НА ФАКУЛЬТЕТЕ ПСИХОЛОГИИ

Ю.П. Зинченко, Н.Б. Березанская, И.А. Володарская, О.А. Тихомандрицкая, А.Г. Шмелев

ОПЫТ ВНЕДРЕНИЯ КОМПЬЮТЕРИЗИРОВАННЫХ ТЕСТОВЫХ ИСПЫТАНИЙ В СИСТЕМУ ИТОГОВОЙ ГОСУДАРСТВЕННОЙ АТТЕСТАЦИИ СТУДЕНТОВ-ПСИХОЛОГОВ

В статье освещаются результаты многолетнего опыта внедрения компьютеризированных тестов на выпускных курсах дневного, вечернего и спецотделения факультета психологии МГУ им. М.В. Ломоносова. Дается описание методики тестового испытания, характеристика созданного банка тестовых заданий, рассматриваются технологические и организационные вопросы (регламент для студентов-экзаменуемых, регламент учета результатов для преподавателей-экзаменаторов). Основное внимание уделяется статистическому анализу полученных взаимозависимостей (статистических корреляций) между результатами тестовых испытаний, оценками на традиционном устном госэкзамене (по билетам) и суммарным баллом диплома, полученным студентом за все годы обучения (сумма оценок за все экзамены). В связи с систематической работой над качеством тестовых материалов корреляция между тестовыми баллами и суммарным баллом диплома (выбранным в качестве главного критерия валидности) год от года росла и достигла значений 0.72 (коэффициент Пирсона) и 0.86 (коэффициент Гилфорда для крайних групп). В то же время аналогичная корреляция для оценок на устном экзамене осталась на прежнем уровне (0.55—0.59 — коэффициент Пирсона в разные годы).

Ключевые слова: образовательное тестирование, компьютеризация, компьютеризированное тестирование, традиционные экзамены, банк тестовых заданий, валидность.

Зинченко Юрий Петрович — докт. психол. наук, профессор, зав. кафедрой методологии психологии, декан ф-та психологии МГУ. *E-mail*: dek@psy.msu.ru

Березанская Наталья Борисовна — канд. психол. наук, доцент кафедры общей психологии ф-та психологии МГУ. *E-mail*: bereznat@mail.ru

Володарская Инна Андреевна — канд. пед. наук, доцент кафедры психологии образования и педагогики ф-та психологии МГУ. *E-mail*: umspsy@mail.ru

Тихомандрицкая Ольга Алексеевна — канд. психол. наук, доцент кафедры социальной психологии ф-та психологии МГУ. *E-mail*: tihomandr@rambler.ru

Шмелев Александр Георгиевич — докт. психол. наук, профессор кафедры психологии труда и инженерной психологии ф-та психологии МГУ. *E-mail*: ags06@ht.ru

The article describes the results of the several years introduction of computerized testing in the outcome examinations for the final year students of psychological faculty of the Lomonosov Moscow State University. The procedure of testing, the features and the structure of the test item bank, technological and organizational issues are discussed including the procedural rules for students and professors. The main focus of the article is statistical analyses of the obtained correlations between test scores, traditional oral exam scores and academic history scores for all years in the university. As a result of the systematic improvements of test item bank the correlation between test scores and academic history scores increased each year and is equal last year 0.72 (the Pearson's linear correlation coefficient) and 0.86 (Gulford's coefficient for extreme groups). At the same time the analogical correlation for oral exams remained at the same level (0.55—0.59 the Pearson's linear correlation coefficient).

Key words: educational testing, computerization, computerized testing, traditional examinations, test item bank, validation.

1. Организационный и научно-методический контекст[1]

В отечественном среднем и высшем образовании в последние два десятилетия наблюдается процесс активного внедрения метода тестов для измерительного контроля качества образования (Аванесов, 1994; Болотов, Шмелев, 2005). Это не в последнюю очередь вызвано тем обстоятельством, что приход в вузы и школы массовой вычислительной техники позволил автоматизировать существенные элементы в процессах подготовки, проведения и обработки именно результатов тестирования. Одной из знаменательных вех в этом процессе служит широко внедренный и широко дискутируемый общероссийский проект ЕГЭ (Болотов и др., 2002; Челышкова, Шмелев, 2004), существенные элементы которого опираются на информационные технологии (хотя сам испытуемый получает, как правило, тестовые материалы в бланковой форме).

Но нужно ли внедрять метод тестов по принципу замещения — замены существующих традиционных форм контроля? Авторы настоящей статьи ставят своей целью обосновать подход к итоговой аттестации, при котором реализуется другой принцип — принцип дополнения. Практически это означает, что тестовые испытания внедряются на факультете психологии МГУ не как единственный и не как важнейший, а лишь как один из вспомогательных этапов государственных экзаменов. На втором (решающем) этапе по-прежнему проводится устный традиционный экзамен. Такая схема явилась для авторов статьи — организаторов этого проекта на факультете — не только и не столько формой политического

[1] Настоящая статья была написана в октябре 2009 г., после завершения этапа освоения и внедрения тестовых компьютеризированных технологий в рамках госэкзаменов на факультете психологии МГУ. О следующем этапе — внедрении письменного экзамена в структуру госэкзаменов — мы будем писать в отдельной статье. (*Примеч. авторов.*)

компромисса между сторонниками и противниками тестовой формы контроля, сколько возможностью проведения полнокровного многолетнего эксперимента, дающего возможность сравнивать и статистически измерять достоинства и недостатки тестовой и традиционной форм не только по отношению друг к другу, но и по отношению к другим внешним критериям.

