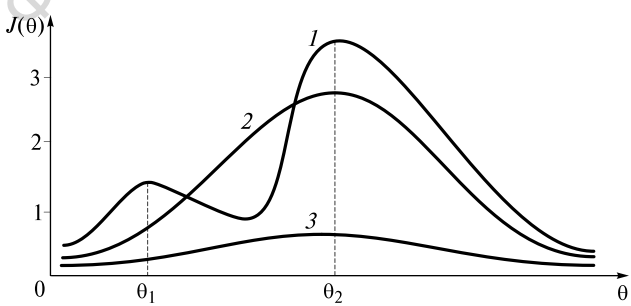

Рис. 31. Информационные кривые трех тестов

Северо-Восточный федеральный университет им. М. К. Аммосова

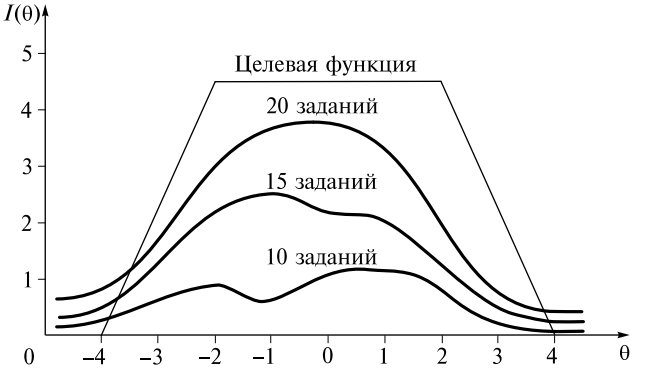

Рис. 32. Информационные кривые моделируемого теста

– выбор заданий из банка с информационными кривыми, удовлетворительно заполняющими пространство под целевой информационной кривой теста;

– сложение ординат информационных кривых тестовых заданий в каждой точке оси латентной переменной θ и пошаговое построение информационной кривой получаемого теста;

– продолжение процесса выбора заданий до тех пор, пока площадь под целевой кривой не будет заполнена с заданной степенью точности;

– проверка абсолютного значения разности между максимальной суммой ординат информационных кривых заданий и планируемым максимумом на целевой кривой в разных точках оси θ.

Имея информационные функции, можно сравнить эффективность различных моделируемых тестов с исходным эталонным без предварительного сбора эмпирических данных. Для этого используют функцию сравнительной эффективности, представляющую собой отношение двух информационных функций: эталонной функции и функции моделируемого теста. Вычисление значений функции сравнительной эффективности позволяет оценить эффект при удалении из теста заданий определенной трудности, при замене заданий средней трудности на легкие или более трудные задания, а также решить ряд других вопросов, возникающих у создателя тестов.

10.5. Современные программные средства для разработки педагогических тестов

Программные средства для конструирования тестов: общая характеристика. Интенсивное развитие программных средств, реализующих алгоритмы IRT, и классической теории для конструирования новых тестов началось в конце 80-х гг. XX в. и продолжа-

Полесский государственный университет

ется в настоящее время. Программные продукты для конструирования тестов нередко путают с инструментальными средствами для компьютерного тестирования, хотя они имеют разное назначение. Первые создаются для анализа эмпирических данных тестирования в целях коррекции характеристик тестов, обеспечения высокого качества педагогических измерений, калибровки заданий при наполнении банков, шкалирования и выравнивания для построения стандартных шкал по данным педагогических измерений. Вторые выполняют исключительно функцию поддержки при проведении компьютерного тестирования и обеспечивают формирование вариантов тестов, их предъявление, накопление баз данных по результатам тестирования и оценку результатов учащихся для выдачи им тестового балла. При правильном положении вещей оба блока программных продуктов должны работать совместно, поскольку информацию о результатах тестирования, накапливаемую в инструментальной оболочке, необходимо передавать дальше для совершенствования характеристик теста.

Часть разрабатываемых в мире программ для конструирования тестов носит закрытый характер и используется исключительно для собственных нужд, как, например, в ETS. Другая часть попадает на рынок благодаря каталогам и Интернету. Материалы, размещенные в Интернете крупнейшими разработчиками и распространителями программного обеспечения для конструирования тестов, включают описания технологий, перечень возможностей программ, демонстрационные версии и т.д.

Виды программных продуктов для конструирования тестов. К числу наиболее интересных программ, созданных мировым лидером в компьютерном тестировании Assessment Systems Corporation (ASC), можно отнести RASCH, RASCAL, Quest, ConQuest, а также XCALIBRE, ASCAL, LOGIMO, MSP, PARELLA и многие другие [91]. Некоторые из разработок корпорации ASC, например, MieroCAT, CAT, позволяют реализовывать адаптивные варьирующие алгоритмы с переменным шагом и осуществлять процессы генерации адаптивных тестов. В настоящий момент наибольший интерес для разработчика и пользователя тестов представляют следующие программы:

– AGREE, предназначенная для расчета согласованности оценок в номинальных шкалах данных в случае, когда два и более экспертов классифицируют объекты по некоторым категориям номинальной шкалы;

– ASC Item and Test Analysis Package — пакет программ ASC для анализа качества тестов и шкалирования результатов их выполнения. Программы пакета позволяют осуществить простейший анализ качества заданий по классической теории тестов, в том числе и углубленный дистракторный анализ (ITEMAN), получить оценки параметров подготовленности испытуемых и трудности

заданий по однопараметрической модели теории IRT (RASCAL), провести калибровку заданий (XCALIBRE), получить данные входного тестирования для эвалюации (Test Pre-Evaluation), оценить валидность теста (Test Validation), провести автоматизированную оценку результатов выполнения заданий и получить шкалированные баллы учащихся (Test Scoring) и т.д.;

– BILOG^{NEW}, позволяющая получить оценки параметров тестовых заданий на основе теории IRT использованием одно-, двух- и трехпараметрических логистических моделей и перейти к наиболее эффективным оценкам параметров по методу максимального подобия с помощью реализации итерационных процессов (только для дихотомических оценок по отдельным заданиям теста);

– ConQuest — универсальная программа, включающая различные модели IRT, в том числе для политомических оценок по заданиям в случае одномерных и многомерных измерений;

– FastTEST^{NEW} — 32-bit Windows система для поддержки банка калиброванных тестовых заданий и моделирования различных тестов, поддерживающая режимы бланкового и компьютерного предъявления параллельных вариантов тестов и оценки их качества с помощью моделей IRT для дихотомических данных по заданиям;

– FastTEST Professional^{NEW} — система для адаптивного тестирования с использованием теории IRT. Считается наиболее продвинутой системой компьютерного адаптивного тестирования в мире;

– ITEMAN — статистика заданий и тестов с использованием классической теории тестов и данных опросов (по типу Лайкерта). Наиболее популярный в мире программный продукт для анализа тестов;

– MicroCAT — полная оболочка для создания тестов и проведения тестирования (как компьютерного, так и бланкового) и последующего анализа результатов;

– PARSCALE^{NEW} — предназначена для шкалирования результатов учащихся на основе теории IRT, позволяет проводить анализ качества заданий и построение рейтинговых шкал;

– XCALIBRE, позволяющая получить оценки наибольшего правдоподобия на основе алгоритмов EM для небольших выборок испытуемых или коротких тестов для двух- и трехпараметрических моделей IRT.

Практические задания

1. Укажите два наиболее существенных недостатка классической теории тестов, которые вынудили специалистов обратиться к созданию современной теории конструирования тестов.

2. В представленной ниже таблице даны оценки параметров шести заданий, полученных по современной теории конструирования тестов.

Северо-Восточный федеральный университет им. М.К.Аммосова

Задание	Трудность (β)	Дифференцирующая способность (*a*)	Угадывание (*c*)
1	1,0	1,8	0,00
2	1,0	0,7	0,00
3	1,0	1,8	0,25
4	–0,5	1,2	0,20
5	0,5	1,2	0,00
6	0,0	0,5	0,10

Для каждого задания вычислите по трехпараметрической модели вероятность $P(\theta)$ для $\theta = -3; -2; -1; 0; 1; 2; 3$. Постройте графики характеристических кривых шести заданий по одно-, двух- и трехпараметрической модели.

Какое задание из представленных в таблице самое легкое?

Какое задание из шести имеет наименьшую дифференцирующую способность, валидность?

Какое из шести заданий для испытуемого с уровнем подготовленности $\theta = 0$ имеет наибольшую вероятность правильного ответа?

3. Докажите, что для трехпараметрической модели вероятность правильного выполнения задания P при $\theta = \beta$ будет равна $P = \frac{1+c}{2}$.

4. Вероятность правильного ответа $P(\theta)$ при заданных величинах θ (верхняя строка) для трех заданий приведена в таблице. Постройте характеристические кривые трех заданий.

№	-3,0	-2,5	-2,0	-1,5	-1,0	-0,5	0	0,5	1,0	1,5	2,0	2,5	3,0
1	0,0	0,0	0,0	0,0	0,0	0,1	0,2	0,3	0,5	0,6	0,7	0,8	0,9
2	1	1	2	4	7	3	2	5	0	5	8	7	3
3	0,0	0,0	0,0	0,0	0,1	0,2	0,5	0,7	0,8	0,9	0,9	0,9	0,9

Для заданий № 1 и 2 $c = 0$. Определите с помощью графиков характеристических кривых величины β — оценки трудности этих двух заданий.

Как можно сопоставить величины параметра *a* из графиков характеристических кривых трех заданий?

Северо-Восточный федеральный университет им. М.К. Аммосова

ГЛАВА

11

ОЦЕНИВАНИЕ НАДЕЖНОСТИ И ВАЛИДНОСТИ ПЕДАГОГИЧЕСКИХ ТЕСТОВ

11.1. Оценивание надежности ретестовым методом (двукратное тестирование)

Общие замечания о надежности и методах ее оценивания. Вводные представления о надежности педагогических измерений были изложены в разделе 4.5, содержащем определение, концептуальное обоснование и анализ надежности в контексте ее связи с дисперсией наблюдаемых баллов и ошибкой измерения. В том же разделе приведено определение коэффициента надежности теста — $r_н$, позволяющее сделать некоторые выводы о факторах, влияющих на ее величину.

Оценка надежности нормативно-ориентированных тестов проводится различными методами, которые по способу осуществления можно условно разделить на две группы [1; 60]. Первая группа методов базируется на двукратном тестировании, проводимом с помощью одного и того же теста или с помощью двух параллельных форм теста. Вторая группа методов предполагает однократное тестирование при оценке надежности теста. На практике стараются использовать вторую группу методов, поскольку организация повторного тестирования и разработка параллельных форм всегда сопряжены с определенными трудностями и дополнительными затратами со стороны создателей тестов. Обычно вне зависимости от метода оценка надежности строится на подсчете корреляции между двумя наборами данных. Логика рассуждений при этом довольно проста: чем выше корреляция, тем надежнее тест.

Для маленькой выборки корреляцию можно оценить визуально (табл. 13). В рассматриваемом гипотетическом примере три теста *A*, *B* и *C* из 10 заданий дважды выполняла одна и та же выборка из 10 учеников.

Тест *A* обладает оптимальной надежностью, так как результаты 10 учеников остались прежними: баллы и места учеников не изменились после повторного выполнения теста. Подсчет корреляции результатов первого и второго тестирования даст коэффициент корреляции, равный единице. Тест *B* полностью ненадежен: тот, кто имел самые высокие баллы в первом тестировании, получает самые низкие баллы во втором тестировании после по-

Полесский государственный университет им. П.М. Машерова

Таблица 13

Результаты двукратного выполнения трех тестов

Номер ученика	Тест *A*		Тест *B*		Тест *C*	
	1-е тестирование	2-е тестирование	1-е тестирование	2-е тестирование	1-е тестирование	2-е тестирование
1	10	10	10	1	10	6
2	9	9	9	2	9	4
3	8	8	8	3	8	8
4	7	7	7	4	7	9
5	6	6	6	5	6	3
6	5	5	5	6	5	1
7	4	4	4	7	4	5
8	3	3	3	8	3	7
9	2	2	2	9	2	2
10	1	1	1	10	1	10

вторного применения этого же теста. Полное отсутствие воспроизводимости баллов испытуемых указывает на минимальную надежность теста, близкую к –1. Тест *C* обеспечивает в целом хаотичное изменение результатов, хотя баллы отдельных учеников (3-го и 9-го) будут воспроизведены при повторном выполнении теста. Скорее всего, надежность третьего теста близка к нулю.

Естественно, что рассмотренные гипотетические ситуации не встречаются в практике. Обычно коэффициент надежности принимает положительные значения, но никогда не бывает равен единице. Это относится даже к существующим десятилетия тестам, получившим всеобщее признание.

Подсчет коэффициента надежности. Ретестовый метод оценки надежности (test-retest reliability) основан на подсчете корреляции индивидуальных баллов испытуемых, полученных в результате двукратного выполнения ими одного и того же теста. Обычно повторное тестирование проводится через 1 — 2 недели, когда испытуемые еще не успели забыть учебный материал и незначительно продвинулись в усвоении новых знаний. При таких условиях повторного предъявления теста низкая корреляция между результатами тестирования будет следствием не изменения состояния испытуемых, а применения ненадежного теста.

Для подсчета коэффициента надежности по методу повторного тестирования используется формула

Северо-Восточный федеральный университет имени М.К. Аммосова

$$(r_{\text{н}})_{\text{рет}} = \frac{N\sum_{i=1}^{N} X_i Y_i - \left(\sum_{i=1}^{N} X_i\right)\left(\sum_{i=1}^{N} Y_i\right)}{\sqrt{N\sum_{i=1}^{N}(X_i)^2 - \left(\sum_{i=1}^{N} X_i\right)^2}\sqrt{N\sum_{i=1}^{N}(Y_i)^2 - \left(\sum_{i=1}^{N} Y_i\right)^2}}, \qquad (16)$$

где $(r_{\text{н}})_{\text{рет}}$ — коэффициент надежности теста по ретестовому методу; X_i — индивидуальный балл i-го испытуемого в первом тестировании; Y_i — индивидуальный балл i-го испытуемого во втором тестировании (i = 1, 2, ..., N).

Пример подсчета. Используя данные табл. 9 (первое тестирование) и добавляя к ним гипотетические данные второго тестирования, можно с помощью табл. 14 подсчитать коэффициент надежности ретестовым методом.

После подстановки чисел из нижней строчки таблицы в формулу (16) коэффициент надежности будет равен $(r_{\text{н}})_{\text{рет}} =$

$$= \frac{10 \cdot 309 - 50 \cdot 55}{\sqrt{10 \cdot 312 - 50^2}\sqrt{10 \cdot 337 - 55^2}} = \frac{340}{\sqrt{620}\sqrt{345}} \approx 0,78.$$ Значение $r_{\text{н}} = 0,78$ указывает на невысокую надежность теста.

Применение ретестового метода может привести к ошибочным оценкам надежности в тех случаях, когда проводится слишком близкое по времени повторное применение теста. Учащиеся запоминают ответы к заданиям и при повторном тестировании значительно повышают свои результаты, что искажает оценку надежности теста.

Таблица 14

Пример данных для оценки надежности

Номер ученика i	Балл при первом тестировании X_i	Балл при втором тестировании Y_i	X_iY_i	$(X_i)^2$	$(Y_i)^2$
1	6	5	30	36	25
2	2	4	8	4	16
3	1	2	2	1	4
4	9	7	63	81	49
5	4	6	24	16	36
6	4	3	12	16	9
7	5	7	35	25	49
8	4	6	24	16	36
9	9	7	63	81	49
10	6	8	48	36	64
	$\Sigma X_i = 50$	$\Sigma Y_i = 55$	$\Sigma X_iY_i = 309$	$\Sigma(X_i)^2 = 312$	$\Sigma(Y_i)^2 = 337$

Северо-Осетинский федеральный университет им. К.Л. Хетагурова

11.2. Метод параллельных форм

Метод параллельных форм (parallel-form reliability) малоэффективен в тех случаях, когда при тестировании используется один вариант теста. В некоторых странах, например в США, благодаря соблюдению всех требований к проведению тестирования применение единственного варианта не снижает необходимый уровень информационной безопасности и обеспечивает при этом высокую сопоставимость результатов выполнения теста. Если тест только один, то для оценки надежности методом параллельных форм приходится создавать параллельный вариант теста, затем с затратами сил, средств и времени на апробацию доказывать правомерность гипотезы о параллельности и только потом оценивать надежность исходного теста.

Если параллельные варианты теста разрабатываются изначально, как в ЕГЭ, оценка надежности методом параллельных форм также требует значительных трудозатрат. Необходима тщательная ротация вариантов в группе испытуемых для обеспечения сходных выборок учащихся на параллельных вариантах теста. Даже при стратификации выборки испытуемых и ротации вариантов достоверность оценок надежности снижается из-за того, что параллельные формы — это скорее теория, чем реальность, поскольку на практике, несмотря на все усилия авторов, как правило, обнаруживаются статистически значимые различия в характеристиках параллельных вариантов. Для оценки надежности методом параллельных форм используется формула (16). В ней X_i ($i = 1, 2, ..., N$) — индивидуальные баллы испытуемых в первой форме, а Y_i ($i = 1, 2, ..., N$) — индивидуальные баллы во второй форме. Далее все вычисления с точностью повторяют подробно рассмотренный пример (см. табл. 9).

Сибирский государственный федеральный университет им. М.К. Аммосова

11.3. Метод расщепления теста (однократное тестирование)

Описание метода. Метод оценивания надежности, основанный на расщеплении результатов по тесту на две части (split-half method), наиболее распространен из-за своего удобства. Он позволяет вычислить коэффициент надежности при однократном выполнении учениками теста. Для оценки надежности результаты тестирования делят на две части: в одну включают данные испытуемых по четным, а в другую — по нечетным заданиям, считая при этом, что получены сходные по содержанию части теста. Правда, деление на две части не единственный способ, возможны и другие варианты, когда выделяют большее число частей при оценке надежности теста.

Таблица 15

Сводная таблица для оценки надежности (метод расщепления)

Номер ученика i	Баллы по четным заданиям X_i	Баллы по нечетным заданиям Y_i	X_iY_i	$(X_i)^2$	$(Y_i)^2$
1	X_1	Y_1	X_1Y_1	$(X_1)^2$	$(Y_1)^2$
2	X_2	Y_2	X_2Y_2	$(X_2)^2$	$(Y_2)^2$
...	...	...	...	...	...
N	X_N	Y_N	X_NY_N	$(X_N)^2$	$(Y_N)^2$
	$\sum_{i=1}^{N} X_i$	$\sum_{i=1}^{N} Y_i$	$\sum_{i=1}^{N} X_iY_i$	$\sum_{i=1}^{N} (X_i)^2$	$\sum_{i=1}^{N} (Y_i)^2$

Подсчет коэффициента надежности. Для оценивания надежности методом расщепления результаты учеников заносят в табл. 15.

Далее для таблицы данных используют формулу (16), в которой роль результатов в первом тестировании выполняют данные по четным, а во втором — по нечетным заданиям. Использование метода расщепления дает заниженные оценки надежности в силу того, что она оценивается для укороченного в 2 раза теста.

Коррекция коэффициента надежности. Для коррекции оценки надежности в соответствии с длиной исходного теста используется формула Спирмена — Брауна $r_{\text{н}} = \dfrac{2(r_{\text{н}})_{\text{расщ}}}{1+(r_{\text{н}})_{\text{расщ}}}$, где в числителе и знаменателе дроби стоит коэффициент надежности для половины заданий теста, а слева — скорректированный коэффициент надежности с учетом всех заданий теста.

Приведенный метод оценивания надежности имеет свои ограничения в применении. Он основан на допущении параллельности двух половин теста, что не всегда и не в полной мере может оказаться верным. Корреляция двух половин возрастает по мере роста гомогенности теста. В этой связи метод расщепления нередко называют методом оценки внутренней состоятельности (согласованности) теста (Internal-Consistency Method).

11.4. Метод Кьюдера — Ричардсона (для дихотомических оценок по заданиям теста)

Описание метода. Метод Кьюдера — Ричардсона для оценки надежности, так же как и метод расщепления теста, основан на

Северо-Восточный федеральный университет им. М.К. Аммосова

однократном тестировании, но в отличие от него не зависит от искусственных допущений о полной параллельности двух частей теста. Однако сфера его применения ограниченна, так как он годится лишь при использовании дихотомических оценок по результатам выполнения заданий гомогенных тестов.

Формула Кьюдера — Ричардсона. Формула Кьюдера — Ричардсона (KR-20) имеет следующий вид:

$$(r_{\text{н}})_{\text{KR-20}} = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} p_j q_j}{S_X^2}\right), \qquad (17)$$

где p_j — доля правильных ответов на j-е задание; q_j — доля неправильных ответов, $q_j = 1 - p_j$; S_X^2 — дисперсия по распределению наблюдаемых баллов; n — число заданий теста [87].

Для матрицы данных, представленных в табл. 9, подсчитанная ранее исправленная дисперсия равна $S_X^2 = 6{,}89$, а доли правильных ответов получаются делением чисел R_j в последней строке матрицы на 10. Сумма произведений долей правильных и неправильных ответов в таком случае будет равна $0{,}9 \cdot 0{,}1 + 0{,}8 \cdot 0{,}2 + 0{,}7 \cdot 0{,}3 + 0{,}6 \cdot 0{,}4 + 0{,}5 \cdot 0{,}5 + 0{,}5 \cdot 0{,}5 + 0{,}3 \cdot 0{,}7 + 0{,}4 \cdot 0{,}6 + 0{,}2 \cdot 0{,}8 + 0{,}1 \cdot 0{,}9 = 1{,}9$,

а коэффициент надежности — $(r_{\text{н}})_{\text{KR-20}} = \frac{10}{10-1}\left(1 - \frac{1{,}9}{6{,}89}\right) \approx 0{,}79$.

Общие рекомендации по применению метода Кьюдера — Ричардсона. В целом при оценке надежности нельзя полагаться лишь на один показатель, поскольку каждый из них имеет свои ограничения, смещающие оценки надежности теста в сторону завышения или занижения. Для достоверной проверки качества теста следует учитывать несколько показателей надежности, подсчитанных по разным формулам. В данном пособии приведена лишь небольшая их часть.

В качестве нижнего предела допустимых значений надежности обычно выбирают 0,7. При более низком значении использование теста вряд ли целесообразно в силу большой погрешности измерения. Если тест разрабатывается профессионалами, то к нему предъявляют более жесткие требования. Как правило, тесты с надежностью менее 0,8 считаются непригодными в профессионально организованных службах и центрах тестирования. Значения коэффициента надежности, превышающие 0,9, свидетельствуют о высоком качестве теста. Они желательны, но встречаются редко. Обычно в тестологической практике надежность тестов колеблется в интервале (0,8; 0,9).

Северо-Восточный федеральный университет имени М.К. Аммосова

11.5. Надежность и стандартная ошибка измерения

Связь между стандартной ошибкой измерения и надежностью теста. Один из аспектов применения коэффициента надежности связан с определением стандартной ошибки измерения. Для установления связи между стандартной ошибкой измерения и надежностью теста необходимо преобразовать формулу (1) для коэффициента надежности теста, выделив в левой части S_E^2. После преобразования формулы относительно S_E^2 получится выражение

$$S_E = S_X\sqrt{1 - r_н},$$

где S_X — стандартное отклонение по распределению индивидуальных баллов; $r_н$ — коэффициент надежности теста; S_E — стандартная ошибка измерения. Это выражение обычно используется для вычисления S_E по известным величинам $r_н$ и S_X.

Для лучшего уяснения смысла показателя S_E можно представить гипотетическую ситуацию, когда i-й испытуемый выполнял много раз один и тот же тест. Если предположить, что эффект запоминания отсутствует, то результаты тестирования образуют нормальное распределение вокруг истинного балла T_i со стандартным отклонением S_E. На практике S_E рассматривается как статистическая величина, отражающая степень точности отдельных измерений, поэтому величину S_E используют для определения границ доверительного интервала, внутри которого должен находиться истинный балл оцениваемого ученика группы.

Построение доверительного интервала. Общераспространен подход, когда доверительный интервал выстраивается вокруг наблюдаемого показателя ученика как две симметричные окрестности (левая и правая), хотя это не совсем верно, поскольку речь должна идти об окрестностях, расположенных слева и справа от истинного балла. Тем не менее этот факт вынужденно игнорируется в прикладных исследованиях в силу отсутствия истинного балла, и доверительный интервал при заданном риске допустить ошибку $t = 0{,}05$ (в пяти случаях из ста) принимается равным ($X_i - 1{,}96S_E$; $X_i + 1{,}96S_E$), где X_i — наблюдаемый балл i-го испытуемого; 1,96 — константа, табличное число, используемое при $t = 0{,}05$.

Численный пример. Для рассматриваемого ранее примера матрицы тестовых результатов (см. табл. 9), коэффициента надежности $r_н = 0{,}78$ и стандартного отклонения $S_X = 2{,}62$, вычисленного ранее для матрицы, S_E будет равно $S_E = 2{,}62\sqrt{1 - 0{,}78} \approx 1{,}23$. В данном случае доверительный интервал для истинного балла первого ученика со значением $X_1 = 6$ будет составлять (6 – 1,23; 6 + 1,23) или (4,77; 7,23). Истинный балл первого ученика может находиться в любой точке этого интервала.

Очевидно, что с ростом S_E границы доверительного интервала будут раздвигаться, и вместе с тем будут увеличиваться возможные пределы отклонения истинного балла от наблюдаемых результатов измерения (величина отклонения наблюдаемых баллов от истинной компоненты измерения).

Предсказание истинных баллов на основе регрессионной модели. Методы регрессионного анализа позволяют прогнозировать оценки истинных баллов испытуемых по распределению наблюдаемых баллов и коэффициенту надежности теста. Прогноз получается путем подстановки в регрессионное уравнение

$$T_i = \bar{X} + r_{\text{н}} \left(X_i - \bar{X} \right),$$

где T_i — истинный балл; X_i — индивидуальный балл i-го испытуемого; $\bar{X}$ — среднее значение баллов испытуемых [60].

11.6. Валидность гомогенных тестов

Общие замечания. Валидность педагогических измерений рассматривалась ранее в разделе 4.5. Как правило, постановка целей создания теста носит комплексный характер, поэтому часто валидность стараются проверить с разных позиций сообразно различным направлениям использования теста. Например, нормативно-ориентированный тест для приема абитуриентов в вузы должен служить цели дифференциации испытуемых и прогностическим целям, так как необходимо не только выделить лучших абитуриентов в момент приема, но и спрогнозировать успешность дальнейшего обучения зачисленных в вузы абитуриентов.

Критерии для оценки валидности. Как было отмечено ранее, оценивание валидности всегда проводится путем соотнесения характеристик результатов измерения с внешними критериями [1; 69; 86]. В качестве таких критериев могут выступать оценки экспертов при анализе содержания теста и его адекватности целям измерения (содержательная валидность), результатов по другим тестам (конструктная валидность), успешности дальнейшего обучения (прогностическая валидность).

Высокая корреляция между анализируемыми результатами испытуемых и внешними критериями подтверждает высокую валидность теста. Основная трудность при такой валидизации носит не практический, а методологический характер, поскольку она состоит в выборе значимого внешнего критерия.

Связь надежности и валидности. Для повышения полноты охвата содержания и роста содержательной валидности теста желательно отбирать задания с малыми коэффициентами интеркорреляции. К противоположному выводу легко прийти, если стараться

Полесский государственный университет

повысить надежность теста. Отбирая задания с большими коэффициентами интеркорреляции, можно обеспечить высокую однородность содержания и надежность теста. Это противоречие, получившее название «парадокс Ф. Лорда», приводит к возникновению серьезных проблем при конструировании теста.

Таким образом, при конструировании гомогенного теста следует стремиться к повышению в разумных пределах его надежности, чтобы не снизить существенным образом содержательную валидность теста. Поэтому при отборе заданий в тест необходимо иметь четкое представление об их содержании и о множестве других факторов, а не просто отдавать предпочтение тем заданиям, которые высоко коррелируют друг с другом и обеспечивают хорошую надежность теста. По мнению Кэттела и Клайна, максимум валидности может быть получен тогда, когда все задания слабо, но положительно коррелируют друг с другом, однако каждое из них имеет высокую корреляцию с критерием по тесту [26]. Поэтому повышению валидности способствует включение заданий, для которых характерны большие коэффициенты бисериальной корреляции с суммой баллов по тесту.

По рассматриваемой выше проблеме существует и другая точка зрения. Так, Гилфорд и Ньюнелли [26] полагают, что внутренняя согласованность теста — непременное условие его высокой содержательной валидности, и потому высокая надежность является предпосылкой оптимальной валидности теста.

Количественные оценки валидности. При количественных оценках валидности для педагогических тестов в качестве критерия обычно берутся оценки экспертов, выставленные ими при традиционной проверке знаний учеников без использования тестов. Процесс валидизации осложняется необходимостью установления меры согласованности оценок экспертов, которых обычно бывает не менее трех человек. Если мера согласованности достаточно высока, то для оценки валидности используется формула

$$r_{\text{в}} = \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(X_{m_i} - \bar{X}_{\text{э}}\right)}{N_m\sqrt{S_X^2 \cdot S_{m_x}^2}},$$

где $X_i - \bar{X}$ — отклонение тестового балла i-го ученика от среднего балла по тесту; $X_{m_i} - \bar{X}_{\text{э}}$ — отклонение балла i-го ученика у экспертов от $\bar{X}_{\text{э}}$ — среднего арифметического экспертных оценок; S_X^2 — дисперсия баллов учеников по тесту; $S_{m_x}^2$ — дисперсия баллов m-го эксперта; m — число экспертов.

Бывают случаи, когда педагог заинтересован в оценке прогностической валидности, указывающей меру вероятности прогноза успешности дальнейшего обучения по результатам выполнения теста. В этом случае результаты по тесту коррелируют с результата-

Гомельский государственный университет имени Ф. Скорины

ми поступивших абитуриентов после окончания первого года обучения в вузе. Высокая корреляция означает, что разработанные тесты для отбора абитуриентов в вуз прогностичны.

Источники повышения валидности теста. Для повышения содержательной валидности теста необходимы:

– подбор оптимальной трудности заданий;
– экспертиза качества содержания теста;
– расчет оптимального времени выполнения теста;
– подбор валидных заданий с высокой дискриминативностью.

Практические задания

1. Используя данные тестирования, полученные вами, рассчитайте ретестовую надежность теста.

Вычислите надежность теста по формуле KR-20.

2. Вычислите надежность теста методом расщепления с использованием формулы Спирмена — Брауна.

Северо-Восточный федеральный Университет им. М.К.Аммосова

ГЛАВА
12
ПОДГОТОВКА К ТЕСТИРОВАНИЮ, ПРОВЕДЕНИЕ ТЕСТИРОВАНИЯ И ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ

12.1. Подготовка к тестированию

Стандартизация условий и материалов. При тестировании не должны возникать непредвиденные обстоятельства, снижающие надежность результатов выполнения тестов. Поэтому на этапе подготовки следует максимально стандартизовать те условия, обеспечивающие единообразие процедуры тестирования, а также те, в которых выполняются тесты. Стандартизация процедуры тестирования требует разработки правильных инструкций для всех участников, выбора оптимального времени дня для проведения тестирования, создания подходящей окружающей обстановки и планирования размещения испытуемых для предотвращения списывания в процессе выполнения теста [1; 20; 28; 35; 60]. Важно продумать, как скомпоновать тестовые материалы для выдачи учащимся, заготовить специальные запасные материалы: бланки ответов, ручки или карандаши и т.д., чтобы не отвлекать учащихся от работы над тестом.

При проведении апробации теста необходимо предусмотреть формирование репрезентативной выборки испытуемых и обеспечить равномерную ротацию вариантов по различным стратам выборки. Для стратификации нужна предварительная информация об учащихся, которую нередко собирают с помощью анкетирования. До начала тестирования желательно провести репетиционное занятие, ознакомив учащихся с формами заданий и действиями по заполнению бланков ответов, особенно если учащиеся никогда ранее не выполняли тесты.

Требования к бланкам для ответов на задания теста. Бланк должен быть разработан таким образом, чтобы сделать его максимально понятным испытуемым и свести к минимуму время, затрачиваемое на поиск места для ответов. В заданиях с выбором ответа обязательно должен быть предложен символ, с помощью которого отмечается правильный ответ. Необходимо также определить возможные исправления, которые имеют право сделать учащиеся на бланке в случае изменения ответа.

Если регистрационная форма не прилагается отдельно, то в начале бланка ответов следует оставить место для данных учаще-

Северо-Восточный федеральный университет им. М.К. Аммосова

гося. В массовом тестировании, когда результаты учащихся используются для принятия административно-управленческих решений, бланк ответов должен быть зашифрован, а данные для идентификации учащегося должны указываться на отдельном регистрационном листе, как, например, при проведении ЕГЭ.

Условия, в которых должно проводиться тестирование. Предварительная подготовка к тестированию касается не только используемых материалов, но и окружающей обстановки. Необходимо заранее выбрать помещение для тестирования. Оно должно быть достаточно тихим, иметь хорошее освещение, вентиляцию и удобные рабочие места. При тестировании больших групп учащихся помещение должно обеспечивать экзаменаторам свободный доступ к посадочным местам учащихся, предусматривать возможность одноместной посадки испытуемых и исключать ситуации, удобные для списывания, когда у впереди или сбоку сидящего учащегося оказывается тот же самый вариант теста.

Если тестирование проводится в специализированных аудиториях (классы математики, физики и т. п.), то необходимо убрать (закрыть) стенды, плакаты и прочие материалы со справочной информацией по соответствующим дисциплинам. На двери обычно вывешиваются списки учащихся, находящихся в помещении для тестирования, и предупреждающие знаки, запрещающие входить в помещения при любых обстоятельствах, кроме чрезвычайных случаев. У дверей помещений для прекращения доступа в помещение опоздавшим учащимся после начала выполнения тестов необходимо поставить помощников.

Северо-Осетинский государственный университет им. М.К. Аммосова

12.2. Инструкции по тестированию и процедура его проведения

Требования к экзаменаторам и их помощникам. На тестировании даже в небольшой аудитории на 15—20 человек должно присутствовать два преподавателя. Необходимо, чтобы один из них преподавал проверяемый учебный предмет на тот случай, если в содержании тестов встретится ошибка. По сложившейся в ЕГЭ терминологии их называют организаторами по проведению тестирования, или экзаменаторами. Если в помещении для тестирования находятся примерно 50 учащихся, то экзаменаторов должно быть не менее четырех человек для поддержания дисциплины и минимизации потерь времени у испытуемых при возникновении вопросов.

Важно, чтобы тестирование проводил экзаменатор, который никогда не общался с учащимися. Экзаменатору нужно предварительно ознакомиться с текстом инструкций и усвоить специфические приемы для установления контактов с испытуемыми. Если тестируются дети младших классов, желательно расположить их к

себе дружеской манерой общения, поскольку для них характерна боязнь незнакомых людей. При тестировании школьников постарше лучше выбрать нейтральную форму общения. Например, в условиях ЕГЭ желательно напомнить о правилах поведения и о том, что подсказки товарищам снижают собственные шансы на поступление в престижный вуз, обратиться к свойственному выпускникам школ духу соревновательности, а в дальнейшем лишь поддерживать дисциплину и никак не вмешиваться в работу. В целом экзаменатору необходимо помнить о том, что малейшее отступление от требований стандартизации в поведении на тестировании повлечет за собой снижение объективности результатов выполнения теста.

Экзаменатора, проводящего тестирование, следует отличать от педагога, использующего результаты учащихся по тесту. Экзаменатор должен владеть правилами поведения на тестировании, психологическими приемами установления контакта с испытуемыми, хотя он сам может ничего не знать о тестировании. Хорошо подготовленный педагог-пользователь выбирает тесты, которые подходят для поставленных целей контроля. Он знаком с научной литературой по тестовой проблематике, способен оценить такие характеристики, как надежность и валидность теста. Педагог делает выводы и дает рекомендации только связав тестовые результаты с другой имеющей отношение к цели тестирования информацией об испытуемом, что позволяет избежать ошибочных заключений.

Инструкция для экзаменатора. Разработка корректных инструкций для педагога, руководящего процессом применения теста, и для учеников, выполняющих тест, имеет большое значение для повышения надежности измерений [1; 20; 60].

В инструкции для экзаменатора содержатся рекомендации по подготовке группы учащихся к выполнению теста, примерные обязанности педагога на этапе проведения тестирования, рекомендации по подготовке краткого отчета о выполненной процедуре предъявления теста, перечень вопросов, на которые учащимся можно давать ответы.

Инструкции для учащихся. Для подготовки группы учащихся к тестированию обычно разрабатывают две инструкции, одна из которых — развернутая, зачитываемая педагогом на репетиционном тестировании или раздаваемая заблаговременно, задолго до начала экзамена, а другая — краткая, выдаваемая вместе с тестом или непосредственно перед началом тестирования и лежащая на столе у каждого ученика.