Еще в 1980-х гг. в рамках лекционно-практического курса «Основы психодиагностики» А.Г. Шмелев начал накапливать опыт применения компьютеризированных тестов как одной из процедур для оценки уровня усвоения студентами материалов учебного курса (Шмелев, 2002). Первые работы по внедрению такого подхода в систему госэкзаменов на факультете психологии МГУ были произведены А.А. Чумаковым (2007). В тестировании тогда принимали участие в основном студенты-добровольцы. Затем с появлением нацпроекта «Инновационное образование» в 2006—2007 гг. на факультете психологии МГУ была выполнена наиболее интенсивная работа по созданию банка тестовых заданий прежде всего по общей психологии (по программе госэкзамена для студентов-психологов), по упорядочению определенных подходов к проведению и обработке результатов тестовых испытаний. Главное заключалось в том, что применение тестовых испытаний стало обязательным для всех студентов, а учет их результатов на устных экзаменах стал осуществляться в соответствии с официальным регламентом и был закреплен в ежегодно обновляемом приказе по факультету. Приобретение тестовыми экзаменами такого официального статуса потребовало тесного взаимодействия всех соавторов настоящей статьи, отвечавших за разные аспекты данного проекта: Ю.П. Зинченко — общее руководство проектом в статусе декана факультета; Н.Б. Березанская — организация работы коллектива авторов тестовых заданий; И.А. Володарская — руководство Учебно-методическим советом факультета, на котором обсуждались и внедрялись принципиальные модели учета результатов тестирования; О.А. Тихомандрицкая — руководство Учебной частью, практическая организация взаимодействия различных групп исполнителей при реализации проекта; А.Г. Шмелев — методическое и технологическое обеспечение основных процедур, комплексная автоматизация процессов подготовки, проведения и обработки результатов тестовых испытаний.

Тем самым подход по принципу дополнения создал условия для реализации научно-практического эксперимента, итоги которого в данной статье авторы впервые попытались систематизировать.

2. Методика

Этап 1. Авторский цикл. Первым и существеннейшим методическим инструментом в рамках данного проекта явилась строго определенная методика и технология работы с авторами банка тестовых заданий. Была применена технология, отработанная А.Г. Шмелевым и сотрудниками

его лаборатории «Гуманитарные технологии» в течение многих лет, начиная с 1997 г., при подготовке компьютеризированной олимпиады «Телетестинг» для старшеклассников и абитуриентов (Шмелев и др., 1998). В ходе авторского цикла данного проекта к работе были привлечены 22 независимых автора — опытные преподаватели и научные сотрудники ф-та психологии МГУ, которые создали в течение двух лет более 700 заданий по основным разделам общей психологии. Опыт разработки заданий по экспериментальной психологии, психодиагностике, клинической, инженерной и экстремальной психологии в данной статье описываться не будет. Данная статья посвящена целиком разработке и проведению тестовых испытаний по общей психологии. В Приложении 1 к данной статье приводится подробный текст инструкции для автора, который позволит читателю не только познакомиться с основным содержательным смыслом авторского цикла в разработке заданий, но и при желании воспроизвести подобный подход.

Этап 2. Экспертный цикл. Существенным вторым этапом работы был так называемый «экспертный цикл». В нем принимали участие в значительной степени те же самые лица, что и в авторском цикле, но в данном случае они выступали экспертами при оценке заданий, созданных коллегами. Важно подчеркнуть, что экспертные работы были еще более формализованными и происходили в режиме диалога с ЭВМ: каждый эксперт решал и оценивал задания, которые предъявлялись ему на экране персонального компьютера (в режиме персонализированного парольного доступа к удаленному серверу с банком заданий)[2]. Задания, при решении которых эксперты не находили правильного ответа согласованным образом (по принципу квалифицированного большинства в 66% ответов), либо корректировались, либо удалялись из банка заданий. В результате экспертного цикла были получены 600 заданий, удовлетворивших большинство экспертов по качеству и дающих согласованные ответы. Это десятикратное превышение численности рабочего банка заданий над числом заданий в индивидуальном варианте(600 против 60) явилось залогом обеспечения достоверности[3] (информационной безопасности) при реализации тестирования.

Этап 3. Проведение тестирования. Каждый год в течение 5 лет (начиная с 2005 г.) каждая экзаменационная сессия преследовала две цели: а) оценить знания студентов-выпускников данного года, б) усовершенствовать банк заданий на основе статистики ответов. Возмож-

---

[2] Данные работы осуществлялись с использованием системы Интернет-сервисов для сбора и анализа экспертных оценок на сайте www.ht-line.ru, разработанных в лаборатории «Гуманитарные технологии» и любезно предоставленных ф-ту психологии МГУ.

[3] Напомним, что достоверностью называется такое свойство теста, которое обеспечивает защиту процедуры от преднамеренных и непреднамеренных искажений. Перемешивание заданий, выбранных случайным образом из банка, обеспечило в нашем проекте защиту от применения грубых шпаргалок в формате «номер вопроса — номер ответа».

ность самосовершенствования методики на основе статистики — это одно из принципиальных достоинств метода тестов, позволяющего формализовать количественно-статистическую обработку результатов и проанализировать качество каждого отдельного тестового задания (вопроса). Поэтому само проведение тестирования в данном случае является одним из ключевых источников информации для повышения качества методики на следующий год.

Тестирование производилось в компьютерных классах факультета психологии в последние годы в течение 1—2 дней в феврале (второй день — резервный) примерно за 10—14 дней до устного этапа госэкзамена. В ходе теста каждому студенту на экране компьютера предъявлялись 60 заданий[4], случайным образом отобранных из банка заданий, причем ровно по 10 заданий на каждый из 6 крупных разделов (см. перечень разделов в Приложении 1). Время на выполнение задания из одного раздела ограничивалось 8 мин., тем самым общее время на тест не могло превышать 48 мин. и в среднем занимало порядка 40 мин. Важно подчеркнуть, что подготовленность студентов к такому испытанию обеспечивала целая серия курсов в предыдущие годы обучения (среди них «Основы психодиагностики», «Экспериментальная психология» и др.), где подобная процедура используется в качестве одного из инструментов текущего и итогового контроля. Также важно, что в течение месяца перед тестовым экзаменом студенты получали неограниченный доступ на интернет-портал факультета www.psy.msu.ru, на котором размещается открытая демоверсия теста.