В инструктировании перед тестированием, которое можно проводить за 2—3 дня или накануне экзамена, педагогу необходимо:

1) объяснить учащимся, зачем нужен тест, сообщить, как будут использованы его результаты;

2) объяснить, почему испытуемые должны приложить максимум усилий для выполнения теста, акцентировать внимание испытуемых на возможности проверки своих сил и подчеркнуть соревновательный мотив;

3) медленно, четким голосом прочесть инструкцию к тесту с примерами, если последние имеются;

4) дать возможность испытуемым потренироваться, решить самостоятельно одну или более задач-образцов. Проверить, правильно ли понята инструкция, проанализировав прямо на консультации результаты выполнения примеров заданий;

5) сообщить о временно́м ресурсе выполнения теста на экзамене, о правилах исправления допущенных ошибок, рассказать о том, к кому обращаться в случае возникновения вопросов, объяснить, на какие вопросы не следует ждать ответов.

В целом инструкции для предварительной подготовки к тестированию бывают довольно длинными и подробными, особенно в тех случаях, когда они предназначены для самостоятельной работы с тестом дома. Например, в инструкции к тесту для самоконтроля можно рассказать о целях работы над тестом, детально описать его содержание, дать краткий обзор процесса развития теста, объяснить стратегию выполнения заданий и правило подсчета баллов, привести таблицу для самооценки и сравнения результата учащегося с результатами других учеников.

Краткая инструкция для испытуемого, выдаваемая перед началом экзамена вместе с тестом, должна содержать в основном правила по заполнению регистрационного бланка и бланка для ответов. Она может иметь различный вид, который зависит от цели тестирования и формы заданий, содержащихся в тесте. Например, в ЕГЭ такая инструкция по выполнению КИМ, включающих три части — *A*, *B*, *C* — с заданиями различной формы, объясняет правила заполнения бланков ответов.

Северо-Восточный федеральный университет им. М.К.Аммосова

12.3. Подготовка учащихся, ее влияние на изменение результатов тестирования

Нужно ли готовить учащихся к тестированию? Ответ на этот вопрос можно получить с различных позиций: с позиции педагога, психолога и специалиста по теории педагогических измерений [1]. Часть учителей, мало знающих о возможностях тестов, скорее всего будут возражать против подготовки к тестированию, поскольку обычно подготовка ассоциируется у них с «натаскиванием». На самом деле такая точка зрения характерна для сторонников традиционного контроля, в котором при рассекречивании содержания предстоящей контрольной работы, состоящей из 5—7 заданий, нельзя говорить о достоверности результатов проверки.

Иначе складывается ситуация при подготовке к выполнению тестов, длина которых обычно колеблется от 30 заданий (в математике и физике) до 50—60 заданий в гуманитарных предметах. Очевидно, что учащиеся, выполнившие в процессе подготовки к тестированию 100—150 заданий, подобных тем, которые будут в тесте, вполне заслуживают высоких оценок, поскольку при таком объеме усвоенного учебного материала речь уже должна идти не о натаскивании, а о хороших знаниях по предмету.

Влияние подготовки на результаты выполнения теста. В целом исследования специалистов показали, что степень улучшения результатов тестирования зависит от способностей и знаний учащихся, количества и вида предварительных занятий и особенностей тестов. Наиболее полезной подготовка к тестированию оказалась для слабых учащихся, в то время как на результаты сильных учеников, обладающих достаточными знаниями для выполнения теста, она повлияла незначительно. Понятно, что степень положительного влияния предварительной подготовки на изменение результатов тестирования находится в прямой зависимости от тесноты связи между содержанием тренировочных заданий и содержанием теста. Ограничивается ли улучшение только результатами по конкретным заданиям, которые использовались при подготовке, или позитивное влияние предварительного тестирования распространяется на качество подготовки учащихся в целом, пока не известно ни педагогам-теоретикам, ни педагогам-практикам.

Конечно, не каждое содержание заданий теста может быть отражено в тренировочных заданиях. Когда речь идет о проверке алгоритмических навыков или репродуктивных умений, то нет ничего плохого в том, что учащийся повторит перед тестированием изученный материал и выполнит предварительно совокупность заданий, похожих на задания теста. Другое дело, когда в тесте предлагаются творческие задания, требующие от испытуемых смекалки и нестандартных решений. Тогда любая тренировка перед тестированием рассекретит содержание теста и сведет на нет все элементы творчества при его выполнении.

Психологи положительно относятся к тренировочному тестированию, поскольку установлено, что предварительная подготовка снижает тревожность в поведении испытуемых во время тестирования и мотивирует к выполнению теста. Слабые учащиеся наиболее подвержены фактору тревожности, который не мотивирует их как сильных испытуемых, а, наоборот, заставляет забыть то немногое, что было выучено при подготовке к экзамену. Поэтому точка зрения психологов во многом совпадает с мнением педагогов: тренировка перед тестированием приносит наибольшую пользу учащимся с недостаточным запасом знаний, она заставляет их поверить в свои силы и мобилизует к выполнению теста.

Северо-Восточный федеральный университет им. М.К. Аммосова

Специалисты по педагогическим измерениям также считают необходимым проводить тренировки перед тестированием, но их интересует не подготовка к содержанию контролируемых вопросов, а обучение учащихся работе с различными формами тестовых заданий, минимизация потерь времени при занесении ответов в бланки, умения исправить случайные описки в бланках, не обращаясь к экзаменатору, и т.д. У тестологов, как всегда, на первом плане находится проблема оценки истинного балла каждого испытуемого, поэтому репетиционное тестирование рассматривается ими с позиций минимизации ошибок измерения. Таким образом, небольшая ориентировка по содержанию тестов и несколько практических занятий по выполнению различных форм заданий и заполнению бланков ответов просто необходимы, если вы хотите повысить объективность результатов выполнения тестов.

12.4. Этические и социальные проблемы тестирования

Этические нормы и принципы тестирования. Этические нормы и принципы тестирования в образовании продиктованы кодексом профессиональной этики педагога [1]. Они являются частью совокупности общечеловеческих норм, регламентирующих поведение учителя на тестировании в соответствии с требованиями долга, профессиональной честности, принципами гуманизма и т.д.

Деятельность учителя, выступающего в качестве тестолога, предъявляет ряд специфических требований, вытекающих из особенностей его профессиональных функций в роли специалиста по педагогическим измерениям. В отличие от психолога учитель не является исследователем человеческой личности, в сферу его интересов не попадают скрытые стороны духовной жизни человека. Поэтому требования к этике поведения тестолога в образовании не такие серьезные, как в психодиагностике, где постоянно приходится думать о профессиональной этике в работе с обследуемыми. Однако и в образовании несоблюдение норм специальной этики при проведении тестирования и интерпретации его результатов может привести к нежеланию учащихся выполнять тесты, к непониманию и отсутствию поддержки тестирования со стороны коллег, особенно если результаты тестирования кажутся им необоснованными, затрагивающими их компетентность в преподавании.

Важное требование к деятельности тестолога в образовании — соблюдение *принципа конфиденциальности*, предполагающего неразглашение сведений о результатах тестирования без согласия ученика, поскольку тест может выявить такие пробелы в подготовке учащихся, которые он предпочел бы скрыть от своих одноклассников. Поэтому до начала тестирования учащихся следует

Московский государственный открытый педагогический университет им. М.А. Шолохова

проинформировать относительно сферы использования его тестовых баллов, объяснить, в какой форме они будут выдаваться и кто будет иметь к ним доступ.

Часто при проведении массового тестирования, например национальных экзаменов, в начале теста учащегося просят заполнить анкетные данные и сведения о родителях, которые позволяют использовать в дальнейшем результаты экзаменов при оценке качества образования. Запрос таких сведений регламентируется принципом осведомленного согласия и принципом соответствия. Сведения, которые собирают в анкетах, должны соответствовать целям тестирования, о которых следует рассказать ученикам, заверив их в неразглашении информации и объяснив уровни ее использования.

При тестировании учащихся следует принимать во внимание *принцип доступности*, связанный с правом учащегося на получение доступа к содержательной интерпретации тестовых результатов, анализу проблем и неудач в выполнении отдельных заданий теста. Такая же информация, но в интегрированной форме, необходима учителям в целях коррекции методов преподавания. Результаты тестирования должны быть представлены в доступном для понимания виде, они не должны содержать специальные термины и профессиональную лексику в соответствии с задачами повышения качества образования.

Принцип обоснованности и динамического отражения развития учащегося предполагает систематическое обновление данных о подготовленности учащихся, полученных с помощью тестов. Информация о результатах тестирования учащихся должна накапливаться длительное время, поскольку она может оказаться полезной для правильного понимания учителем особенностей развития каждого школьника, динамики прироста его знаний и индивидуальной работы с ним. Однако наличие прежних результатов тестирования не должно приводить к ошибочным выводам. Было бы нелепо, например, ссылаться на плохие результаты по чтению ученика в I классе, объясняя его отставание по математике в III классе.

Учебно-методический комплекс Университет им. Н.А. Минского

12.5. Интерпретация результатов педагогических тестов, использование результатов на различных уровнях управления качеством образования

Цели интерпретации, проблема корректности при анализе данных тестирования. Интерпретация данных тестирования может иметь различные цели и проводиться разными группами лиц, заинтересованных в использовании результатов выполнения тестов. Ее осуществляют тестологи, которые анализируют данные те-

стирования для коррекции теста. В итоговом контроле в целях повышения качества образования и выявления тенденций в изменении качества интерпретации подвергаются шкалированные баллы учащихся. В текущем контроле педагоги анализируют результаты тестирования для коррекции процесса обучения и диагностики причин отставания отдельных учеников и т.д.

Случаи неправильной или упрощенной интерпретации данных тестирования встречаются при анализе, проводимом без учета дополнительных факторов, например социально-экономических, значимо влияющих на результаты выполнения тестов. Наблюдаемые по данным тестирования отставания в итоговых результатах групп учащихся отдельных школ района могут быть следствием неучтенной слабой материально-технической оснащенности учебного процесса, низкой квалификации преподавательского состава, высокого уровня безработицы среди родителей учащихся, а также ошибок при формировании выборочной совокупности учеников для проведения анализа результатов. Результаты интерпретации, получаемые путем намеренного исключения из анализа данных тестирования результатов слабых учащихся, могут быть искусственно завышены.

Некорректная интерпретация данных тестирования в целом приводит к возникновению недоверия к возможностям тестов. Особенно это относится к тем педагогам и руководителям, чья деятельность незаслуженно получает негативные оценки.

Уровни интерпретации результатов тестирования. В тех случаях, когда рассматриваются результаты не тестирования в школе или отдельном классе, а национальных экзаменов типа ЕГЭ, можно выделить несколько уровней интерпретации результатов учащихся, определив права доступа каждой группы пользователей к различным видам информации. В частности к таким группам пользователей следует отнести:

– общество, учащихся и их родителей;

– педагогов, директоров школ;

– руководителей и работников органов управления образованием районного уровня;

– руководителей и работников органов управления образованием регионального (областного) уровня;

– руководителей и работников органов управления образованием федерального уровня.

Модели для анализа и интерпретации данных тестирования. Зарубежные специалисты по измерениям в образовании полагают, что корректное использование результатов тестирования и повышение справедливости управленческих оценок достигаются на основе применения динамических моделей для анализа данных в сочетании с лонгитюдными измерениями [80]. Неоднократное тестирование учащихся на протяжении определенного периода обу-

Электронный архив библиотеки МГУ имени А.А. Кулешова

чения, позволяющее оценивать скорость прироста учебных достижений, а затем сопоставлять школы на основе средних скоростей прироста, дает более достоверные оценки качества образования по сравнению с одноразовыми измерениями.

Интерпретация данных тестирования, отражающая динамику изменения качества подготовленности учащихся, менее восприимчива к просчетам при формировании выборочных данных для анализа и косвенно принимает во внимание начальные (входные) данные учеников.

Влияние репрезентативности выборки на обоснованность интерпретации результатов тестирования. Репрезентативность выборки является важнейшим фактором, влияющим на обоснованность интерпретации результатов тестирования и качество измерений. На федеральном уровне для формирования репрезентативной выборки выпускников школ регионов необходимо определить генеральную совокупность, включающую всех выпускников текущего года в средних учебных заведениях общего образования Российской Федерации. Затем следует выделить основания по стратификации выборки (например, регион, район, центр региона, город разного типа, поселок, село, школа) и построить планируемую репрезентативную выборку из генеральной совокупности учащихся России, собирая сводную статистику по отдельным стратам.

Требования к использованию результатов тестирования в управлении качеством образования. Проблему использования результатов тестирования в управлении качеством образования лучше всего рассмотреть на примере ЕГЭ. В последнее время в ряде регионов предпринимаются отдельные попытки интерпретации данных ЕГЭ для анализа различных проблем, связанных с качеством образования. Корректная интерпретация результатов ЕГЭ в первую очередь зависит от правильной постановки основной цели их использования в управлении качеством образования. Процесс управления предполагает целенаправленную деятельность по повышению эффективности системы образования и всех ее компонентов. Поэтому приоритетными целями использования результатов тестирования на разных уровнях управления являются анализ, распространение и внедрение в практику функционирования образовательной системы России лучших технологий, опыта и методов работы всех субъектов системы, выявленных на основе анализа и интерпретации результатов ЕГЭ. Таким образом, данные по всем уровням анализа результатов тестирования должны оказывать позитивное воздействие на систему образования путем управленческой, методической и финансовой помощи отстающим образовательным учреждениям, а также внедрения в практику их работы лучшего опыта отечественного образования.

При использовании результатов тестирования в управлении качеством образования необходимо обеспечить сопоставимость ре-

зультатов выпускников школ по годам, которая требует выполнения ряда специальных требований со стороны теории педагогических измерений к содержанию тестов, их статистическим характеристикам и используемым методам шкалирования результатов учащихся. Помимо этого сопоставимость регламентирует определенные правила анализа данных тестирования для пользователей в управлении качеством образования. В частности для обеспечения сопоставимости данных по годам управленцам необходимо учитывать многие дополнительные факторы, отслеживать их изменение на протяжении анализируемого временно́го промежутка и их влияние на качество образования.

Для каждого уровня управления и региона необходимо выявить основные показатели, быстро изменяющиеся с течением времени. Таким образом, среди показателей — характеристик качества учебного процесса, кадрового состава, форм и методов дополнительного образования, уровня образованности и обеспеченности родителей учащихся, их социального статуса, миграционных процессов среди населения России и др. — необходимо выделить главные факторы, наиболее существенно влияющие на качество образования в данном регионе. При выполнении прочих условий, предъявляемых теорией педагогических измерений, сопоставление результатов тестирования по годам на каждом уровне управления возможно лишь в ситуации равенства главных факторов.

Надежность и генерализуемость данных тестирования. Для принятия эффективных управленческих решений необходима высокая объективность и обоснованность информации, базирующейся на результатах измерения. Среди управленцев и педагогов, использующих результаты тестирования, может сформироваться мнение о том, что надежность данных педагогических измерений является проблемой разработчиков тестов. Эта точка зрения верна лишь отчасти, поскольку надежность является составной частью более общей теории генерализации данных педагогических измерений.

Теория генерализации обосновывает качество широкого диапазона компонентов измерений: самих тестов, шкал учебных достижений, моделей измерения, процедур проведения тестирования, методов обработки результатов, поведения экзаменаторов и экспертов и многих других компонентов процессов тестирования, влияющих на величину ошибки при принятии определенных управленческих решений [80].

Помимо этого теория генерализации обеспечивает оценку возможности распространения результатов тестирования, полученных на выборочной совокупности выпускников, на генеральную совокупность выпускников России или региона. В этой связи перед принятием административных решений на региональном и федеральном уровнях образования по данным тестирования необ-

Электронный архив библиотеки Челябинского государственного университета

ходимо оценить их генерализуемость, которая характеризует возможность обобщения различных тенденций в образовании, выявленных при интерпретации результатов выполнения тестов.

Последовательность работ по использованию результатов тестирования. Использование результатов тестирования для управления качеством образования должно включать определенные этапы, к которым относятся:

– выбор целей анализа и интерпретации результатов тестирования;

– определение объектов и уровней анализа;

– выбор показателей качества и переменных измерения;

– построение репрезентативных групп испытуемых вероятностно-пропорциональным методом для формирования матриц эмпирических данных тестирования;

– формирование матриц эмпирических данных по испытуемым выборки;

– планирование и проведение дополнительного эксперимента для сбора эмпирических данных (если есть потребность согласно задачам исследования);

– выбор фасетов, формирование дизайна исследования для оценки генерализуемости эмпирических данных (на федеральном или региональном уровнях анализа);

– оценка коэффициентов генерализации для данных тестирования (на федеральном или региональном уровнях анализа) и эмпирических данных дополнительного эксперимента (если они есть);

– оценка надежности и валидности эмпирических данных;

– применение количественных (качественных) методов или их сочетания для анализа и обработки эмпирических данных;

– интерпретация результатов обработки;

– подготовка выводов и рекомендаций для принятия управленческих решений в целях повышения качества образования.

Среди перечисленных этапов трудно выделить менее или более важный. Все они обязательны в ситуации принятия административных управленческих выводов на муниципальном, региональном или федеральном уровнях управления качеством образования. При использовании результатов тестирования в школе, и тем более в классе, часть этапов можно опустить.

В частности в школе по результатам тестирования могут быть построены радиальные диаграммы для сравнительного анализа тематической структуры усвоения содержания учебной дисциплины отдельными учащимися (на уровне класса или всей школы). Для различных групп учащихся может быть проведено сравнение результатов со средними баллами по школе или по району, а также установлена степень затруднений или успешности обучения отдельных учащихся или целых классов, выявлены слабо усвоенные разделы предметов и причины таких отставаний и т.д.

Дальневосточный федеральный университет имени Г.И. Амосова

Для муниципального уровня управления качеством образования с учетом временно́го фактора можно проследить динамику среднего балла по районам и по различным предметам, провести рейтинг школ внутри районов города и другие сравнительные исследования.

В целом следует отметить, что разработка методики использования результатов тестирования для управления качеством образования — актуальная задача, поскольку в настоящий момент в условиях множества учебных программ, допущенных к использованию, и многообразия школ система управления качеством образования нуждается в проведении различных сравнительных исследований. Различаются квалификации преподавателей, условия учебного процесса, в том числе и бытовые, комплектность школ, национальный и гендерный состав и др. Согласно международным исследованиям, все эти факторы серьезным образом сказываются на качестве подготовки выпускников учебных заведений. Поэтому принятие обоснованных управленческих решений сегодня невозможно без использования данных педагогических измерений.

Северо-Восточный федеральный университет им. М.К. Аммосова

Практические задания и вопросы для обсуждения

1. Какие условия и процедуры тестирования подлежат стандартизации?

2. Разработайте инструкции для тестирования в классе при текущем контроле для:

а) случая, когда в тесте есть только задания с выбором одного или нескольких правильных ответов;

б) случая, когда в тесте совмещены задания разных форм.

3. Соблюдение принципа конфиденциальности предполагает неразглашение сведений о результатах тестирования без согласия ученика. Как вы думаете, правильно ли сообщать результаты ЕГЭ учителям и родителям ученика без его согласия?

4. Сформулируйте административно-управленческие задачи в образовании, при решении которых необходимо проверять данные тестирования на генерализуемость.

ГЛАВА

13

ШКАЛИРОВАНИЕ РЕЗУЛЬТАТОВ ТЕСТИРОВАНИЯ

13.1. Постановка задачи шкалирования

Для чего и когда следует использовать процедуру шкалирования. Для обоснованного сопоставления результатов учащихся между собой тестовые баллы в соответствии с рядом критериев и норм (число правильно выполненных заданий при дихотомической оценке результатов выполнения каждого задания, сумма оценок по отдельным заданиям при политомической, или взвешенной, оценке) переводятся в производные показатели при помощи процедуры, которая получила название шкалирования.

Таким образом, процесс шкалирования состоит в преобразовании сырых баллов в производные показатели, обеспечивающие адекватную интерпретацию и сравнение результатов выполнения педагогических тестов [1; 21; 22; 60].

Современная трактовка процесса шкалирования. Процесс шкалирования включает в себя различные процедуры. В простейшем случае под шкалированием понимается отображение сырых баллов на готовую шкалу, производимое по определенным правилам.

Перевод сырых баллов в производные показатели и их размещение на готовой шкале не могут повысить надежность и валидность данных по тесту.

В современной литературе по теории педагогических измерений встречается расширенное понимание процедуры шкалирования, в которую включают конструирование шкалы по определенным правилам и последующее преобразование исходных эмпирических данных для помещения их на данную шкалу. Таким образом, согласно расширенной трактовке, шкалирование включает ряд последовательных этапов, охватывающих все компоненты педагогических измерений, и имеет связь с качеством результатов.

13.2. Этапы построения шкал для педагогических измерений

Этапы шкалирования. При трактовке процесса шкалирования в расширенном варианте можно выделить четыре основных этапа

Северо-Восточный федеральный университет им. М.К.Аммосова

построения измерительных шкал в образовании для ситуации бланкового тестирования и обобщенного случая измерений:

Этап 1 — определение цели измерения, выбор конструкта, размерности и содержательной области, адекватно описывающей конструкт.

Этап 2 — разработка заданий и экспертное обоснование их качества, экспертное оценивание адекватности содержания заданий конструкту, определение первоначальной длины теста.

Этап 3 — апробация, эмпирический анализ качества теста, чистка и коррекция измерителя для повышения надежности и валидности шкалы, проверка размерности пространства измерений или доказательство одномерности теста.

Этап 4 — подтверждение качества шкалы и анализ возможности ее использования для представления результатов учащихся по тесту.

Последний этап начинается с построения устойчивой шкалы, выбранной в соответствии с целями измерения и подходом к созданию теста. При последующем использовании теста сырые баллы учеников отображаются на готовой шкале. Особую важность на данном этапе имеет процедура выравнивания результатов педагогических измерений, полученных учащимися по разным вариантам теста.

Необходимость выравнивания может быть не совсем понятна педагогу-практику, поскольку в школе принято выдавать существенно различающиеся по трудности варианты контрольных работ, а затем присваивать одинаковые оценочные эквиваленты разным, зачастую несопоставимым, результатам учащихся. В практике педагогических измерений утвердилась другая норма сравнения и интерпретации результатов испытуемых, основанная на выравнивании, которое представляет собой статистический метод преобразования оценок испытуемых по различным вариантам для обеспечения их сопоставимости.

Федеральный восточный университет имени Н.Н. Миклухо-Маклая

13.3. Виды шкал в образовании

Общие цели шкалирования. Процесс шкалирования реализует разные цели в зависимости от подхода, выбранного к разработке теста. При нормативно-ориентированном подходе шкалированные показатели позволяют уточнить место, занимаемое результатом испытуемого относительно норм, или сравнить результаты испытуемых, установив место результата каждого учащегося по отношению к результатам остальных учащихся, выполнявших этот тест.

При критериально-ориентированном подходе шкалированный балл показывает процент освоенного содержания и место резуль-

тата учащегося в сравнении с критериальным баллом. Перечисленным целям отвечают разные шкалы, которые можно построить по результатам выполнения теста.

Шкала перцентильных рангов. Перцентильный (процентильный) ранг для каждого балла определяется процентом испытуемых, которые выполнили столько же или меньше заданий теста. Например, если 30 % учащихся выполнили верно по 20 заданий теста и получили за каждое из них по одному баллу, то сырой балл «20» соответствует 30-му перцентилю. Таким образом, перцентиль показывает относительное положение испытуемого в выборке учащихся, которая выполняла тест. Чем ниже перцентильный ранг результата испытуемого, тем хуже его результаты по сравнению с другими тестируемыми группы.

Перцентили выше 50-го представляют результаты выше среднего по выборке, а перцентили ниже 50-го — ниже среднего, если в качестве средней нормы выступает медиана, которой соответствует 50-й перцентиль. Для 25-го и 75-го перцентилей существуют специальные названия: 1-й и 3-й квартили соответственно. Они отсекают нижнюю и верхнюю четверть распределения тестовых баллов, поэтому их выделение удобно для сравнения результатов данного тестирования с распределениями результатов по другим тестам.

Если шкала перцентилей построена на выборке стандартизации, то, используя ее, легко определить ранг каждого учащегося, выполнявшего в другое время тот же тест. Для этого достаточно подсчитать его сырой балл и по готовой таблице соответствия найти соответствующий перцентиль. Первичный балл, который ниже любого результата в выборке стандартизации, будет иметь нулевой перцентильный ранг. Результат, превышающий любой другой в выборке, получит перцентильный ранг 100. Конечно, оба эти результата не говорят о нулевом или абсолютном результате выполнения теста. Перцентили не следует путать с обычными процентными показателями, которые при дихотомическом оценивании результатов выполнения отдельных заданий представляют собой выраженную в процентах долю правильно выполненных заданий теста. В отличие от обычных процентов перцентиль является производным показателем, который оценивается в единицах процента испытуемых.

Перцентили имеют несомненные достоинства — они удобны в подсчете и просты в интерпретации. Помимо достоинств перцентильные ранги имеют два существенных недостатка. В о-п е р в ы х, они являются значениями порядковой шкалы, так как показывают относительное положение каждого индивида в нормативной выборке, а не определяют величину истинного различия между результатами отдельных испытуемых группы. В о-в т о р ы х, перцентили не только не отражают, но даже искажают реальные раз-

Электронный архив библиотеки МГПУ имени А.А. Кулешова

личия в результатах выполнения теста. Это связано с особенностями распределения перцентилей, имеющего прямоугольный характер. В этой связи небольшие отклонения от среднего в центре распределения наблюдаемых баллов будут значительно увеличены перцентилями, в то время как относительно большие отклонения на краях кривой нормального распределения будут сжаты.

Стандартные показатели. *Z*-шкала. При выборе метода шкалирования часто обращаются к стандартным показателям, указывающим отличие индивидуального результата испытуемого от среднего балла по выборке в единицах стандартного отклонения. Эти показатели используются для установления места первичного балла каждого испытуемого в сравнении с результатами других на основе подсчета нормированных отклонений и называются *z*-оценками. Результат отображения *z*-оценок на числовую ось образует *Z*-шкалу.

Для перевода в *Z*-шкалу сырой балл *i*-го испытуемого преобразуется по формуле

$$Z_i = \frac{X_i - \bar{X}}{S_x}, \qquad (18)$$

где X_i — сырой балл *i*-го испытуемого; $\bar{X}$ — среднее значение индивидуальных баллов *N* испытуемых группы; S_x — стандартное отклонение. Поскольку среднее значение $\bar{X}$ вычитается из каждого исходного значения X_i, то новое среднее в *Z*-шкале — *z* — будет равно нулю, а стандартное отклонение благодаря нормированию будет равно единице.

Если величина разности $X_i - \bar{X}$, стоящей в числителе дроби, больше 0, то результат *i*-го испытуемого выше среднего по тесту. В противном случае индивидуальный балл *i*-го испытуемого ниже среднего. В силу линейного характера преобразований при получении *z*-оценок все свойства исходного распределения сырых баллов переносятся на множество шкалированных баллов.

Использовать *Z*-шкалу можно для любого распределения индивидуальных баллов. Особенно удобны *z*-оценки в случае близости распределения первичных баллов к требованиям нормального закона, поскольку можно заранее предсказать процент результатов, лежащих в пределах одного и двух стандартных отклонений под кривой нормального распределения. Несомненным достоинством *Z*-шкалы является общая средняя арифметическая и общая мера вариации данных, позволяющие достичь сравнимости результатов по разным тестам.

Однако помимо явных достоинств есть и недостатки. Отрицательные и дробные *z*-оценки, которые нередко получаются при вычитании среднего и деления на стандартное отклонение, малопригодны для сообщения результатов тестирования испытуемых

Государственный университет им. А.А. Кулешова

группы. Поэтому применяются специальные методы линейного преобразования z-оценок для перевода их на множество целых положительных чисел.

Шкалы стандартных оценок, полученных на основе линейных преобразований *Z*-шкалы. Для перевода z-оценок в область положительных целых чисел выбираются новые значения среднего арифметического (M) и стандартного отклонения (σ). Они сохраняют все различия между баллами испытуемых, выявленные в Z-шкале, но позволяют избавиться от отрицательных и дробных значений z благодаря умножению каждой z-оценки на одно и то же число, а также прибавлению общей константы и последующему округлению. Для преобразования z-оценок используется формула

$$z_1 = M + \sigma_z, \tag{19}$$

где M — новое среднее арифметическое; σ — новое стандартное отклонение.

В качестве значений M и σ в формуле (19) можно использовать любые удобные числа. Например, для шкалы IQ эти значения равны 100 и 15. Поэтому $z_{IQ} = 100 + 15z$. Другое линейное преобразование с $M = 50$ и $\sigma = 10$ переводит значения z в стобалльную T-шкалу по формуле $T = 50 + 10z$. Эта шкала позволяет избавиться от дробных и отрицательных значений только в том случае, если значения z лежат в интервале от –5 до +5 и имеют один знак после запятой. В противном случае, если z-показатели подсчитаны с точностью до сотых, необходимо последующее округление T-показателей, что может привести к снижению дифференцирующего эффекта теста.

Для шкалы CEEB по тестам SAT (Scolastic Aptitude Test), разработанным Советом по приемным экзаменам в колледжи, z-оценки пересчитываются со средним $M = 500$ и $\sigma = 100$ по формуле $z_{CEEB} = 500 + 100z$. Значению $z = -1$ будет соответствовать значение $z_{CEEB} = 500 + 100(-1) = 400$. А при $z = +1$ — $z_{CEEB} = 600$. Таким образом, в шкале CEEB все дробные z-оценки превращаются в целые и попадают в интервал (0; 1000) в тех случаях, когда Z лежит в интервале (–5; +5). Так же в тысячебалльную шкалу переводятся оценки результатов выполнения таких известных в мире тестов, как GRE (Graduate Record Examination) и др.

Сопоставимость и выравнивание. Поскольку обеспечение сопоставимости результатов педагогических измерений является одной из главных причин перехода от сырых баллов к производным показателям в процессе шкалирования, то возникает вопрос о возможности сравнения z-оценок, полученных на основе различных вариантов теста. Ответ на этот вопрос на теоретическом уровне носит, несомненно, положительный характер в тех случаях, когда сравниваются z-оценки по параллельным вариантам одного

Гомельский государственный университет имени Ф. Скорины

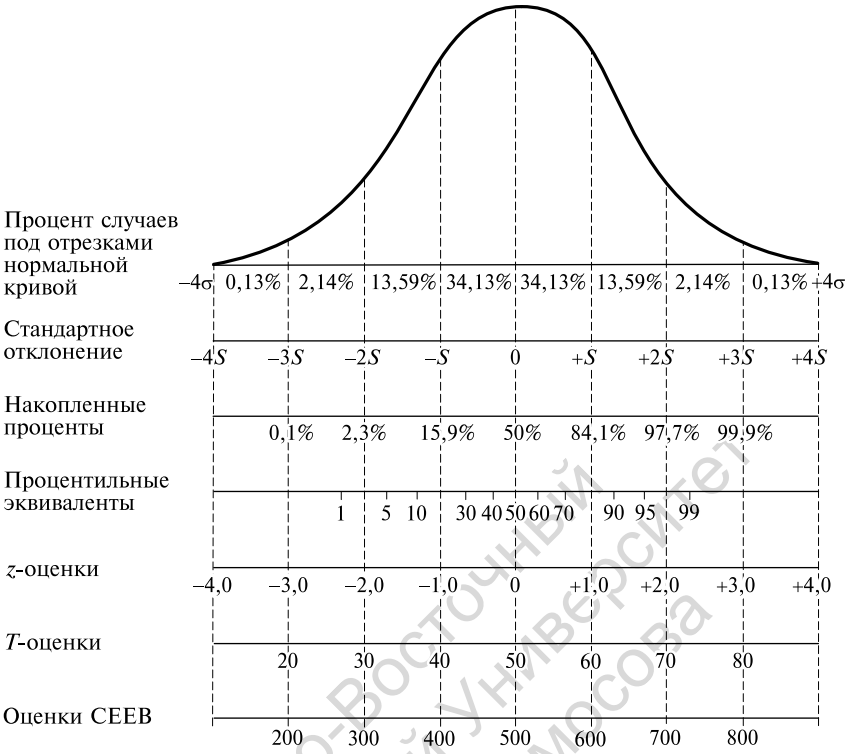

Рис. 33. Сопоставление шкал

и того же теста. Однако на практике из-за неизбежных отклонений от требований параллельности и существования ошибок измерения для повышения сопоставимости оценок испытуемых обычно используют процедуру выравнивания.

В отдельных случаях возникает необходимость сравнения относительного положения испытуемых, полученного в различных шкалах и по различным тестам. Если результаты тестирования имеют нормальное распределение, а выстроенные шкалы основаны на идентичных выборках испытуемых, такое сравнение можно провести с помощью рис. 33.

Чтобы добиться сопоставимости результатов тестирования в ситуации отличия распределений баллов от нормального закона, необходимо преобразование, изменяющее вид кривой распределения с целью приближения ее к виду нормальной кривой.

Нормализация данных тестирования. Для нормализации данных тестирования используется нелинейное преобразование, позволяющее придать эмпирическому распределению желаемую форму нормальной кривой. С этой целью вводятся нормализованные стан-

Северо-Восточный федеральный университет им. М.К. Аммосова

дартные показатели, соответствующие распределению, преобразованному так, что оно аппроксимируется формой нормальной кривой. Их значения могут быть найдены с помощью таблиц, в которых приводится процент случаев различных отклонений в единицах от среднего значения для нормальной кривой.

Преобразование сырых баллов к нормальному распределению осуществляется способом, получившим название *пробит-преобразования* [1; 18]. В рамках процедуры преобразования баллов сначала для каждого сырого показателя определяется кумулированная частота, которая представляет собой сумму всех частот, лежащих ниже данного сырого показателя. Затем к ней добавляется половина количества испытуемых, имеющих этот сырой балл. По этим данным вычисляется кумулированная доля путем деления полученной суммы на общее число испытуемых выборки. Затем по статистическим таблицам, содержащим значения площади под кривой нормального распределения, находят значения нормализованных стандартных показателей для каждой кумулированной доли [63].

Нормализованный стандартный показатель, как и линейно преобразованный стандартный показатель, имеет среднее значение «0», а стандартное отклонение — «1». Результат учащегося в «–1» балл можно интерпретировать как превосходящий приблизительно 16 % результатов группы, а в «+1» балл — как превосходящий 84 % всех результатов.

Шкала станайнов, стенов и другие шкалы. Нормализованным стандартным показателям, так же как и линейно преобразованным, стараются придать удобную форму, пригодную для сообщения испытуемым. Для этого используют шкалы стандартных десяти или девяти единиц. Разбиение нормального распределения на девять интервалов приводит к *шкале станайнов*, имеющей девять стандартных единиц. Название «станайн» связано с тем, что оценки в этой шкале принимают значения от «1» до «9». При оценке результатов испытуемых по тесту 4 % самых худших результатов присваивается станайн 1, а самых лучших — станайн 9. Следующим за худшими и лучшими 7 % результатов присваивают станайны 2 и 8 соответственно. Далее 12 % результатов — станайны 3 и 7. Следующим 17 % присваивают станайны 4 и 6 и, наконец, 20% средних результатов — станайн 5 (табл. 16).

Помимо описанной шкалы станайнов существуют еще две шкалы, имеющие некоторое преимущество перед девятибалльной в смысле различающей способности. Одна из них — шкала стандартных десяти единиц, называемая также шкалой Кэттелла, или *шкалой стенов* (*sten*). Как следует из названия, весь массив результатов делится на десять частей с интервалом 0,5 стандартного отклонения. В шкале стенов среднее арифметическое принимается равным 5,5, а расстояние между двумя соседними стандартными единицами равно $0{,}5S_x$.

Таблица 16

Таблица соответствия процентов и станайнов

Процент	4	7	12	17	20	17	12	7	4
Станайн	1	2	3	4	5	6	7	8	9

Какие шкалы использовать в педагогических измерениях. Многие из шкал, приведенных выше, используются исключительно психологами, другие нашли свое применение в образовании. В практике деятельности зарубежных тестовых служб в образовании чаще всего обращаются к *стобалльной* или *тысячебалльной шкале*, полученным на основе преобразования *z*-оценок. Хотя тысячебалльная шкала обладает высокими дифференцирующими возможностями, обычно ее концы оказываются не работающими в силу специального подбора по трудности заданий теста для приближения частотных распределений оценок трудности к виду нормальной кривой. Поэтому, как правило, оценки испытуемых распределяются в интервале от 200 до 800 баллов. Но даже использование менее протяженного диапазона оценок, чем тысячебалльная шкала, требует специальных профессиональных навыков по интерпретации баллов учащихся.

Как осмыслить свой результат, если он, например, равен 570 или 650 баллам? Как отнести его к категории плохих или хороших результатов на столь широком диапазоне баллов? Другое дело, если результат испытуемого составляет 5 или 6 баллов по девятибалльной шкале. Поэтому к растянутым шкалам обычно обращаются профессиональные тестовые службы для массового тестирования в образовании, когда большое число испытуемых требует повышения дифференцирующей способности шкалы.

В России при шкалировании данных ЕГЭ была выбрана стобалльная шкала, в которую переводятся оценки выпускников. Конечно, стобалльная шкала — это своего рода компромисс между потребностью в хорошем дифференцирующем эффекте шкалы из-за значительного числа тестируемых во всех регионах и постепенным переходом от пятибалльной шкалы, существовавшей на протяжении многих лет в России, к более растянутым шкалам.