Этап 4. Обработка результатов тестирования и коррекция заданий. Для облегчения восприятия результатов студентами и преподавателями простейшая обработка сводилась к подсчету числа правильных ответов (сырого балла) по тесту в целом и по тематическим разделам. В течение ряда лет студентам и преподавателям предъявлялись также процентильные баллы (процент испытуемых, показавших балл ниже данного в текущем году), но затем от этой практики решено было отказаться. Напротив, для облегчения сопоставимости в учете результатов тестирования и результатов устного экзамена баллы тестирования переводились в традиционные оценки, исходя из интервалов, согласованных в ходе заседания комиссии из авторов заданий и руководителей проекта. В последние 2 года границы перевода сырых баллов в экзаменационные оценки не изменялись: 44—60 правильных ответов - оценка «отлично»;

---

[4] В данной статье описывается процедура тестирования и схема госэкзаменов, которая уже в 2010—2011 гг. претерпела изменения. Как уже отмечалось выше, к тесту и устному экзамену добавилось третье испытание — письменный экзамен. При этом количество заданий в тесте выросло до 90 заданий, 12 из которых в 2011 г. стали предъявляться на иностранном языке. Баллы по тесту вносят в итоговую оценку вклад в размере 20%. Такая модификация открывает новый цикл практических исследований, заслуживающих отдельного анализа.

34—43 — «хорошо»; 25—33 — «удовлетворительно»; 0—24 правильных ответов — оценка «неудовлетворительно».

Переход к такой фиксированной конверсионной таблице (таблице перевода баллов в оценки) стал возможным благодаря определенной стабилизации процедуры, а главное — повышению качества тестовых заданий.

Для понимания мотивации студентов к выполнению тестов (а это является с психологической точки зрения важнейшим аспектом любой методики тестирования) тут же важно сформулировать, как учитываются результаты тестирования в итоговой оценке за госэкзамен. В последние три года действует мягкий принцип учета, который можно назвать «принципом подстраховки»: низкие баллы по тесту никак не ограничивают студентов в возможности получить в конечном счете любой высокий итоговый балл, а вот высокий балл на тестовом испытании дает определенные гарантии, поскольку устные комиссии, ознакомленные с результатами тестирования, не должны понижать балл более чем на 1 градацию по традиционной 5-балльной шкале. Это практически означает, что «отличники» по тесту не должны получать итоговую «тройку», а «хорошисты» по тесту — итоговую «двойку». Такая мягкая форма учета результатов тестирования (в пользу студентов), утвержденная в приказе, сняла определенное социально-психологическое напряжение, которое неизбежно возникло бы в коллективе (и среди студентов, и среди преподавателей) в случае применения более жестких форм учета.

Для повышения качества тестовых заданий для каждого задания регулярно (каждый год) рассчитываются индексы трудности (процент правильных ответов), дискриминативности (разность в пропорции правильных ответов в высокой и низкой группе испытуемых, сгруппированных по величине балла)[5]. Авторам заданий также передавались данные о проценте выбора дистракторов (ложных ответов). Это позволяло либо удалять низкокачественные задания из банка (с отрицательным или низким КД — коэффициентом дискриминативности), либо корректировать вопрос и ответы с целями:

а) понижения трудности для слишком трудных заданий, б) повышения трудности для слишком легких заданий, в) повышения дискриминативности для таких заданий, по которым либо слабые испытуемые догадывались о правильном ответе, либо сильные запутывались в дистракторах.

Пример тестовых заданий, достаточно легких для того, чтобы проиллюстрировать степень неосведомленности тех студентов, которые отвечают на них неправильно, приводится в Приложении 2.

---

[5] О том, что такое коэффициент дискриминативности и другие характеристики тестовых заданий, в настоящее время на русском языке уже вышла вполне доступная литература (Аванесов, 1996; Клайн, 1994; Майоров, 2001).

Получение эмпирически-статистических индексов трудности для каждого задания позволило применять начиная с 2008 г. такой алгоритм блочной рандомизации[6], который обеспечивает сбалансированный уровень трудности для каждого студента (каждого сеанса тестирования).

В целом примененная технология регулярного повышения качества тестовых материалов — это не какое-то особое изобретение авторов данного проекта, а стандартный подход в рамках известных, давно апробированных и внедренных в странах, развитых в плане тестологической культуры. Так что применение данной технологии может обеспечить любой вузовский коллектив, изучивший классические работы по научной теории тестов.

3. Результаты и их анализ

3.1. Анализ распределения тестовых баллов и оценок

В 2007—2009 гг. процедура тестирования (включая мотивацию студентов) и сам банк заданий оказались практически стабилизированы, так что в дальнейшем имеет смысл принимать во внимание именно результаты за эти три года. За это время в условиях реального экзамена компьютеризированный тест по общей психологии выполнили студенты ф-та психологии МГУ, завершившие обучение по программам дневного и вечернего отделений, — всего 621 человек. Гистограмма распределения оценок изображена на рис. 1. Анализ распределения сырых тестовых баллов (числа правильных ответов) показал, что по критерию Колмогорова—Смирнова гипотеза о нормальности

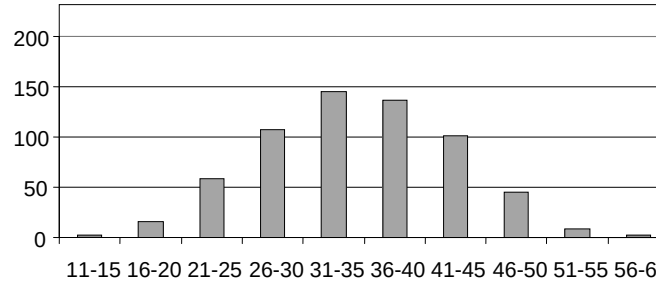

Рис. 1. Гистограмма распределения частот сырых тестовых баллов (числа правильных ответов) по результатам тестирования 621 студента-выпускника (за 2007—2009 гг.). Выбран интервал равнозначности в 5 сырых баллов

[6] Блочная рандомизация — это подход, при котором варианты формируются путем случайной выборки заданий из банка с заданным числом заданий в тематических блоках.