13.4. Шкалирование результатов тестирования на основе теории IRT

Построение шкалы с помощью современной теории тестов. Рассмотренные в предыдущем разделе шкалы позволяют сопоставить результаты тестирования и служат удобной формой их интерпретации, но они не повышают уровень измерений в силу того, что

Северо-Восточный университет им. М.К. Аммосова

используют статистический аппарат классической теории тестов. Порядковая шкала сырых баллов испытуемых переходит в порядковую шкалу производных стандартизированных показателей, не позволяющих интерпретировать разность результатов двух испытуемых, выполнявших один и тот же тест. Зарубежные исследования конца 80-х гг. XX в. показали возможность построения интервальной шкалы результатов педагогических измерений в том случае, если для создания теста и шкалирования результатов его выполнения используется теория IRT.

Условно процесс шкалирования в IRT можно подразделить на три этапа. Первый предполагает построение шкалы логитов для латентного параметра подготовленности испытуемых, второй — шкалы логитов для оценок латентного параметра трудности заданий. Третий этап позволяет свести две шкалы в общую шкалу стандартных оценок для обоих латентных параметров.

Связь шкалы логитов и шкалы Гутмана. Процедура построения шкалы латентных переменных связана с так называемым шкалированием по Гуттману (*Guttman — type scale*), в которой задания отбираются в порядке нарастания их трудности по определенным, тщательно структурированным элементам содержания дисциплины. Отличительной особенностью шкалы Гуттмана является существование стойкого кумулятивного эффекта, означающего, что любой испытуемый с правильной структурой знаний, справившийся с *j*-м заданием, может наверняка успешно выполнить все предыдущие, более легкие задания теста. В понимании Гуттмана совершенная шкала существует в том случае, если по последнему правильному ответу испытуемого можно воспроизвести все его ответы на более легкие задания теста.

Конечно, стойкий кумулятивный эффект наблюдается далеко не всегда. В основном он характерен для заданий, довольно тесно связанных по содержанию. Для иллюстрации идей Гуттмана в англоязычной методической литературе популярен следующий пример заданий на умножение:

$$1)\ \times\frac{17}{2} \quad 2)\ \times\frac{236}{12} \quad 3)\ \times\frac{1437}{382} \quad 4)\ \times\frac{57342}{7412}$$

Он вполне ясно, хотя и довольно упрощенно, показывает, как реализуется эффект кумулятивности на практике. Действительно, если испытуемый умеет умножать на четырехзначное число, то он тем более справится с умножением на трех-, двух- и однозначные числа.

Шкалирование на основе теории IRT в определенной степени преодолевает ограниченность предположений шкалы Гуттмана, поскольку является вероятностной версией и отражает сущность тестовых процессов, неизбежно связанных с ошибками измере-

Федеральный институт развития образования имени К.Д. Ушинского

ния. Согласно моделям IRT о правильном выполнении любого задания испытуемым, можно прогнозировать успешность лишь в том случае, если эта вероятность близка к единице.

Преимущества и проблемы шкалирования по теории IRT. Инвариантность оценок параметров испытуемых относительно трудности заданий теста, достигаемая благодаря возможностям IRT, позволяет реализовать эффект специфической объективности, который способствует повышению точности оценок параметра подготовленности учащихся. Благодаря единой шкале интервального типа в IRT разности оценок латентных параметров испытуемых приобретают вполне интерпретируемый смысл, поскольку их можно считать мерой отличия в подготовленности испытуемых по предмету. Таким образом, теория IRT повышает возможности педагогической интерпретации шкалированных баллов учащихся. С ее помощью можно сопоставить приращения в обученности учащихся и повысить надежность их оценок по тесту.

Однако реализовать преимущества теории IRT довольно сложно. Для этого необходимо обеспечить выполнение ряда условий ее применимости, без которых эффект инвариантности не имеет места. В частности нужно обеспечить конструирование теста на основе теории IRT, подтвердить соответствие эмпирических данных тестирования требованиям моделей измерения или удалить неподходящие данные по результатам выполнения теста. Необходимо также обеспечить нормальный характер распределения сырых баллов учащихся, оценок трудности заданий теста, ошибок измерения и реализовать требование локальной независимости отдельных заданий теста. Немало проблем вызывает расходимость итерационных процессов, работающих в методе максимального правдоподобия при переходе от начальных оценок к наиболее эффективным оценкам параметров испытуемых и трудности заданий теста. Поэтому теория IRT в шкалировании используется далеко не всегда, только в случаях массового тестирования для принятия административно-управленческих решений в образовании, когда есть смысл тратить силы на разработку и применение теста.

Преобразования шкалы логитов. Поскольку оценки параметров подготовленности учащихся и трудности заданий теста в шкале логитов обычно лежат в интервале (–5; 5) и имеют несколько знаков после запятой, они малопригодны для сообщения испытуемым без приведения к целому неотрицательному виду. Поэтому необходимы линейные преобразования оценок в другую, более удобную для сообщения результатов шкалу подобно тому, как это происходит с z-оценками.

Сначала все значения параметров умножают на один и тот же множитель для перевода результатов в область целых чисел и округляют результат до целых. Затем переносят все значения параметров на множество положительных чисел путем прибавления

некоторой константы, определяющей новую точку отсчета на шкале, для того чтобы избавиться от отрицательных оценок параметра подготовленности θ. Примеры таких преобразований приведены в специальной литературе по шкалированию результатов педагогических измерений.

13.5. Шкалирование в критериально-ориентированном тестировании

Виды шкал в критериально-ориентированном тестировании. Виды шкал в критериально-ориентированном тестировании выбираются в зависимости от предназначения теста. Если тесты используются для оценки степени освоения содержательной области (domain-referenced tests), отображение которой в тесте условно можно принять за 100 %, то каждый балл учащегося показывает процент освоенного содержания. Процесс шкалирования осуществляется достаточно просто: балл, набранный учащимся, делят на максимально возможный балл по тесту и полученную величину умножают на 100 %. Упорядочение найденных результатов и их нанесение на ось позволяют построить шкалу, каждая точка которой соответствует проценту усвоенного содержания для учащегося или группы учеников.

В другом случае, когда критериально-ориентированный тест применяется для деления тестируемых на две или несколько групп с помощью порогового (критериального) балла (mastery test), строится номинальная шкала. Например, подобное деление происходит при аттестации: в одну группу попадают аттестованные, а в другую — не аттестованные учащиеся, как не выполнившие запланированный процент заданий теста. Основная трудность при таком шкалировании заключается в установлении порогового балла для отсечения группы учащихся, не показавшей достаточного владения содержанием теста.

Методы выбора критериального балла. Для установления порогового балла используются три метода. В первом случае балл устанавливается экспертным путем, априорно, на основе анализа целостного содержания теста. Во втором случае эксперты выбирают пороговый балл на основе анализа содержания тестовых заданий и присвоения им априорных оценок трудности, с помощью которых выделяется критерий отбора в группу аттестованных учащихся. В третьем случае для определения порогового балла анализируются эмпирические данные по результатам апробации теста на репрезентативной выборке учащихся и используется метод контрастных групп.

Для получения валидного значения критериального балла третьим методом прежде всего необходимо провести предваритель-

Электронный архив Восточно-Казахстанского государственного университета им. С. Аманжолова

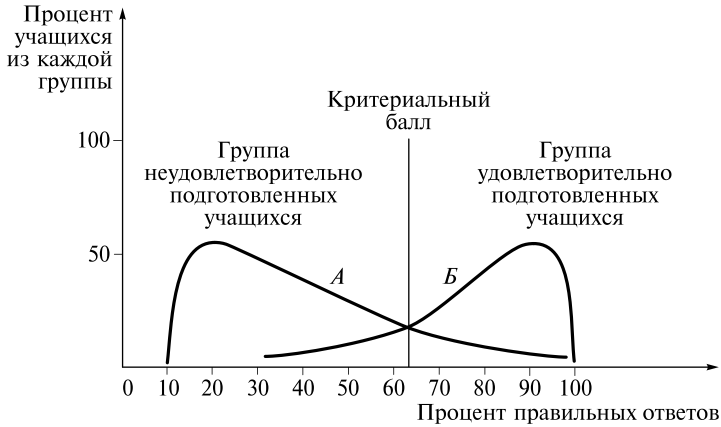

Рис. 34. Сглаженные частотные распределения баллов по тесту для контрастных подгрупп

ное тестирование на близком по содержанию входном претесте или отобрать группу экспертов, хорошо представляющих подготовленность тестируемой выборки учащихся. По результатам претеста или экспертизы из группы учащихся выделяются две контрастные подгруппы: заведомо не готовых к тесту самых слабых — 27 % и 27 % самых сильных, хорошо подготовленных к тестированию. В совокупности получаются две контрастные по подготовленности выборки учеников. Затем каждой подгруппе (слабой и сильной) выдается критериально-ориентированный тест, распределение баллов по которому строится на одном графике отдельно для слабых и сильных учащихся (сглаженные кривые — рис. 34, экспериментальные кривые — рис. 35).

Северо-Восточный федеральный университет им. М.К. Аммосова

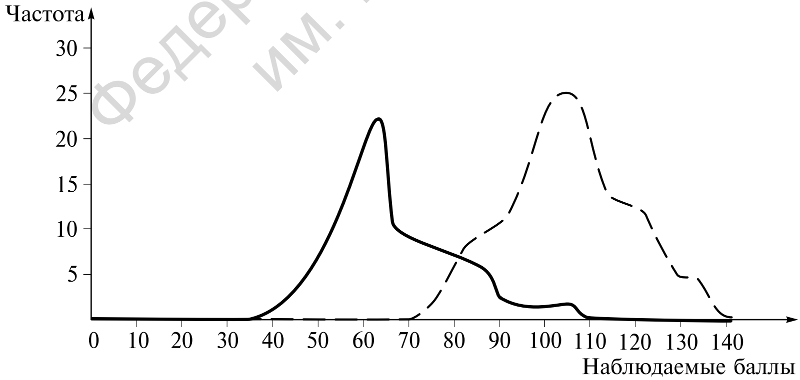

Рис. 35. Эмпирические частотные распределения баллов по тесту для контрастных подгрупп

После проведения тестирования на репрезентативной выборке учащихся и построения частотных распределений для контрастных групп устанавливается критериальный балл в точке, соответствующей на горизонтальной оси пересечению кривых распределения баллов. Эта точка пересечения, спроецированная на рис. 35 на горизонтальную ось, наиболее четко разделяет группы не аттестованных и аттестованных учащихся, поскольку в ней наблюдается наименьший процент ошибочных решений — одновременно минимизируется число учащихся, обладающих достаточно высокой подготовкой, но попавших в группу не аттестованных (часть кривой *A* слева от вертикальной прямой) и число неподготовленных учеников, ошибочно отнесенных к категории прошедших за пороговый балл (часть кривой *B* справа от вертикальной прямой). Полученный пороговый балл обладает наибольшей достоверностью по сравнению с его аналогами, определенными экспертными методами.

Уровневые шкалы, совмещающие нормативно-ориентированный и критериально-ориентированный подходы. Для получения надежных и обоснованных результатов итоговой аттестации выпускников учебных заведений тестовый балл иногда дополняют развернутой содержательной интерпретацией, описывающей характеристики уровня подготовки учащегося в терминах освоенных элементов содержания.

Рис. 36. Пример уровневой шкалы

Такие шкалы, позволяющие совместить интерпретацию оценки испытуемого по отношению к результатам остальных тестируемых и к уровням освоения содержания, выделенным по критериальному принципу, получили название уровневых. Пример уровневой шкалы приведен на рис. 36, на котором диапазоны тысячебалльной шкалы, выбранные гипотетически, соотносятся с уровнями подготовки.

На рисунке выделен базовый и промежуточный уровни вместе с уровнем высокой компетентности. Для построения уровневой шкалы обычно шкалируют результаты репрезентативной группы учащихся в рамках нормативно-ориентированного подхода и строят стандартизованную шкалу тестовых баллов. Затем на шкале выделяют диапазоны и выявляют совокупности содержательных элементов, освоенных учащимися в каждом диапазоне, дополняя детальным описанием освоенных знаний и умений.

Федеральный институт развития образования Северо-Восточный

13.6. Рейтинговые шкалы

Упрощенная трактовка рейтинговой шкалы. В российской системе высшего и среднего образования нет устоявшихся определений, позволяющих однозначно определить рейтинговый балл учащегося. В основном под ним понимают накопленный балл, полученный в результате простого или взвешенного суммирования оценок в порядковых шкалах, которые строятся на основе субъективного выставления и учета баллов учащегося в соответствии с различными уровнями учебной деятельности, временны́ми промежутками в обучении или уровнями усвоения. Нередко к сумативным оценкам, характеризующим успеваемость, прибавляют поощрительные баллы за своевременную сдачу заданий, активность на занятиях, хорошую посещаемость и т. д.

Такая упрощенная трактовка, далекая от педагогических измерений, таит в себе, по меньшей мере, две серьезные ошибки: в о-п е р в ы х, операция суммирования является недопустимой на порядковом уровне измерений, в о-в т о р ы х, происходит бессмысленное объединение баллов по различным переменным, что исключает возможность какой-либо корректной интерпретации результатов подобного объединения. Вполне возможна ситуация, когда в сумме баллов, накопленной учащимся за определенный период обучения, будут доминировать оценки по второстепенным переменным, не имеющим заметного отношения к целям образования.

Таким образом, за видимой простотой операции получения рейтингового балла скрывается серьезная опасность: по результатам обучения могут быть признаны лучшими те учащиеся, которые не обладают творческим мышлением, но вовремя сдают домашние задания, не пропускают уроков и не нарушают дисциплины в классе.

Обращение к рейтинговой шкале в связке с контрольными заданиями для модулей, построенным на деятельностной основе в русле идей модульного обучения, немного повышает корректность приведенной выше упрощенной трактовки. По крайней мере выделение модулей происходит на содержательной основе и позволяет накапливать оценки уровней усвоения конкретных предметных знаний, что способствует обоснованной интерпретации суммарной оценки.

В целом рейтинговые баллы при корректном подходе к их подсчету и интерпретации могут оказать позитивное влияние на контрольно-оценочную систему в образовании. Они способствуют систематической работе учащихся, снижают роль случайности при сдаче экзаменов и снимают нервное напряжение во время экзаменов благодаря заблаговременному накоплению оценок результатов обучения.

Федеральный университет им. М. Ахмедова

Корректный подход к построению рейтинговых шкал на основе теории педагогических измерений. Для корректного построения рейтинговых шкал необходимо выполнять ряд условий. В зарубежной литературе к ним относят:

– концептуальное выделение переменных измерения;

– использование тестов с высокой содержательной и конструктной валидностью для получения баллов учащихся по каждой переменной;

– построение отдельных рейтинговых шкал для каждой переменной измерения;

– интеграцию результатов по отдельным шкалам (количественного характера) в единую рейтинговую шкалу с использованием весовых коэффициентов, определенных с помощью регрессионного анализа и методов выравнивания шкал для тестов различной длины при последующем объединении взвешенных количественных баллов по отдельным шкалам.

В целом необходимо отметить, что построение рейтинговых шкал требует от учителя определенной методической подготовки, наличия тестов и систематической работы по корректному построению отдельных шкал. При этом повышается нагрузка педагога, поэтому обманчивая простота рейтингования на деле при правильном подходе оборачивается значительными трудозатратами. Под вопросом остается общий эффект, поскольку пока неясно, оправданны ли такие затраты энергии со стороны педагогов или нет.

Северо-Восточный федеральный университет им. М.К. Аммосова

Практические задания и вопросы для обсуждения

1. Приведите известные вам примеры использования шкал, отличных от пятибалльных, в российском образовании. Удобны ли они для обучающихся?

2. Предположите, что группа студентов выполняла ранжированные по нарастанию трудности задания теста. Если индивидуальные баллы четырех студентов таковы, что $X_1 = 5$, $X_2 = 10$, $X_3 = 40$, $X_4 = 45$, то имеет ли смысл интерпретировать равенство $X_2 - X_1 = X_4 - X_3$ при сопоставлении результатов студентов? Какую шкалу необходимо построить для интерпретации разности баллов?

3. Можно ли выбрать единую шкалу тестовых баллов и пользоваться ею всегда при интерпретации результатов любого теста?

4. Можно ли путем шкалирования повысить надежность результатов выполнения теста?

5. Переведите в Z-шкалу сырые баллы 10 учеников: $X_1 = 2$, $X_2 = 7$, $X_3 = 1$, $X_4 = 5$, $X_5 = 5$, $X_6 = 11$, $X_7 = 9$, $X_8 = 2$, $X_9 = 15$, $X_{10} = 3$, выполнивших 25 заданий теста.

ГЛАВА

14

ЕДИНЫЙ ГОСУДАРСТВЕННЫЙ ЭКЗАМЕН, ЕГО КОМПОНЕНТЫ, ТЕХНОЛОГИЯ ПРОВЕДЕНИЯ, ШКАЛИРОВАНИЕ И ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ

14.1. Цели и задачи эксперимента по введению Единого государственного экзамена, его участники

Основные цели эксперимента по введению Единого государственного экзамена. Эксперимент по введению 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. Впервые в истории отечественного образования предпринята попытка не директивным, а экспериментальным путем определить целесообразность фундаментальной перестройки деятельности учебных заведений и системы управления качеством образования.

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 [9; 10]. По результатам ЕГЭ выпускники школ получают две оценки, одна из которых выставляется в школьной пятибалльной шкале, а другая — в стобалльной шкале для предоставления приемным комиссиям вузов и ссузов.

Задачи, решаемые с помощью ЕГЭ. К числу основных задач, решаемых в ЕГЭ, относятся:

– объективизация оценки качества образования на основе создания механизма внешнего оценивания и использования педагогических измерений;

– повышение доступности профессионального образования, в первую очередь для молодежи из малообеспеченных семей и из отдаленных от вузовских центров мест проживания;

– объективизация требований к общеобразовательной подготовке абитуриентов учебных заведений в системе профессионального образования;

– снижение психологической нагрузки на выпускников общеобразовательных учреждений за счет уменьшения числа экзаменов;

Северо-Восточный федеральный университет им. М.К. Аммосова

– развитие преемственности общего и профессионального образования, обеспечивающей готовность выпускников школ продолжить образование;

– совершенствование системы государственного контроля и управления качеством образования на основе независимой оценки качества подготовки выпускников.

Решение последней задачи, в частности, позволяет говорить о существенном вкладе ЕГЭ в становление и развитие Общероссийской системы оценки качества образования (ОСОКО), создание которой происходит сейчас в России. Выполняя свои основные задачи, ЕГЭ является важным структурным элементом ОСОКО и одним из центральных звеньев в развитии и совершенствовании системы управления качеством образования.

Решение перечисленных задач обеспечивается выполнением работ по ряду направлений, к которым относятся: формирование нормативно-правовой базы эксперимента; создание КИМ; разработка технологии проведения ЕГЭ; обеспечение информационной безопасности процедур, материалов и результатов экзамена; формирование информационных систем сопровождения эксперимента и использования его данных в образовании; подготовка специалистов для кадрового обеспечения эксперимента; мониторинг ЕГЭ; анализ и интерпретация данных мониторинга в целях повышения качества образования.

Схемы участия регионов в ЕГЭ и развитие эксперимента. После принятия решения об участии в ЕГЭ регион на добровольной основе выбирает одну из возможных схем проведения экзамена, которая может отличаться от других количеством и составом предметов, добровольностью или обязательностью аттестации выпускников школ в форме ЕГЭ, долей компьютерного тестирования и другими специфическими особенностями, обусловленными финансовыми возможностями региона, его материальной базой, уровнем развития коммуникаций, числом вузов, их готовностью к участию в эксперименте и т.д.

В соответствии с основной линией на отсутствие жесткой директивности со стороны федеральных органов управления образованием в ЕГЭ сохраняется максимально возможная вариативность схем проведения, которые нередко меняют сами регионы в процессе развития эксперимента.

Несмотря на существующую нормативную базу, определяющую перечень обязательных предметов при итоговой аттестации выпускников школ, многие регионы в рамках работ по совершенствованию схемы проведения ЕГЭ отказываются от обязательных предметов и предоставляют выпускникам право самостоятельного выбора числа экзаменов и самого участия в ЕГЭ. Показательно, что по мере расширения сферы действия принципов вариативности и добровольности число участников эксперимента неуклонно

Электронный архив библиотеки МГПУ им. И.М. Ситникова

растет вместе с проявлением позитивного отношения общества к эксперименту.

Анализ данных мониторинга ЕГЭ убедительно свидетельствует о резком увеличении числа регионов, участвующих в эксперименте, и повышении доверия к ЕГЭ со стороны органов управления образованием, профессионального сообщества вузов и школ, учащихся и их родителей. В частности число регионов — участников эксперимента — изменилось с 6 в 2001 г. до 78 в 2005 г., а количество выпускников этих регионов, сдававших в 2005 г. экзамены в форме ЕГЭ по математике и русскому языку, превысило 50 % [41]. Значительно расширился список вузов и ссузов, принимающих абитуриентов по результатам ЕГЭ. В 2005 г. в эксперименте приняли участие 1543 вуза (включая филиалы) и 1765 ссузов, тогда как в 2001 г. в нем участвовали лишь 16 вузов. Резко увеличилось общее количество предметов, выносимых на ЕГЭ в регионах. Согласно данным мониторинга и региональной статистики, за годы эксперимента существенно возросло число выпускников из сельских районов, поступивших в вузы по результатам ЕГЭ. В целом тенденции эксперимента говорят о том, что связываемые с ЕГЭ надежды на расширение доступности высшего образования для жителей сельских и отдаленных районов России полностью оправдываются. Это находит свое подтверждение в данных мониторинга ЕГЭ.

Участие учителей в проведении ЕГЭ. Учителя нередко выступают организаторами проведения ЕГЭ в школах, однако они работают только с теми выпускниками, которые не являлись ранее их учениками. В этом случае педагоги принимают участие в подготовке к проведению экзамена, организуют саму процедуру тестирования, обеспечивают сбор и отправку материалов экзамена.

При подготовке к тестированию учителя получают все необходимые материалы, включающие комплекты КИМ с запасными материалами на случай опечатки, списки выпускников, формы протоколов проведения тестирования, в которых отмечаются все отклонения от требований стандартизации, и руководство по проведению тестирования, включающее описание всех требований к проведению ЕГЭ.

После завершения экзамена организаторы собирают материалы: использованные и неиспользованные тесты, заполненный протокол проведения тестирования и прилагают списки учащихся. Эти материалы передаются в специальные центры для проверки и дальнейшей обработки, получившие в системе ЕГЭ название региональные центры обработки информации (РЦОИ). Задания с выбором ответов подвергаются автоматизированной проверке на местах или в Москве, а для проверки заданий с развернутыми ответами приглашаются эксперты, работа которых организуется по месту проведения экзамена в условиях полной информационной безопасности и независимости экспертных суждений.

Электронный архив библиотеки МГУ имени А.А. Кулешова

14.2. Контрольные измерительные материалы

Структура КИМ. В структуре КИМ выделяют три части (*A*, *B*, *C*), имеющие различное число заданий в зависимости от предмета. Например, заданий по биологии и истории может быть 50, а по русскому языку — 40.

В части *A* по всем предметам содержатся только задания с выбором ответов. Задания части *B* значительно варьируют по форме и предполагают возможность краткого регламентированного ответа, установления соответствия между элементами двух множеств и правильной последовательности различных процессов, явлений, объектов. Часть *C* предназначается для свободного конструирования ответа. Например, при тестировании по русскому языку дается задание написать небольшую письменную работу (сочинение), по математике — дать развернутые решения заданий и т.д.

Содержание КИМ. Результаты ЕГЭ используются при итоговой аттестации учащихся и отборе абитуриентов. Содержание КИМ предназначено для получения персонифицированной информации о достижении выпускниками школ базового и повышенного уровней подготовки по предметам. Поскольку при аттестации требуется проверить освоение выпускниками большинства элементов предметной подготовки, а время экзамена ограничено, для повышения репрезентативности охвата содержания образования приходится варьировать спецификации КИМ ЕГЭ незначительно внутри одного года и более существенно по годам. Благодаря специальному планированию, осуществляемому предметными комиссиями по разработке КИМ, за 2—3 года удается охватить все необходимые содержательные элементы. Из-за вариации спецификаций КИМ снижается сопоставимость результатов выпускников, но расширяются возможности использования результатов ЕГЭ на федеральном и региональном уровнях в мониторинге качества образования. Таким образом, на основе анализа результатов по отдельным годам ЕГЭ удается получить обоснованную информацию о тенденциях в изменении общеобразовательной подготовки выпускников школ России.

В целом содержание КИМ отбирается на основе спецификаций, которые ежегодно обновляются в Интернете и включают обобщенные планы экзаменационных работ со ссылками на соответствующие позиции кодификаторов (пронумерованных перечней содержательных элементов по предметам). В содержании КИМ отображаются только предметные знания и умения, которым обучают в школе, хотя это противоречит современным воззрениям на приоритеты в обучении, принятым во многих странах. Вследствие этого российские учащиеся показывают невысокие результаты в международных сравнительных исследованиях качества образования. Они не умеют выполнять задания, требующие приме-

нения знаний в ситуациях, имитирующих жизненные, или междисциплинарных умений.

Анализ результатов выполнения КИМ в контексте содержательной интерпретации дает важную информацию для совершенствования требований ГОС и повышения качества образования. Согласно результатам анализа, проведенного предметными комиссиями по разработке КИМ и опубликованного в 2006 г. в аналитическом докладе [41], задания базового уровня трудности выполняют в основном выпускники, получившие по ЕГЭ хорошие и отличные оценки. Процент учащихся этой группы, справившихся со всеми заданиями базового уровня, несколько выше по математике и русскому языку (более 90 %) и ниже по остальным предметам (не более 85 %). Около половины выпускников, получивших по результатам ЕГЭ «два» и «три», не осваивают и половины планируемого к изучению материала. Этот результат помогает выявить проблемы, связанные с завышенными требованиями ГОС и излишним объемом содержания школьного образования.

14.3. Технология разработки контрольно-измерительных материалов, организации и проведения Единого государственного экзамена

Структуры, участвующие в организации и проведении ЕГЭ. К числу структур, осуществляющих организационно-управленческую и научно-методическую поддержку при реализации технологии ЕГЭ, относятся федеральные и региональные органы управления образованием, Федеральный центр тестирования (ФЦТ), Федеральный институт педагогических измерений (ФИПИ), компания «КРОК», региональные центры обработки информации и различные организации (вузы и другие образовательные структуры), выполняющие на конкурсной основе исследовательские проекты для формирования научно-методических основ технологии проведения ЕГЭ и совершенствования КИМ.

Технология разработки КИМ. Технологический ежегодный цикл разработки КИМ по предмету начинается с пересмотра кодификатора элементов содержания, подлежащих проверке, а также с создания спецификации и демонстрационного варианта КИМ. Обычно по 13 предметам, заявленным в ЕГЭ, ежегодно разрабатывается не менее 10 тысяч новых заданий. Помимо целевого заказа на задания для пополнения банка ЕГЭ объявляется конкурс, в котором принимают участие все желающие авторы тестовых заданий из различных регионов России. Материалы, представленные на конкурс, проходят экспертизу, по результатам которой осуществляется отбор и покупка тестовых заданий для банка ЕГЭ.

Образовательный Интернет-Университет www.intuit.ru

Специалисты ФИПИ, используя банк, формируют варианты КИМ. Затем проводится экспертиза качества содержания вариантов, анализ их параллельности и соответствие заданий требованиям тестовой формы. После коррекции, основанной на результатах экспертизы, КИМ передаются на апробацию, проводимую на репрезентативных выборках учащихся школ России. Обработка данных апробации и их анализ позволяют выполнить очередную коррекцию, после завершения которой получаются эквивалентные варианты КИМ с комплектами сопроводительной документации для проведения ЕГЭ, созданные в условиях высокого уровня информационной безопасности и хранящиеся в специальных помещениях до начала ЕГЭ.

После проведения экзамена и обработки данных специалисты ФИПИ готовят аналитический отчет, содержащий рекомендации по совершенствованию КИМ, которые учитываются при выполнении очередного ежегодного технологического цикла.

Технология проведения ЕГЭ. При проведении ЕГЭ основной технологией является бланочная, предполагающая выдачу заданий выпускникам на бумаге. Реализация бланочной технологии зависит от организации, занимающейся ее осуществлением. Так, ФЦТ все материалы печатает в Москве, а затем перед началом экзамена передает пакеты с КИМ и бланками ответов в регионы. КРОК поддерживает технологию, при которой бланки ответов распечатываются по месту применения в регионах.

При использовании компьютерного тестирования в ЕГЭ технология носит смешанный характер, поскольку задания части *С* с развернутыми ответами выполняются на бумажных бланках. Единая технология проведения ЕГЭ пока еще не сложилась. Каждый из вариантов проведения эксперимента имеет свои достоинства и недостатки, выявить которые в полной мере можно лишь по его окончанию.

Федеральный институт им. М.А. Шолохова

14.4. Шкалирование результатов Единого государственного экзамена и использование их в управлении качеством образования

Виды используемых шкал в ЕГЭ. В соответствии с заявленными целями и решаемыми задачами по данным ЕГЭ выставляются две оценки — в сто- и пятибалльной школьной шкале. Первую получают специалисты ФЦТ путем шкалирования первичных данных ЕГЭ, преобразования их в стандартную шкалу логитов на основе современной теории тестов. Вторая, школьная, выбирается специалистами группы по шкалированию, организуемой Рособрнадзором во время экзаменов и состоящей из тестологов, представителей предметных комиссий ФИПИ и сотрудников ФЦТ.

Для выбора границ интервалов стобалльной шкалы и установления их соответствия оценкам четырехбалльной шкалы результаты ЕГЭ по каждому предмету и всем регионам подвергаются многоаспектному анализу, включающему проблемы содержательной валидности школьных баллов и сопоставимости оценок по предмету в ЕГЭ разных лет.

Сопоставимость результатов ЕГЭ разных лет. Для обеспечения содержательной сопоставимости результатов тестирования многие страны, в которых есть национальные экзамены в форме тестов или другие виды массового тестирования, вводят стандартизованные уровневые шкалы (см. раздел 13.5), позволяющие проследить связь между содержанием тестов по годам. Важное преимущество уровневых шкал заключается в расширении возможности интерпретации результатов экзаменов, поскольку тестовый балл можно дополнять содержательным описанием подготовки выпускников школ в терминах, освоенных и не освоенных тематических элементов предмета.

Для введения содержательной интерпретации в практику ЕГЭ необходимо, прежде всего, провести цикл работ по совершенствованию структуры ГОС на основе идей уровневой дифференциации [66] и содержательно описать различные уровни общеобразовательной подготовки по всем школьным предметам. Затем на репрезентативной выборке выпускников школ России следует построить устойчивую шкалу стандартных тестовых баллов и выделить диапазоны шкалы путем соотнесения их с содержанием заданий, выполненных на каждом уровне.

В терминах теории педагогических измерений сопоставимость означает возможность переноса результатов тестирования различных лет на единую шкалу. Сопоставимость обеспечивает выпускнику возможность однократной сдачи ЕГЭ и использования своих результатов при неудачном поступлении в вуз в последующие годы, а управленцам — возможность проведения сравнительных исследований качества образования.

Традиционный подход к решению проблемы сопоставимости основан на определении норм так называемого якорного теста для составления таблиц эквивалентности баллов по разным тестам. В качестве якорного выбирается тест, который стандартизуется на национальной репрезентативной выборке испытуемых, тщательно сформированной из генеральной совокупности в масштабах всей страны.

На основе якорного теста определяются репрезентативные общенациональные нормы по различным предметам. Данные, собранные на национальной выборке по якорному тесту, служат для построения шкалы единых показателей. Каждый новый тест, разрабатываемый в последующие годы, калибруется относительно якорного теста, в результате чего можно установить, какой

Федеральный государственный образовательный стандарт

результат испытуемого в последующие годы эквивалентен показателю в якорном тесте.

Для построения таблиц перевода в общем случае на одной и той же выборке испытуемых обычно используют метод равных перцентилей, согласно которому показатели считаются эквивалентными, если они имеют равные перцентили и получены на нормальной группе сравнения по параллельным вариантам тестов. К недостаткам метода можно отнести высокую стоимость, неизбежную коррекцию и совершенствование содержания тестов, изменение содержания образования и приоритетов в учебных достижениях, а к достоинствам — стабильность и высокую точность, поскольку результаты по всем последующим версиям тестов сравниваются с одним якорным тестом.

В связи с появлением теории IRT стали использовать более эффективные технологии, обеспечивающие сопоставимость результатов тестирования по различным годам. Эта технология основана на включении в тесты различных лет групп якорных (общих) заданий, связывающих цепочкой результаты по более поздним и более ранним версиям теста. Отсчет ведется от шкалы, построенной на эталонной группе сравнения. Тем самым каждый вариант очередной версии теста благодаря блоку общих заданий связывается с предыдущим и последующим вариантами батареи тестов.

Для пользователей тестов обычно разрабатываются необходимые разъяснительные материалы, обеспечивающие перевод сырых баллов в производные показатели, сопоставимые с результатами эталонной шкалы. Дополнительные меры по повышению сопоставимости результатов тестирования обеспечиваются специальной теорией выравнивания тестов. Аналогичные современные методы планируется использовать в ЕГЭ.

Филиал Восточный федерального университета им. М.К. Аммосова

14.5. Единый государственный экзамен и Общероссийская система оценки качества образования

Что понимают под Общероссийской системой оценки качества образования. Под Общероссийской системой оценки качества образования понимается совокупность организационных и функциональных структур, которая обеспечивает основанную на единой концептуально-методологической базе оценку образовательных достижений граждан, а также выявление факторов, влияющих на образовательные результаты [10; 11].

Основная цель создания ОСОКО заключается в повышении объективности и обоснованности информационной основы системы управления качеством образования в России, а также обеспечении всех участников образовательного процесса и об-

щества в целом достоверной информацией о качестве образования в стране.

В соответствии с поставленной целью к основным задачам ОСОКО можно отнести:

– оценку качества учебных достижений обучаемых и выпускников учебных заведений на различных образовательных ступенях;

– разработку научно-методического обеспечения процедур и инструментария для оценки качества образования;

– создание структур, обеспечивающих качество инструментария и технологий педагогических измерений (центров сертификации);

– создание центров статистической обработки и анализа данных о качестве образования, организацию системы подготовки и переподготовки педагогических кадров и т.д.

Процесс создания ОСОКО в России пока не завершен, поэтому число основных задач по мере ее становления может меняться.

Построение ОСОКО предполагает широкое использование существующих организационных структур, механизмов и процедур: ЕГЭ, центров сертификации, аттестации и аккредитации, центров качества образования и мониторинга, работающих на единой научно-методической основе в рамках стратегии развития общероссийской системы оценки качества образования.

ЕГЭ в ОСОКО. ЕГЭ является неотъемлемым структурным элементом ОСОКО и обеспечивает объективную информацию о реальном состоянии качества школьного образования. Данные ЕГЭ позволяют получить оценки состояния образовательных достижений выпускников школ, выявить тенденции и динамику изменения системы среднего образования, сформировать совокупность основных факторов, влияющих на качество образования в различных регионах, и оценить меру их влияния.

ЕГЭ не является единственным информационным источником ОСОКО. Несомненно, что помимо собранной в процессе экзамена информации о предметных знаниях и умениях выпускников школ необходимы дополнительные данные, характеризующие состояние качества образования на различных ступенях (оценки умений применять знания, собранные с помощью портфолио и других средств аутентичного оценивания; оценки междисциплинарных, общеучебных и коммуникативных умений; данные об уровне воспитанности обучаемых и результативности воспитательных воздействий; результаты участия учащихся в олимпиадах и творческих конкурсах и т.д.).

Важным требованием, обеспечивающим корректность использования данных ЕГЭ в ОСОКО, является учет различных факторов при интерпретации результатов экзамена, лежащих зачастую за пределами влияния школы, но существенно влияющих на качество образования.

Практическое задание и вопросы для обсуждения

1. Можно ли использовать оценки, полученные на ЕГЭ, при аттестации учителей школ и самих школ? Каковы, по вашему мнению, должны быть условия такого использования?

2. Проанализируйте содержание демонстрационных вариантов КИМ по предметам. Какие задания вы бы хотели включить дополнительно?

3. Стоит ли облегчить КИМ, чтобы уменьшить процент «двоечников» на ЕГЭ и привести его в соответствие с традиционной практикой не ставить двойки на выпускных экзаменах в школе?

4. Участвовали ли вы в ЕГЭ? Какие меры повышения дисциплины на ЕГЭ вы можете предложить?

Северо-Восточный федеральный Университет им. М.К.Аммосова

ГЛАВА
15
МОНИТОРИНГ КАЧЕСТВА ШКОЛЬНОГО ОБРАЗОВАНИЯ

15.1. Мониторинг в образовании, его достоинства и недостатки

Цели и функции мониторинга. Под мониторингом следует понимать систему постоянного сбора данных о наиболее значимых характеристиках качества образования, их обработку, анализ и интерпретацию с целью обеспечения общества и системы образования достоверной, достаточно полной и дифференцированной по уровням использования информацией о соответствии процессов и результатов образования нормативным требованиям, происходящих переменах и прогнозируемых тенденциях. Иначе говоря, мониторинг — это стандартизированное наблюдение за образовательным процессом и его результатами, позволяющее создавать историю состояния объекта во времени, количественно оценивать изменение субъектов обучения и образовательной системы, определять и прогнозировать направления их развития. Основная *цель* создания *системы мониторинга* — повышение качества образования [20].