распределения не отвергается. Отсутствуют значимые значения асимметрии и эксцесса. Тем самым можно считать, что мы имеем дело с нормальным распределением с параметрами: среднее = 34.18, стандартное отклонение = 8.11. Так что формальный способ выставления оценок по интервалам, определенным параметрами $\{X_{ср}-S, X_{ср}+S\}$, дал бы в нашем случае границы традиционных оценок, вполне близкие к тем, которые были утверждены комиссией (см. выше).

На рис. 1 наблюдается легкая асимметричность, но если бы мы сдвинули интервалы квантования на 1—2 сырых балла вправо (тем самым сцентрировав шкалу вокруг реальной медианы), то мы получили бы практически симметричную колоколообразную кривую. На рис. 2 можно увидеть, с какой частотой выставлялись традиционные оценки по тесту в течение трех разных лет и каков разброс (колебания) в численности студентов с разными оценками в разные годы.

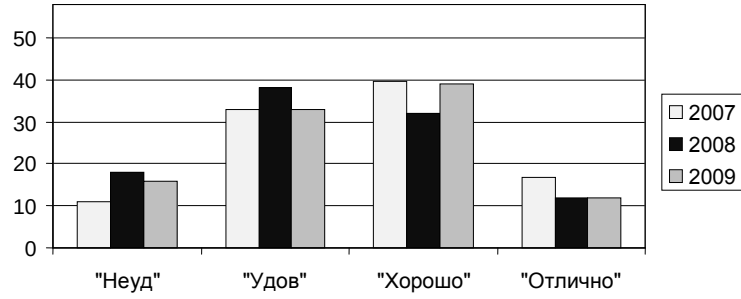

Рис. 2. Пропорции оценок за тестирование в разные годы
(за 100 процентов в каждом году принято общее количество протестированных)

На глаз заметен спад результатов студентов в 2008 г. (по сравнению с 2007 г.), тогда как в 2009 г. наметился определенный подъем. Основным фактором падения результатов в 2008 г. следует считать заблаговременное информирование студентов о том, что в этом году было решено не показывать тестовые оценки членам комиссии на устном экзамене (это было предпринято с целью проверки возможного влияния тестовых оценок на мнение устной комиссии), тогда как в 2009 г. было решено вновь вернуться к ознакомлению комиссии с тестовыми оценками — прямо в процессе приема экзамена.

3.2. Сопоставление тестовых оценок с оценками на устном экзамене

Одним из ключевых направлений анализа результатов явилось сопоставление тестовых оценок и оценок на устном (традиционном) экзамене. На рис. 3 можно видеть гистограмму, аналогичную той, которая приведена на рис. 2, но только применительно к устным оценкам.

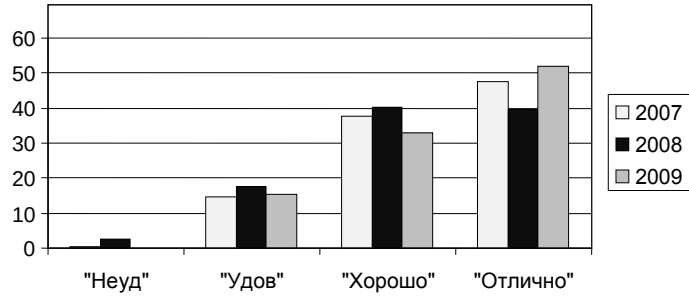

Рис. 3. Пропорции оценок на устном экзамене в разные годы (за 100 процентов в каждом году принято общее количество проэкзаменованных)

Бросается в глаза резкая асимметрия распределения в пользу высоких оценок на устном экзамене, особенно на фоне достаточно симметричного распределения тестовых оценок.

Анализ согласованности тестовых и устных оценок далее производился двумя способами: с помощью расчета традиционных коэффициентов линейной корреляции по нормализованным устным и тестовым оценкам, а также с помощью матриц сопряженности.

На фоне высокой асимметрии устных оценок немудрено, что линейная корреляция нормализованных тестовых баллов и устных оценок оказывается относительно невысокой (хотя и вполне значимой). В 2007 г. она достигала величины 0.51, в 2008 — 0.45, в 2009 г. — 0.47. Это не так плохо с учетом серьезных различий в этих процедурах (и прежде всего с учетом их различий в мотивации в силу их принципиально разного вклада в итоговую оценку, что расхолаживало некоторых студентов при выполнении теста).