К составляющим мониторинга относятся: объекты и субъекты образовательного процесса, комплекс показателей качества образования, инструментарий, базы данных для накопления информации, методики анализа, переработки и интерпретации информации, программно-инструментальные средства обработки данных. Ведущими *функциями мониторинга* в образовании являются: информационная, диагностическая, сравнительная и прогностическая. Основная *сфера практического применения мониторинга* — информационное обеспечение управления качеством образования, осуществляемого различными методами, в том числе и прямым административным вмешательством.

Информация, накапливаемая в системе мониторинга, может использоваться для идентификации проблем в обучении, связанных с недостатками в методах преподавания, искажениями в пропорциях учебных планов, просчетами авторов школьных учебников и др. Она помогает оценивать последствия инноваций в образовании, осуществляемых в государстве, регионе, районе или внутри отдельной школы. Данные мониторинга мотивируют руковод-

Северо-Восточный федеральный университет им. М.К. Аммосова

ство школ и преподавателей к улучшению своей деятельности и способствуют повышению ответственности за результаты учебного процесса.

Достоинства и недостатки мониторинга. Совокупность показателей мониторинга всегда согласована с наиболее общими тенденциями в образовании, выработанными правительством страны и другими органами управления образованием. Сбор обширного количества данных и накопление описательной (дескриптивной) статистики не только важны для образования, но и имеют политический смысл.

Статистика, описывающая состояние системы образования, может использоваться для демонстрации потребности в образовательных реформах, служить доказательством плохой работы предыдущего управленческого состава в образовании или показывать преимущества проводимых в государстве реформ. Некоторые критики систем мониторинга утверждают, что аналитики специально выбирают для сообщения нужную информацию, которая зависит от политических целей в образовании и способствует усилению централизованного контроля со стороны органов управления образованием за результатами обучения.

Манипулирование данными мониторинга становится затруднительным, если они сравниваются с некоторыми стандартными количественными критериями или нормами выполнения тестов, установленными на уровне школы, района или региона. Подобное сопоставление будет давать обоснованные результаты при условии статистической коррекции данных мониторинга с учетом дополнительных факторов, выравнивающих показатели каждой школы относительно средних показателей по району.

Мониторинг позволяет решать многие повседневные задачи диагностического характера. Данные, накапливаемые в школьном мониторинге, помогают выявить систематические трудности в усвоении отдельных разделов дисциплин, оценить эффективность инновационных методов работы учителей, диагностировать причины неудач отдельных учащихся, обоснованно связав их с предметными, социально-экономическими или другими факторами. В целом школьная система мониторинга обеспечивает обратную связь, позволяющую судить о сильных и слабых аспектах системы обучения.

Условия эффективного проведения мониторинга. Специалисты по мониторингу выделяют ряд условий его эффективности, среди которых:

– использование системного подхода, обеспечивающего слаженную работу механизма по сбору, обработке, анализу и интерпретации информации;

– сочетание количественных и качественных методов измерения в мониторинге;

– репрезентативная совокупность показателей мониторинга, учет различных, в том числе и косвенно влияющих на результаты обучения, факторов;

– корректная интерпретация данных мониторинга с учетом различных влияний и связей между показателями;

– репрезентативность выборочных совокупностей учащихся, принимающих участие в мониторинге;

– привлечение к проведению мониторинга квалифицированных специалистов и преподавателей школ;

– использование качественного инструментария и современного программного обеспечения для обработки и анализа данных мониторинга;

– методическая и финансовая помощь со стороны органов управления образованием разного уровня.

15.2. Виды мониторинга

Классификация видов мониторинга. В основу классификации видов мониторинга могут быть положены разные основания, к числу которых относятся: цели проведения мониторинга; его основные функции; область применения данных; инструментарий; модель или технология проведения мониторинга и др.

Чаще всего классификация видов мониторинга проводится в соответствии с его основными функциями: информационной, диагностической, сравнительной и прогностической.

Информационный мониторинг. Проведение информационного мониторинга нацелено на сбор, накопление, анализ, структуризацию и интерпретацию данных по выделенной совокупности показателей при условии, что анализ носит не сопоставительный или прогностический, а констатирующий характер. Отличительной чертой информационного мониторинга является отсутствие анализа эффектов связи и взаимного влияния показателей, сопоставления результатов мониторинга на различных уровнях управления качеством образования, выявления тенденций в образовании и прогнозирования их влияния на качество образования.

В рамках информационного мониторинга анализ направлен на выявление степени согласованности данных с некоторыми нормами и стандартами. Например, анализ данных мониторинга аттестационного тестирования выпускников школ нацелен на установление соответствия учебных достижений выпускников требованиям ГОС [36; 50]. При анализе таких данных следует принимать во внимание средние размеры классов по школам внутри района, количество учащихся, приходящихся на одного преподавателя в школе, ежегодные расходы на учебные материалы, размер библиотечных фондов, данные о квалификации преподавателей в

Полесский государственный университет имени А.Кулешова

школах, число сотрудников вспомогательного состава и т.д. Эти и другие данные в усредненном виде могут быть отнесены к некоторым районным или областным стандартам школьного образования.

Диагностический мониторинг. Диагностический мониторинг предназначен для определения того, как справляются с различными темами или разделами учебного плана большинство учащихся. Диагностический мониторинг может проводиться на различных уровнях. Преподаватели выявляют проблемы усвоения учебного материала и осуществляют деятельность по диагностическому мониторингу на уровне класса. В районе диагностические системы мониторинга нацелены на определение отдельных слабо усвоенных умений и навыков.

Сбор данных для диагностического мониторинга обычно проводится с помощью педагогических измерений. В качестве основного инструментария используются корректирующие критериально-ориентированные тесты, которые сопровождаются диагностическими тестами для установления причин пробелов в усвоении учебного материала.

В диагностическом мониторинге не учитываются входные характеристики учащихся, поскольку главная его цель состоит в том, чтобы идентифицировать сильные и слабые стороны в учебных достижениях и образовательной деятельности независимо от характеристик учащихся и их возможностей усвоения материала. Поэтому данные диагностического мониторинга не используют для проведения сравнений между школами или районами.

Сравнительный мониторинг. Сравнительный мониторинг отличается от других видов мониторинга специфическим анализом данных, который направлен на сопоставление количественных оценок по совокупности показателей для регионов, областей, районов, школ, отдельных преподавателей и других участников образовательной деятельности. Сравнения проводятся либо по вертикали (территории, регионы, образовательные учреждения), либо по горизонтали (рейтинг школьников, рейтинг территорий и т.д.) на основе анализа количественных оценок по одинаковым показателям и с учетом различных факторов, смещающих оценки. По результатам сравнительного мониторинга обычно принимаются административные решения.

При проведении сравнений особое значение имеет анализ связей между показателями и их взаимного влияния. Необходимы доказательства достаточной полноты совокупности показателей, используемой в сравнительном мониторинге, и репрезентативности выборочных совокупностей учащихся, принимающих участие в проведении мониторинга. Специфика анализа данных в сравнительном мониторинге и особая ответственность за его результаты вынуждают предъявлять повышенные требования к качеству ин-

Северо-Восточный федеральный университет имени М.К. Аммосова

струментария, его надежности и валидности. В сравнительном мониторинге обычно используют количественные шкалы, стандартизированные тесты учебных достижений и профессионально разработанные анкеты для сбора дополнительной информации о факторах, находящихся вне деятельности школ, но существенно влияющих на результаты их образовательной деятельности.

Проведение сравнительного мониторинга нередко включает измерения входных данных и итоговых результатов обучения, поскольку многие из них значимо влияют на итоговые учебные достижения учащихся. Хотя возможности мониторинга этого вида у многих учителей вызывают сомнения и часто являются предметом критики, мировой опыт свидетельствует о том, что сравнительный мониторинг приводит к явным позитивным следствиям. Межшкольные, межрайонные или межрегиональные сравнения стимулируют соревнование и мотивируют педагогов к улучшению образовательного процесса.

В сравнительном мониторинге обычно используют нормативно-ориентированные тесты учебных достижений, стандартизированные на представительных выборках в масштабах всей страны или региона. Таким образом, средние оценки по школам или районам могут быть сопоставлены с региональными и национальными нормами. Наибольшие трудности при проведении сравнительного мониторинга связаны со сбором дополнительной информации по ряду показателей, находящихся вне сферы влияния школ, но требующих учета при проведении межшкольных, межрайонных и других сравнений. Выбор этих показателей должен учитывать социальные и экономические особенности региона или района, организационные механизмы формирования состава школ и другие факторы, влияющие на качество образования.

Прогностический мониторинг. Прогностический мониторинг предназначен для выявления и предсказания позитивных и негативных тенденций в развитии образовательных систем. Он очень важен для решения управленческих задач в образовании, связанных с формированием социального заказа и соответствующих потенциалу системы образования. Роль прогностического мониторинга в условиях реформирования российского образования неуклонно возрастает в силу изменений, происходящих в характере управленческих решений на разных уровнях иерархии [5]. Если раньше приоритет принадлежал оперативным управленческим решениям, направленным на текущее функционирование образовательных структур, то сейчас на первый план нередко выходят стратегические решения, нацеленные на развитие системы образования.

Опыт реформирования в образовании показал, что стратегические решения зачастую приходится принимать в условиях отсутствия необходимой информации, в результате не учитываются

Алтайский государственный педагогический университет им. В.М. Шукшина

все возможные последствия проводимых нововведений и преобразований. Примером тому может служить предоставление образовательным учреждениям права на оказание платных услуг, которое, с одной стороны, сыграло положительную роль, а с другой — имело негативные, неучтенные последствия. В частности оно способствовало коммерциализации общего образования, расширению практики использования денежных средств родителей.

При наличии необходимых информационных условий для разработки стратегии на основе анализа внутрисистемных и межсистемных противоречий можно предвидеть многие негативные последствия реформаторства и принять упреждающие меры на момент осуществления нововведений. Стратегические управленческие решения, направленные на развитие системы образования, должны основываться на вероятностной оценке тенденций в образовании, которые следует получать с помощью прогностического мониторинга. Для прогноза обычно широко используются методы регрессионного анализа и другие математико-статистические методы.

Рассмотренные виды мониторинга редко встречаются в практике образования в чистом виде. Их выделение имеет смысл для описания требований к проведению, обеспечивающих корректность сбора, анализа и интерпретации данных. Как правило, в образовании проводят комплексный мониторинг, сочетающий в различных пропорциях перечисленные выше виды.

15.3. Модели проведения мониторинга

Модель соответствия нормам и стандартам. Наиболее простая модель мониторинга нацелена на сбор данных о процессе и результатах образовательной деятельности, включая их анализ путем сопоставления с установленными нормами и стандартами. Достоинство модели соответствия нормам и стандартам — в простоте и оперативности реализации. Недостатки модели — ограниченные возможности интерпретации результатов мониторинга, связанные с отсутствием достаточного числа характеристик образовательной деятельности школ, процессов ее протекания и входных данных об учащихся. В силу отсутствия важной информации модель соответствия нормам и стандартам не позволяет корректно сопоставить результаты обучения и сделать обоснованные выводы для управления качеством образования.

Модель «вход—выход». Классическая модель «вход—выход», используемая в большинстве стран в системах информационного и сравнительного мониторинга, основана на предположении о том, что входные данные учащихся существенно влияют на результаты их обучения в школе. К входным данным относят сово-

Гомельский государственный университет имени Ф. Скорины

купность показателей, характеризующую начальные способности учащегося, социально-экономический статус семей учащихся, ресурсы школы (профессиональный уровень преподавателей школ, размеры финансирования на одного учащегося и т. д.). Выходные данные включают экзаменационные оценки учебных достижений учащихся и выпускников школ и полный перечень освоенных ими знаний и умений.

Учет входных данных позволяет выделить однородные группы школ, начинающих работу в одинаковых условиях, что гарантирует корректность внутригрупповых сравнений по конечным результатам образовательной деятельности. Сравнение школ с различными входными данными проводится с помощью уравнений множественной регрессии. Таким образом, модель «вход — выход» расширяет возможности мониторинга и позволяет проводить внутригрупповые и межгрупповые сравнения образовательных учреждений.

Модель «вход — процесс — выход». В настоящее время большое распространение получила так называемая модель с пятью факторами (модель «вход — процесс — выход»). Она включает в себя комбинацию характеристик не только результатов, но и процесса обучения. К факторам этой модели относятся:

1) сильное административное руководство;

2) наличие стабильного и хорошо организованного внутреннего микроклимата в классе и в школе;

3) преимущественная ориентация в обучении на формирование у учащихся базовых академических навыков;

4) повышенные требования к учащимся со стороны преподавателей, ориентирующие учащихся на постоянный прирост учебных достижений;

5) наличие системы внутришкольного мониторинга, обеспечивающей достоверную информацию о качестве образования.

Таким образом, улучшенная модель системы мониторинга образовательной деятельности — модель «вход — процесс — выход» — включает информацию относительно процессов, протекающих в школе при обучении и воспитании учащихся. В основе расширенной модели лежит предположение о том, что совершенствование процесса обучения запланированным образом неизбежно должно привести к более высоким учебным достижениям, что вполне отвечает реалиям в практике образования.

Динамическая модель мониторинга. Большинство сравнительных оценок деятельности школ основывается на межгрупповом анализе и сопоставлении средних оценок учебных достижений, существенно зависящих от контингента учащихся школы и социально-экономической характеристики района, в котором расположена школа. В силу социальной неоднородности заселения различных районов некоторые школы из года в год получают значи-

Гродненский государственный университет имени Я. Купалы

тельную долю неблагополучного контингента учащихся. Поэтому достижения таких школ при сравнении по конечным результатам обучения могут оставаться незамеченными годами, несмотря на эффективную образовательную деятельность и усилия педагогического коллектива. Преодолеть этот недостаток можно, используя динамическую модель мониторинга, основанную на анализе динамики изменений в образовательной деятельности школы, выявлении их позитивного или негативного характера и степени влияния школы на происходящие изменения.

Наиболее эффективная динамическая модель мониторинга строится на измерениях скорости прироста учебных достижений учащихся в течение некоторого временного периода. Поскольку на темп развития учащихся существенно влияют начальные способности и характеристики семьи, то учет этих показателей обеспечивается без дополнительных усилий при измерениях скорости прироста учебных достижений. В этой связи многие специалисты настоятельно советуют использовать динамическую модель мониторинга, основанную на измерениях изменений в учебных достижениях.

Наработки в статистике, теории педагогических измерений и дизайне тестов в последние десятилетия значительно усиливают преимущества динамического мониторинга. Благодаря теории IRT у создателей тестов появилась возможность построения батарей тестов учебных достижений, включающих отдельные субтесты для учащихся на различных образовательных уровнях без потери сопоставимости результатов тестирования. С этой целью в тесты для двух смежных классов обычно включают блок якорных заданий, одинаковых для обоих тестов. Таким образом достигается связка при построении шкал учебных достижений предыдущих и последующих образовательных уровней. Благодаря этой связке результаты мониторинга размещаются на одной непрерывной шкале, которая охватывает длительный период обучения в школе.

Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования имени Н.А. Некрасова

15.4. Этапы и уровни проведения мониторинга качества образования, пользователи и исполнители, доступ к информации

Основные этапы мониторинга. При проведении мониторинга, как и при любом эмпирическом исследовании, можно выделить три основных этапа, среди которых:

– подготовка исследования;
– сбор информации;
– обработка информации, ее анализ и интерпретация.

Данные этапы носят обобщенный характер, их конкретизация зависит от целей мониторинга, его вида, выбранной модели проведения и управленческого уровня, на котором принимаются ре-

шения (школьный, районный, областной/региональный, федеральный). По мере увеличения охвата учащихся и масштабов в управлении качеством образования развернутый перечень этапов будет меняться в направлении увеличения их числа, а сама система мониторинга будет расширять свои функции и приобретать комплексный характер.

Этапы мониторинга для областного уровня управления. Конкретизация этапов мониторинга для регионального или областного уровней управления включает следующие этапы.

1. Подготовка исследования:

– выбор целей создания системы мониторинга, круга пользователей и вида мониторинга;

– подготовка программы работ и плана их проведения, выбор исполнителей;

– определение модели мониторинга;

– формирование совокупности показателей, их операционализация, распределение по количественным и качественным уровням;

– выбор методов сбора данных, обеспечивающих сочетание количественных и качественных уровней анализа информации;

– разработка инструментария для сбора данных (тестов, анкет и т.д.);

– формирование репрезентативных выборочных совокупностей учащихся для участия в мониторинговых исследованиях;

– проведение пилотных исследований качества инструментария (его надежности и валидности), коррекция инструментария, его стандартизация;

– выбор шкал и методов шкалирования и интеграции данных мониторинга;

– разработка программного обеспечения для обработки данных мониторинга;

– разработка программно-инструментального обеспечения для ввода данных;

– разработка структуры баз данных мониторинга и средств для их ведения.

2. Сбор информации:

– проведение тестирования и анкетирования;

– проведение собеседований;

– работа с документами;

– организация процедур наблюдений за сбором информации для обеспечения ее достоверности;

– организация процедур соблюдения конфиденциальности информации.

3. Обработка информации, ее анализ, интерпретация:

– анализ и чистка сырых данных, обработка данных мониторинга;

Северо-Восточный федеральный университет им. М.К. Аммосова

– оценивание надежности и валидности данных, анализ генерализуемости;

– коррекция и выравнивание данных для обеспечения сопоставимости по группам сравнения;

– шкалирование данных мониторинга;

– анализ данных;

– интерпретации результатов анализа;

– подготовка заключительного отчета по использованию результатов мониторинга в соответствии с целями его проведения.

Для проведения мониторинга внутри школ перечень этапов должен быть существенно сокращен. В частности должна исчезнуть разработка инструментария, поскольку делать его имеет смысл лишь профессионалам в городе или области и стандартизовать на таких же выборках для обеспечения корректности межшкольных сравнений. Не нужны также процедуры обработки, анализа и интерпретации данных, которые следует проводить в специальных информационно-диагностических центрах, и т.д.

Представление данных. При использовании результатов мониторинга на каждом уровне (например, при формировании общественного мнения относительно ЕГЭ) необходимо согласовать формат и перечень предоставляемых данных с возможностями восприятия той группой пользователей, для которой эта информация предназначена. Под согласованием понимается выбор формата представления данных мониторинга, их объема, языка изложения и необходимых поясняющих материалов для интерпретации.

Информационная безопасность. Для обеспечения информационной безопасности данных мониторинга следует учитывать различные аспекты использования информации в районе или в регионе. К ним относятся:

– хранение информации в виде баз данных и на бумажных носителях;

– систематизация информация, ее структурирование;

– статистическая обработка информации;

– анализ результатов обработки;

– интерпретация данных анализа;

– принятие управленческих решений.

15.5. Показатели качества образования и эффективности образовательной деятельности школ

Выбор показателей качества образования. Для реализации выбранной модели мониторинга необходимо составить достаточно полный набор показателей, характеризующих качество образования и эффективность деятельности образовательных учреждений.