Но более интересные результаты дал анализ матриц сопряженности (табл. 1). Только в двух случаях из 621 в течение трех лет экзаменаторы выставляли на устном экзамене оценку, которая оказывалась на 2 градации ниже по традиционной шкале, чем оценка по тесту, т.е. более строгое испытание, каким оказывался тест, достаточно определенно (с точностью до 1 деления на традиционной шкале) информирует о том, что студент вполне подготовлен. Собственно это статистическое наблюдение, выявленное еще в 2005—2006 гг., и стало основой для формулирования того правила, которое действует в настоящее время в виде мягкой поощрительной (подстраховочной) формулы учета результатов тестирования. Только в 28 случаях из 621 экзаменаторы на устном экзамене понижали свою оценку на 1 балл ниже тестовой, т.е. на устном экзамене балл снижался только менее чем у 5% студентов (достигнут стандартный статистический уровень ошибки измерения!). Важно подчеркнуть, что эта закономерность вполне сохранилась и в 2008 г., когда экзаменаторы просто НЕ видели тестовых оценок. Таким

Таблица 1

Три (А, Б, В) частотные матрицы сопряженности тестовых и устных оценок. В каждой клетке каждой матрицы указано количество студентов (частота совместного события), показавших определенное сочетание оценок по тесту и на устном экзамене

А

2009 год	Тест 2	Тест 3	Тест 4	Тест 5
Устный 2	0	0	0	0
Устный 3	6	17	6	0
Устный 4	9	25	28	1
Устный 5	2	19	51	27

Б

2008 год	Тест 2	Тест 3	Тест 4	Тест 5
Устный 2	2	3	1	0
Устный 3	12	27	3	0
Устный 4	12	50	28	6
Устный 5	10	22	43	19

В

2007 год	Тест 2	Тест 3	Тест 4	Тест 5
Устный 2	0	1	0	0
Устный 3	9	13	5	1
Устный 4	11	34	24	3
Устный 5	1	15	47	28

образом, в табл. 1 можно просто увидеть основания для следующего вывода: ***высокий балл по тесту предопределяет высокий балл на устном экзамене с высокой (более 95%!) статистической надежностью.***

Эту эмпирически выявленную закономерность мы предлагаем интерпретировать таким образом: действительно хорошо подготовленные студенты проявляют свою высокую подготовку и в тестовых испытаниях, и на устном экзамене, а средне и слабо подготовленные студенты НЕ могут проявить себя на тестовом испытании, тогда как на устном экзамене им удается добиться высокой оценки.

Низкий балл по тесту не предопределяет постановку на устном экзамене низкой оценки. При этом на 3 балла (с двойки на пятерку) оценка повышалась в 13 случаях из 621 (около 2% от всех оценок). На 2 балла (с двойки на четверку и с тройки на пятерку) оценка повышалась в 78 случаях (почти 13% от всех оценок). Повышение на 1 балл происходит в 277 случаях из 621 (44%). Получается, что завышение на 1 балл встречается даже чаще, чем точное соответствие (35%). На рис. 4 видно, что распределение отклонений устных оценок от тестовых оказывается резко смещенным вправо.

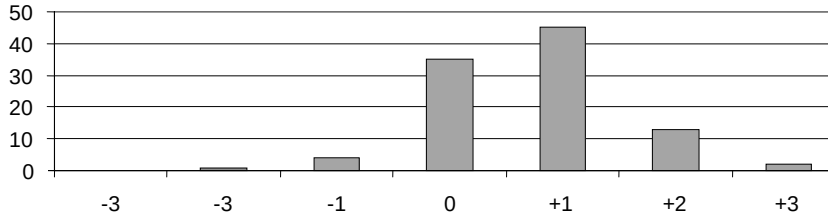

Рис. 4. Гистограмма частот, указывающая на отклонение устных оценок от тестовых (за три года). По оси абсцисс на данном графике отложены величины отклонений устных оценок от тестовых (в единицах традиционной 5-балльной шкалы)

Надо сказать, что постепенно к 2009 г. эта тенденция (можно ее интерпретировать как «тенденцию завышения оценок на устном экзамене») стала нивелироваться. Очевидно, это происходило не без влияния со стороны тестового испытания. Заметно, что число, например, завышений на 3 балла (постановок оценок «отлично» тем, кто получил «неуд» за тест) резко возросло (до 10 человек) именно в 2008 г., когда экзаменаторам вообще не показывали тестовые оценки. Но в целом тенденция сохранилась.

Как же правильнее проинтерпретировать асимметричную связь тестовых баллов и устных оценок? Может быть, за этим стоит вовсе не либерализм преподавателей на устном итоговом экзамене (не «завышение» как таковое), а, например, неумение какой-то части хорошо подготовленных студентов успешно выполнять тестовый экзамен за компьютером? Но... ведь при этом почти ВСЕ студенты, которые успешно выполняют тестовый экзамен, успешно сдают и устный. Так что все-таки основная наша гипотеза, объясняющая асимметричность, заключалась в том, что за асимметрией стоит элементарное стремление преподавателей завышать оценки на устном экзамене (тем более что за пять лет большинству преподавателей студенты стали хорошо известны, многие из них регулярно получали на разных устных экзаменах высокие оценки и тем самым уже как бы «ангажировали» своих экзаменаторов на госэкзаменах на выставление высоких баллов).

Чтобы выяснить, какая же из двух процедур вносит больше искажений (слишком ли жестким и неудобным оказывается тест или слишком мягким и удобным оказывается устный экзамен), надо взять какой-то третий независимый критерий и проверить его статистическую связь с каждой из двух исследуемых переменных.

3.3. Анализ связей тестовых оценок с суммарным баллом диплома

В качестве этого третьего критерия был взят так называемый «суммарный балл диплома» — сумма оценок, полученных каждым студентом на всех экзаменах (их около 40) за все 5 лет обучения на факультете. В

Таблица 2

Показатели линейной корреляции тестовых и устных оценок с критерием «суммарный балл диплома»

Для всей выборки	2005	2007	2008	2009
Тест	0.60	0.7	0.68	0.74
Устный	0.55	0.53	0.52	0.54

Для 16% на краях тест-шкалы	2005	2007	2008	2009
Тест	0.70	0.78	0.84	0.86
Устный	0.55	0.58	0.57	0.59

западной литературе этот критерий известен под названием “academic history score” («балл за всю историю обучения»). Он часто используется в тестологических исследованиях прогностической валидности тестов для абитуриентов и студентов как более объективный инструмент, чем оценка за один единственный экзамен.

В табл. 2 мы видим, что коэффициент корреляции теста с третьим критерием неуклонно рос все эти годы (начиная с 2005 г., который мы тоже привлекли в данном случае для анализа, чтобы показать тенденцию), тогда как тот же коэффициент корреляции между устными оценками и суммарным баллом диплома фактически оказался на одном уровне, немногим более высоком, чем корреляции между тестом и устным экзаменом. Особенно высокий рост корреляций заметен для «краев» тестовой шкалы, т. е. для явных «отличников» (верхние 16% выборки) и для «двоечников» (нижние 16% по тестовому баллу).

Таким образом, работа по совершенствованию тестовых заданий и самой процедуры тестирования дала очевидные плоды: ***валидность именно теста по отношению к внешнему критерию за годы наблюдений возросла, тогда как невысокая валидность устного экзамена осталась на прежнем уровне.***

3.4. Анализ надежности и прогностической устойчивости тестовых результатов

Последний вид анализа, который мы проделали, касается такой ключевой характеристики любой тестометрической процедуры, как надежность, или ретестовая устойчивость. В 2005 г. А.А. Чумаков провел подобный анализ, пригласив сразу после устного экзамена добровольцев — тех студентов, которые откликнулись на просьбу проверить устойчивость теста к перетестированию. В таком перетестировании приняли участие 38 человек. Надо подчеркнуть, что, несмотря на низкую мотивацию студентов (тестирование производилось в тот же день, что устный экзамен, сразу после его сдачи), был получен вполне приличный коэффициент корреляции между первым и повторным тестированием

(r=0.68). Интервал между первым и повторным тестированием в данном случае равнялся 2 неделям — стандарт для такого рода проверочных процедур.

В 2008 и 2009 гг. удалось собрать новые данные такого типа, но на более значительном временно́м интервале между первым и повторным тестированием. В сентябре повторному тестированию по точно такой же программе-методике (при том же банке заданий) подверглись выпускники факультета, поступавшие в аспирантуру (тестирование было включено в программу вступительных экзаменов в аспирантуру). Итого интервал между первым и вторым тестированием составлял уже не 2 недели, а целых 7 месяцев. Подобную проверку можно назвать по-другому. Это уже не столько проверка на надежность, сколько проверка на прогностическую устойчивость тестовых оценок. Ведь за 7 месяцев, да к тому же в такой насыщенный жизненный период, как окончание университета, происходит немало значимых событий (сюда относятся защита дипломной работы, попытки трудоустройства и другие события, возможно, и в личной жизни).

За 2 года были собраны данные по 57 студентам, которые поступали в аспирантуру. Коэффициент линейной корреляции Пирсона для нормализованных тестовых баллов оказался равным 0.72. Такой результат следует признать высоким. Почему? Казалось бы, это весьма умеренное значение на фоне требований к более высокой диахронной устойчивости тестов достижений (устойчивость профессиональных тестов такого рода должна находиться в диапазоне 0.8—0.9). Но следует учесть, что рекомендацию в аспирантуру получают в основном очень сильные студенты. Среди попавших в данную выборку преобладали отличники (уже по результатам первого тестирования). Тем самым речь идет заодно и о различительной прогностичности теста именно в тех диапазонах шкалы, которые прилегают к высокому полюсу. Например, для сравнения аналогичный коэффициент корреляции для устного экзамена (на интервале в 7 месяцев) оказался близким к нулю (всего лишь 0.10, что незначимо отличается от нуля на выборке в 57 человек). Почему мы получили такое превосходство теста в данном случае? Дело в том, что почти все поступающие в аспирантуру получали на устном госэкзамене оценки «отлично», поэтому никакой дифференцирующей прогностической силы эти оценки уже не могли нести. Именно эти причины (низкая дифференцирующая способность устных экзаменов) и подталкивает руководство факультета к внедрению таких процедур, которые гарантируют различение подготовки претендентов в условиях конкурса.

4. Опросы студентов и преподавателей, мониторинг общественного мнения

Любая инновация должна быть так или иначе поддержана большинством сотрудников коллектива. В противном случае коллектив

найдет возможность в явном или неявном виде отказаться от внедрения инновации, т.е. так трансформировать логику проведения и использования формализованных процедур, что они лишатся своего смысла. Поэтому при разработке и внедрении новой технологии оценивания качества подготовки студентов мы постарались по возможности отслеживать динамику отношения к нашей инновации со стороны студентов и преподавателей. Для этого проводились опросы как в режиме онлайн (на портале факультета психологии), так и в ходе очных встреч и собраний (на ряде заседаний Ученого совета факультета, а также, например, на расширенном профессорском собрании в марте 2008 г.).

Из полученной нами статистики ответов на проведенные опросы среди студентов и преподавателей в динамике в течение трех лет следует, что в 2007 г. студенты значительно более скептически относились к тестированию в ходе экзамена, чем преподаватели, которые в целом гораздо более явным большинством поддержали инновацию. Но в 2008—2009 гг. прослеживается положительная динамика в установках студентов. С чем это связано и как это объяснить? Ведь каждый год опрашивались новые выпускники, которым приходилось каждый раз впервые проходить подобную процедуру (на уровне госэкзамена). Одно из объяснений можно свести к достаточно банальному эффекту привыкания. Общественное мнение в студенческой среде тоже инерционно. Доверие к тому, что за низкие баллы по тестам никто не пострадает, возникло не сразу. Но постепенно год от года, по мере того как тест стал приносить скорее позитивные сдвиги в итоговых оценках студентов, отношение стало более позитивным. И что даже важнее: отношение стало меньше зависеть от самооценки успешности собственного выполнения тестового испытания. Например, студенты 2009 г. оценили в целом свою успешность явно ниже, чем студенты 2007 г., но при этом выразили более позитивное отношение к процедуре тестирования.