Северо-Осетинский государственный университет им. М.К. Аметова

Основания по выбору и структурированию совокупности показателей могут быть самыми разными, но в целом совокупность показателей должна:

– обладать достаточной полнотой;

– быть в основной своей части операционализируемой;

– быть признанной и полезной на различных уровнях управления;

– обеспечивать сбор и сообщение информации относительно образовательной деятельности в соответствии с задачами, решаемыми в управлении качеством образования;

– быть нацеленной на ту информацию, которая обладает прогностическими возможностями и является значимой на протяжении нескольких лет;

– обеспечивать надежность, простоту, экономическую целесообразность сбора информации.

В мониторинге качества образования общую совокупность показателей можно разделить на следующие подгруппы:

– показатели качества функционирования образовательных систем;

– показатели качества учебного процесса;

– показатели качества результатов обучения;

– показатели качества инновационной деятельности (использование информационных технологий, передового опыта педагогической науки и практики) при обучении;

– показатели качества, характеризующие материальные вложения в образование;

– показатели качества управления образовательными учреждениями.

Первая группа показателей. Показатели первой группы характеризуют: качество содержания образования, структуры, образовательных программ и форм организации учебного процесса; эффективность реализации целей обучения и воспитания; наличие развитых педагогических технологий, качество работы системы подготовки и переподготовки педагогических кадров и др.

Необходимо также принимать во внимание группу обобщенных показателей качества образовательных систем, к которым можно отнести характеристики оперативности принятия управленческих решений, наличие информационных возможностей, эффективность прямой и обратной связи в системах, прогнозируемость тенденций развития образовательной системы и т. д.

Вторая группа показателей. Показатели второй группы характеризуют уровень доступности и индивидуализации обучения; качество организации образовательного процесса, условия и атмосферу преподавания; вариативность учебных программ; соотношение традиционных и инновационных технологий обучения и контроля; соответствие структуры и содержания обучения акту-

Учебно-методический центр Федеральный университет имени К. Жумабаева

альным тенденциям теории и практики образования; степень внедрения компьютерных методов в обучение, характеристики деятельности образовательного учреждения в основное (урочное) и не основное (внеурочное) время и т.д.

Третья группа показателей. Третью группу составляют показатели, характеризующие результаты контроля и оценки учебных достижений учащихся (внутренние и внешние, полученные в процессе аттестации); качество учебных достижений обучаемых; уровень развития общеучебных и коммуникативных умений; подготовку выпускника школы к жизни в современном обществе высоких технологий; наличие сформированной базы для продолжения образования, умения решать проблемы жизненного характера и т.д.

К остальным группам показателей качества образования относятся характеристики использования передового опыта педагогической науки и практики при обучении, размеры финансирования образования, кадровое, информационное, материально-техническое и методическое обеспечение.

Дополнительные показатели. При оценке качества образования для принятия управленческих решений необходимо уделять внимание дополнительным показателям, характеризующим социально-экономическое развитие страны, региона, области, района, являющимся следствием национальных процессов и связанным с качеством образования.

К ним можно отнести: индекс развития человеческого потенциала (процент грамотного взрослого населения, количество обязательных лет обучения в школе основной доли взрослого населения, количество людей с высшим образованием на 1000 человек); обобщенную характеристику экономической динамики в регионе; характеристику криминогенной обстановки в районе и многие другие.

Структурирование показателей, влияющих на качество образования. Все показатели, влияющие на качество образования, по разным основаниям можно разделить на группы. В частности выделяются не изменяющиеся или слабо изменяющиеся со временем факторы (географическое положение школы, ее расположение в городе или в селе, определенном районе города, социально-экономическое окружение образовательного учреждения или социально-экономические характеристики региона и др.) и факторы, подверженные динамичным изменениям (педагогический состав, формы и методы дополнительного образования, уровень обеспеченности родителей учащихся, их социальный статус, миграционные процессы среди населения России и др.).

Показатели подразделяются на поддающиеся влиянию со стороны органов управления образованием, изменение которых можно вести целенаправленно для повышения качества образования, и инвариантные, не зависящие от управленческих воздействий.

Сетевой Восточный Федеральный университет им. М.К. Аммосова

В первую группу, например, входят: использование передового опыта педагогической науки и практики, размеры финансирования образования, кадровое, информационное, материально-техническое обеспечение. Ко второй группе можно отнести: уровень обеспеченности родителей учащихся, их социальный статус, миграционные процессы среди населения России.

Практическое задание и вопросы для обсуждения

1. Перечислите наиболее значимые показатели качества школьного образования.

2. Считаете ли вы, что мониторинг служит усилению воздействия централизованного управления на систему образования и ограничивает свободу учителя? Приведите аргументы за или против этого утверждения.

3. Какие элементы системы мониторинга существуют в школах и какие следует создавать заново?

4. Можно ли использовать оценки, полученные традиционным путем, в системах мониторинга качества образования? Аргументируйте свою точку зрения.

Северо-Восточный федеральный Университет им. М.К.Аммосова

ЛИТЕРАТУРА

1. *Анастази А.* Психологическое тестирование : в 2 т. / пер. с англ. ; предисл. К. М. Гуревича, В. И. Дубовского. — М., 1982.

2. *Андреев А. Б.* Компьютерное тестирование : системный подход к оценке качества знаний студентов. Усиление практической подготовки студентов к деятельности в условиях рыночной экономики // Материалы Всероссийского практического семинара ректоров (проректоров) вузов. — 2001. — № 21. — Пенза, 2001.

3. *Атаманчук Г. В.* Управление — фактор развития. Размышления об управленческой деятельности. — М., 2002.

4. *Балыхина Т. М.* Словарь терминов и понятий тестологии. — М., 2000.

5. *Басова А. Л.* Перспективы использования методов тестирования абитуриентов для прогноза успешной учебной деятельности в вузах с различными формами обучения. Психология и социология образования. — М., 2001.

6. *Берка К.* Измерения, понятия, теория, проблемы. — М., 1987.

7. *Беспалько В. П.* Программированное обучение : Дидактические основы. — М., 1970.

8. *Боголюбов Л. Н.*, *Дик Ю. И.*, *Иванова Е. О.*, *Ковалева Г. С.*, *Красновский Э. А.*, *Челышкова М. Б.*, *Шмелев А. Г.* О подходах к разработке требований к обязательному уровню подготовки выпускников основной школы. Перспективы развития общего среднего образования: сб. науч. тр. — М., 1998.

9. *Болотов В. А.*, *Шаулин В. Н.*, *Шмелев А. Г.* Единый экзамен и качество образования. — М., 2002.

10. *Болотов В. А.* Единый государственный экзамен : на пути к созданию системы независимой оценки качества образования // Единый государственный экзамен : сб. ст. — М., 2004.

11. *Болотов В. А.* Основные подходы к созданию общероссийской системы оценки качества образования // Единый государственный экзамен : сб. ст. — М., 2005.

12. *Бондаревская Е. В.* Гуманистическая парадигма личностно ориентированного образования // Педагогика. — 1997. — № 4.

13. *Бондаревская Е. В.*, *Кульневич С. В.* Педагогика : личность в гуманистических теориях и системах воспитания. — М.; Ростов н/Д, 1999.

14. *Бордовская Н. В.*, *Реан А. А.* Педагогика : учеб. для вузов. — СПб., 2001.

15. *Венцель К. Н.* Свободное воспитание. — М., 1993.

16. *Выготский Л. С.* Избранные психологические исследования. — М., 1956.

17. *Гальперин П. Я.*, *Решетова З. А.*, *Талызина Н. Ф.* Психолого-педагогические проблемы программированного обучения на современном этапе : материал к Всесоюзной конференции по программированному обучению. — М., 1966.

РЕПОЗИТОРИЙ ВОСТОЧНО-КАЗАХСТАНСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА ИМЕНИ С. АМАНЖОЛОВА

18. *Гласс Дж.*, *Стенли Дж.* Статистические методы в психологии педагогики / пер. с англ. Л.И.Хайрусовой. — М., 1976.

19. *Гутгарц Р.Д.* Особенности дистанционного тестирования в Интернете. Современные проблемы экономики региона : сб. науч. тр. — Иркутск, 2001.

20. *Ефремова Н.Ф.* Современные тестовые технологии в образовании. — М., 2003.

21. *Звонников В.И.*, *Найденова Н.Н.*, *Никифоров С.В.*, *Челышкова М.Б.* Шкалирование и выравнивание результатов педагогических измерений : учеб. пособие. — М., 2003.

22. *Звонников В.И.* Измерение и шкалирование в образовании. — М., 2006.

23. *Зимняя И.А.* Педагогическая психология. — М., 2005.

24. Из опыта разработки качественных и количественных характеристик знаний, умений и навыков : сб. научн. трудов НИИ содержания и методов обучения АПН РСФСР. — М., 1977 (48).

25. Качество знаний учащихся и пути его совершенствования / под ред. М.Н.Скаткина, В.В.Краевского. — М., 1978.

26. *Клайн П.* Введение в психометрическое программирование : справочное руководство по конструированию тестов. — Киев, 1994.

27. *Кларин М.В.* Инновационные модели обучения в зарубежных педагогических поисках. — М., 1994.

28. *Ковалева Г.С.* Зарубежный опыт построения и актуальные проблемы развития тестовых систем. Российский и зарубежный опыт построения систем образовательного тестирования : материалы к семинару «Актуальные проблемы построения системы национальных образовательных стандартов и тестирования». — М., 2000.

29. *Ковалева Г.С.*, *Красновский Э.А.*, *Краснокутская Л.П.*, *Краснянская К.А.*, *Кошеленко Н.Г.*, *Смирнова Е.С.* Результаты российских учащихся в международном исследовании PISA-2000 : материалы международного исследования PISA-2000 «Новый взгляд на грамотность». — М., 2004.

30. Концепция оценки достижения учащимися требований общеобразовательного стандарта / под ред. В.С.Леднева. — М., 1993.

31. *Кривошеев А.О.* Разработка и использование компьютерных обучающих программ // Информационные технологии. — 1996. — № 4.

32. *Кулибаба И.И.*, *Красновский Э.А.*, *Коган Т.Л.* Дидактический анализ качества знаний учащихся. Проблемы и методы исследования качественных и количественных характеристик знаний, умений и навыков учащихся. — М., 1976.

33. *Лернер И.Я.* Качества знаний учащихся. Какими они должны быть? — М., 1978.

34. *Логвиненко А.Д.* Измерения в психологии : математические основы. — М., 1993.

35. *Майоров А.Н.* Теория и практика создания тестов для системы образования. — М., 2001.

36. *Масленников А.С.* Разработка методики проведения педагогических измерений уровня подготовки выпускников в условиях аттестации учеб-

Северо-Восточный федеральный университет им. М.К.Аммосова

ных заведений среднего и высшего профессионального образования / сб. мат. по программе «Научное и научно-методическое функционирование развития системы образования : в 3 ч. — М., 2003. — Ч. 3.

37. *Михайлычев Е. А.*, *Карпова Г. Ф.*, *Леонова Е. Е.* Педагогическая диагностика: история, теория, современность. — Ростов н/Д, 2002.

38. *Найдёнова Н. Н.* Формирование репрезентативной выборки : учеб. пособие. — М., 2003.

39. *Нардюжев В. И.*, *Нардюжев И. В.* Модели и алгоритмы информационно-вычислительной системы компьютерного тестирования. — М., 2000.

40. Народное образование в СССР : Общеобразовательная школа : сб. док. (1917—1973). — М., 1974.

41. Национальные экзамены в системе оценки качества образования. Материалы и тезисы докладов Международной конференции. — М., 2006.

42. Общая психодиагностика / под ред. А. А. Бодалева, В. В. Столина. — М., 1987.

43. *Поддубная Л. М.*, *Татур А. О.*, *Челышкова М. Б.* Задания в тестовой форме для автоматизированного контроля знаний студентов. — М., 1995.

44. *Пойа Дж.* Математика и правдоподобные рассуждения / под ред. С. А. Янковской. — М., 1975.

45. *Полонский В. М.* Оценка знаний школьников. — М., 1981.

46. Проблемы педагогической квалиметрии : сб. тр. / пер. с англ. Л. И. Хайрусовой. — М., 1976.

47. Психологические проблемы неуспеваемости школьников / под ред. Н. А. Менчинской. — М., 1971.

48. *Родионов Б. У.*, *Татур А. О.* Стандарты и тесты в образовании. — М., 1995.

49. *Селевко Г. К.* Технологии развивающего обучения. — М., 1998.

50. *Селезнева Н. А.*, *Байденко В. И.* Проблема качества образования: актуальные аспекты пути решения // Проблемы качества, его нормирования и стандартов в образовании : сб. науч. статей. — М., 1998.

51. *Симонов В. П.* Педагогический менеджмент : учеб. пособие. — М., 1997.

52. *Соколов В. М.* Роль и место тестов достижений в диагностике качества образования // Вестник Нижегородского ун-та. — Н. Новгород, 2006.

53. *Талызина Н. Ф.* Педагогическая психология. — М., 2003.

54. *Талызина Н. Ф.* Теоретические основы контроля в учебном процессе. — М., 1983.

55. *Унт И.* Индивидуализация и дифференциация обучения. — М., 1990.

56. Управление развитием школы : пособие для руководителей образовательных учреждений / под ред. М. М. Поташника, С. В. Лазорева. — М., 1995.

57. *Фирсов В. В.* От базисного плана к стандартам образования // Учительская газета. — 1992. — № 52.

58. *Цетлин В. С.* Неуспеваемость школьников и ее предупреждение. — М., 1977.

59. *Челышкова М. Б.* Адаптивное тестирование в образовании. — М., 2000.

60. *Челышкова М. Б.* Теория и практика конструирования педагогических тестов. — М., 2001 (15).

Социально-Восточный Университет им. М.А. Шолохова

61. *Челышкова М. Б.*, *Шмелев А. Г.* Шкалирование результатов Единого госэкзамена : проблемы и перспективы // Вопросы образования. — 2004. — № 2.

62. *Шмелев А. Г.*, *Бельцер А. И.*, *Харцонов А. Г.*, *Серебряков А. Г.* Адаптивное тестирование знаний в системе «Телетестинг». Школьные технологии. — М., 2001.

63. *Шторм Р.* Теория вероятностей и математическая статистика. Статистический контроль качества. — М., 1970.

64. *Якиманская И. С.* Развивающее обучение. — М., 2000.

65. Applying the Rasch Model: Fundamental Measurement in the Human Sciences / Trevor G. Bond. Christine M. Fox. — N.-J. : Lawrence Erlbaum Associates, 2001.

66. Automated Essay Scoring: A Cross-Disciplinary Perspectives / Ed. by Mark D/ Shermis. Jill Burstein. — N.-J. : Lawrence Erlbaum Associates, 2003.

67. *Baker F. B.* Item Response Theory : Parameter Estimation Techniques. — ASC. Univ. Ave, 2004.

68. *Bateson D.*, *Nicol C.*, *Achroeder T.* Alternative Assessrnent and Tables of Specification for the Third International Mathematics and Science Study. — ICC 64, 1991.

69. *Binet*, *A. Simon T. H.* The development of intelligence in young children. — N.J. : The Training School, 1936.

70. *Blaine R. Wortnen*, *Karl R. White*, *Xitao Fan*, *Richard R.* Sudweeks. Measurement and Assessment in Schools. — N-Y., 1999.

71. *Bloom B. S.* Taxonomy of Educational Objectives : The Classification of Educational Goals. Handbook 1: Cognitive Domain. — N.Y.: David Mc Kay Co, 1956.

72. Computerized Adaptive Testing : Theory and Practice / Ed. by Wim J. van der Linden and Cees A.W. Glas. — London: Kluwer academic publishers, 2003.

73. *Crocker L., Algina J.* Introduction to Classical and Modern Test Theory. — Tallahassy : Univ. of Florida. HBJCP, 1986.

74. Educational Measurement / Ed. by R. L. Linn. — N.-Y. : Macmillan, 1989.

75. Handbook of Modern Item Response Theory / Ed. by Wim J. van der Linden and Ronald K. Hambleton. — ASC, Univ. Ave, 1997.

76. IRT from SSI : Bilog-mg Multilog. Parscale Testfast / Edited bu Mathilda du Toit. — Scientific Software International, 2003.

77. *Parshall C.*, *Spray J.*, *Kalohn J.*, *Davey T.* Springer. Practical considerations in computer based testing. — N.-Y. : SAGE Publication, 2002.

78. Software and Books for all Your Measurement Needs. Catalog Assessment Systems Corporation. — 1996.

79. *Steven J.* Osterlind. Constructing Test Items : Multiple-Choise. — Constructed-Response, Performance, and Other Formats. — Columbia : University of Missouri-Columbia, 2004.

80. Test Theory : A Unified Treatment / Ed. by Roderick McDonald. — N.-J. : Lawrence Erlbaum Associates, 1999.

81. *Walsh W. Bruce*, *Betz Nancy E.* Test Assessment. — N.-J. : Prentice Hall, 2001.

82. *Weiss D. J.* (Ed.) New Horizons in testing. — N.-Y. : Academic Press, 1983.

83. *Wright B. D.*, *Stone M. H.* Best Test Design. — Chicago : Messa Press, 1984.

Учебно-методический кабинет филологического факультета МГУ имени А.А. Кулешова

СОДЕРЖАНИЕ

Северо-Восточный федеральный университет им. М.К. Аммосова

Северо-Восточный федеральный университет им. М.К. Аммосова

Электронный архив Восточный университет им. М.К. Аммосова

Северо-Восточный федеральный Университет им. М.К.Аммосова

Северо-Восточный федеральный университет им. М.К. Аммосова

Учебное издание

Звонников Виктор Иванович,
Челышкова Марина Борисовна

Современные средства оценивания результатов обучения

Учебное пособие

Редактор *И. В. Пучкова*
Технический редактор *О. Н. Крайнова*
Компьютерная верстка: *Р. Ю. Волкова*
Корректоры *Э. Г. Юрга, Т. Г. Дмитриева*

Изд. № 103109593. Подписано в печать 11.11.2008. Формат 60×90/16.
Гарнитура «Таймс». Бумага офсетная. Печать офсетная. Усл. печ. л. 14,0.
Тираж 3000 экз. Заказ №

Издательский центр «Академия». www.academia-moscow.ru
Санитарно-эпидемиологическое заключение № 77.99.02.953.Д.004796.07.04 от 20.07.2004.
117342, Москва, ул. Бутлерова, 17-Б, к. 360. Тел./факс: (495)330-1092, 334-8337.

Отпечатано в полном соответствии с качеством диапозитивов, предоставленных издательством, в ОАО «Саратовский полиграфкомбинат». www.sarpk.ru
410004, г. Саратов, ул. Чернышевского, 59.