Многочисленный выбор преподавателями в 2009 г. ответа «не участвовал в приемке экзаменов» вызван был тем, что именно в 2009 г. круг экзаменаторов впервые (за последние десятилетия) был строго ограничен исключительно профессорским составом, а получить от профессорского состава ответы на онлайн-опрос оказалось делом более трудным. Ответы преподавателей в 2008 г. иллюстрируют явный импульс со стороны коллектива в пользу не только формального учета результатов тестирования в итоговой оценке, но и прямого знакомства экзаменаторов с результатами тестирования непосредственно в ходе экзамена. Именно в результате такого мнения коллектива, высказанного на профессорском собрании в марте 2008 г., в 2009 г. было решено вернуться к практике открытого информирования экзаменаторов о результатах тестирования.

5. Общие выводы и перспективы

По характеру распределения тестовых баллов (близкому к нормальному распределению), по результатам проверки надежности и валидности тестовых испытаний можно сделать однозначный вывод, что мы получили достойный инструмент для педагогических измерений качества образования, который и может, и должен быть использован для итоговой аттестации как необходимый этап в системе различных испытаний.

Есть, конечно, «горячие головы», которые предложили таким образом проинтерпретировать высокую корреляцию между тестом и суммарным баллом за все годы обучения (значение, близкое к 0.9): а не является ли это свидетельством того, что тест вообще не нужен, ведь он фактически дублирует ту информацию, которая собрана за все годы. Но именно тестовое испытание способно на самом последнем этапе выявить те важные 10—15% исключений, когда очень слабые ответы студента на самые очевидные вопросы позволяют усомниться в том, насколько корректными были многие оценки, полученные этим студентом в течение 5 лет. Итоговая аттестация для того и предназначена, чтобы не только мотивировать учащихся на сохранение знаний, которые должны быть предъявлены в определенный контрольный момент (а не учиться по принципу «сдал — забыл»), но и осуществлять контроль доброкачественности выполнения своих профессиональных обязанностей всем педагогическим коллективом образовательного учреждения. Кстати, высокая корреляция между результатами теста и суммой баллов за все годы позволяет «страховать» и тестовые технологии от каких-то «чудовищных ошибок» (вызванных, например, неожиданным провалом сильного студента в результате волнения, плохого самочувствие и т.п.) Все случаи резкого расхождения суммарного балла диплома и результатов тестирования должны рассматриваться государственной комиссией индивидуально и быть предметом особого внимания.

Таким образом, наш опыт параллельного применения двух процедур — традиционного устного экзамена и компьютеризированного тестового испытания — по многим параметрам выявил однозначное превосходство тестового испытания. Но на основании этого результата мы не собираемся настаивать на замене устной процедуры. Наоборот, параллельное использование двух процедур дает возможность использовать тестовые данные как основание для совершенствования устных экзаменов с целью преодоления тех недостатков, которые им давно свойственны и в целом хорошо известны (включая весьма либеральное отношение экзаменационных комиссий к тому, что некоторые студенты отвечают на вопросы билетов, не отрывая глаз от своих записей, которые при проверке оказываются удивительно однотипными).

В свою очередь само тестирование вовсе нельзя считать завершенной и безупречной процедурой, не нуждающейся в улучшении

и обогащении новыми элементами и идеями. Для более надежного моделирования профиля успешности по отдельным темам требуется расширение банка заданий (примерно до 1000 заданий) и удлинение сеанса тестирования до 90 заданий (к примеру, в западных сертификационных тестовых экзаменах психологи выполняют до 300 заданий в течение одного сеанса, продолжающего до 4 часов). Также имеется необходимость моделирования так называемых кейс-тестов, предназначенных для диагностики навыков практического применения знаний при решении прикладных задач.

Приложение 1

Инструкция авторам тестовых заданий

«Вам предлагается принять участие в одном из направлений, активно реализуемом в настоящее время на факультете психологии МГУ им. М.В. Ломоносова в рамках инновационного гранта по модернизации образования, — разработка банка тестовых заданий по курсу общей психологии для предэкзаменационного тестирования выпускников. Одним из наиболее существенных требований к инновационным программам является требование к автору (ответственному преподавателю) курса спланировать определенную систему контроля знаний и успеваемости, отвечающую современным требованиям.

При разработке тестовых заданий вам необходимо ориентироваться на Федеральный компонент государственного стандарта высшего образования и программу курса «Общей психологии» для госэкзамена факультета психологии МГУ им. М.В. Ломоносова. При разработке содержания и формы проверочных заданий необходимо учитывать особенности курса психологии и требования к его усвоению, соотнесенные с целями обучения.

При составлении своего блока заданий мы предлагаем вам руководствоваться следующими рекомендациями:

1. Разработка заданий ведется по **6 разделам** курса общей психологии:
- введение в общую психологию и история психологии,
- психология ощущения и восприятия,
- психология внимания и памяти,
- психология мышления и речи,
- психология мотивации и эмоций,
- психология личности и индивидуальных различий.

2. Минимальная численность набора заданий, который должен предоставить один автор, — 24 (по 4 на каждый раздел). В случае трудности охвата каких-то разделов допустимо создание до 8 заданий по одному из разделов (тематических блоков), но крайне желательно охватить не менее 4 разделов.

3. На данном этапе работ требуются задания ТОЛЬКО закрытого типа с выбором одного ответа из четырех возможных.

4. Форма представления заданий. Для удобства и оперативности обработки Ваших материалов рекомендуем использовать текстовые редакторы (например, MS Word или Блокнот). Задания следует оформлять следующим образом: номер вопроса (со значком номера перед цифровой информацией!), на следующей строке текст вопроса, затем на каждой строке варианты ответа, так что каждый

ответ должен обозначаться символом «тире» в начале каждой строки, тогда как правильный — символом * в начале строки. Пример оформления вопроса приводится ниже (номер вопроса п/п, перечень ответов, обозначение правильного ответа):

№ 1
Текст вопроса
— ответ 1
— ответ 2
— ответ 3
* ответ 4

5. Ссылки на литературные источники и авторство. При формулировании вопроса и вариантов ответа не должно возникать неоднозначных ситуаций, порождающих различные трактовки в зависимости от выбранной концепции. Например, должна быть прописана явная отсылка к источнику: не теория внимания вообще, а теория внимания в работах Рибо (всегда как минимум отсылка к автору, а можно и к книге автора). В противном случае почти всегда найдется автор другой «классической книги», который думал по-другому. Так обстоит реально дело в психологии — еще не вполне стандартизованной системе знаний.

6. Типология вопросов. При разработке заданий проще всего использовать классификацию вопросов с помощью естественных их обозначений, основанных на вопросительных местоимениях естественного языка: Что? Кто? Где? Когда? Как? Сколько? Зачем? Почему? Куда? и т.п., включая менее очевидные сложные вопросы: Как называется? В каком порядке? Чем отличается? Как создается?

Таким образом, следует создавать задания, которые сравнительно равномерно будут представлять примерно 12 типов, полученных путем комбинаторики двух оснований классификации заданий:

По трудности:
Трудные — менее 40% правильных ответов
Средние — от 40% до 70%
Легкие — более 70%
По логико-семантическому (методологическому) статусу:
A) Фактологические (Кто? Что? Где? Когда?)
B) Терминологические (Как называется?)
C) Логико-теоретические (Как объяснить? Как предсказать?)
D) Методические (Как воспроизвести, проверить, сделать?)

7. При формулировании вариантов ответа следует избегать возможности применения тактик угадывания правильного ответа, которые являются артефактами самих формулировок ответов. Такие ответы обычно имеют существенные (явные) отличия от остальных вариантов (дистракторов):

- самый длинный ответ;
- содержит «самые умные» слова;
- содержит термины, встречающиеся в самом вопросе;
- нечто среднее между явно абсурдными крайностями;
- имеет подсказку из содержания другого вопроса;
- нечто развернутое на фоне сверхкратких формальных дистракторов.

Одним из вариантов борьбы с тактиками случайного угадывания правильного ответа может послужить создание дистракторов, запутывающих студентов, прибегающих к угадыванию. Такие дистракторы могут:

- содержать ключевые слова (на уровне словесной рифмы с условием вопроса);
- содержать пафосные суждения;
- содержать длинный перечень логических условий, при котором выполняется основное суждение;

8. На последнем этапе работы Ваши задания будут проходить экспертизу со стороны других авторов, которая будет проходить с использованием как минимум 4 вопросов:

— проверка правильности ответа;

— оценка трудности задания;

— оценка качества задания и вариантов ответа;

— тематическая отнесенность к соответствующему разделу курса общей психологи.

Приложение 2

Образец легких тестовых вопросов с «удивительными» ошибками
(звездочкой отмечены проценты выбора правильных ответов)

№	Текст вопроса и вариантов ответа	% студентов, выбравших каждый из вариантов ответа
1	Сознание в интроспективной психологии определялось как:	
	функция деятельности	4
	высшая форма отражения мира	15
	совокупность явлений, данных в переживании только субъекту	83*
	не подлежащая изучению реальность	0
2	К свойствам аффектов НЕ относится:	
	накопление	6
	навязчивость	12
	торможение сознательного контроля	3
	когнитивная сложность	82*
3	Эффект, полученный в экспериментальном исследовании Б.В. Зейгарник, заключается в том, что:	
	люди проявляют более сильную тенденцию к спонтанному воспоминанию о тех действиях, которые им удалось завершить	5
	люди проявляют более сильную тенденцию к спонтанному воспоминанию о тех действиях, которые привели к хорошему результату	5
	люди проявляют более сильную тенденцию к спонтанному воспоминанию о тех действиях, которые им не удалось завершить	86*
	люди имеют тенденцию запоминать все, что они делают	0

СПИСОК ЛИТЕРАТУРЫ

Аванесов В.С. Научные основы тестового контроля знаний. М., 1994.

Аванесов В.С. Композиция тестовых заданий: Учебная книга для преподавателей вузов, учителей школ, аспирантов и студентов педвузов. М., 1996.

Болотов В.А., Шаулин В.Н., Шмелев А.Г. Единый экзамен и качество образования // Единый государственный экзамен: научные основы, методология и практика организации эксперимента / Под ред. В.А. Болотова. М., 2002. С. 9—32.

Болотов В.А., Шмелев А.Г. Развитие инструментальных технологий контроля качества образования: стандарты профессионализма и парадоксы роста // Высшее образование сегодня. 2005. № 4. С. 16—21.

Клайн П. Справочное руководство по конструированию тестов. Киев, 1994.

Майоров А.Н. Теория и практика создания тестов для системы образования. М., 2001.

Челышкова М.Б., Шмелев А.Г. Шкалирование результатов Единого экзамена: проблемы и перспективы // Вопр. образования. 2004. № 0. С. 168—186.

Чумаков А.А. Методика конструирования тестов профессиональных достижений с использованием интернет-технологий: Автореф. дис. ... канд. психол. наук. М., 2007.

Шмелев А.Г. Компьютерное тестирование знаний у студентов-психологов // Вестн. Моск. ун-та. Сер. 14. Психология. 2003. № 1. С. 35—48.

Шмелев А.Г., Ларионов А.Г., Серебряков А.Г. Телетестинг — инновационная аттестационная технология // Тезисы конф. «Развивающаяся психология — основа гуманизации образования» (Москва, 19—21 марта 1998 г.). Т. 2. С. 126—127.

Поступила в редакцию
09.12.10