Проблема эффективности педагогических измерений

Вадим Аванесов

testolog@mail.ru

Опубликовано в ж. «Педагогические Измерения» №4, 2008 г.

Аннотация

Эффективность - четвёртый основной критерий оценки результатов педагогических измерений. Исследования первых трёх критериев - надёжности, валидности и объективности педагогических измерений - были представлены в статьях журнала ПИ[1] и на сайте автора[2].

Эффективность – комплексный критерий оценки полезности тестов[3], тестовых заданий и тестовых результатов, по широкому кругу экономических, социальных, психологических. Эффективность тестов зависит от целей измерения, методологии, теории, методики и от текущей практики педагогических измерений.

В настоящей работе намечаются основные контуры проблемы. Полное и всесторонне исследование эффективности – дело исследователей следующего поколения. В статье дано новое определение педагогических измерений.

Ключевые понятия: педагогические измерения, эффективность педагогических измерений, теорий педагогических измерений, теста, тестовых результатов, тестовых заданий, тестовой формы, дистрактора.

Постановка проблемы

Исследование проблемы эффективности педагогических измерений требует уточнения существующих понятий. И сделать это лучше в начале статьи. Иначе будет непонятно, откуда и зачем возникла проблема эффективности педагогических измерений.

---

[1] Аванесов В.С. Проблема качества педагогических измерений. ПИ № 2», 2004г. С.3-31; Аванесов В.С. Проблема объективности педагогических измерений. ПИ №3, 2008.

[2] http://testolog.narod.ru

[3] В смысле англ. test utility

С начала XX века в научный оборот зарождавшейся тогда классической теории вошли два ключевых критерия качества тестовых результатов - надежность и валидность. В течение примерно ста лет эти критерии назывались иначе и использовались в словосочетаниях «надёжность теста» и «валидность теста». А потому в классической теории[4] особое внимание уделялось разработке методов обоснования именно надёжности и валидности «тестов».

Слово «тесты» здесь взято в кавычки только потому, что в наше время большинство зарубежных авторов сейчас используют уже другую лексику. А именно, вместо надёжности и валидности «тестов» теперь говорят о надёжности и валидности тестовых результатов. Это гораздо правильнее, потому что надёжность и валидность результатов зависит не только от качества тестов, но и от множества других факторов, таких, как адекватность испытуемых уровню трудности заданий, адекватность трудности заданий уровню подготовленности группы. Влияет также длительность тестирования, уровень мотивации испытуемых, возможности списывания и мн. др. И об этом уже приходилось писать[5].

Иное дело - критерий эффективности. Впервые к его исследованию автор данной статьи обратился в своей докторской диссертации. Именно там появился специальный раздел, называвшийся «Эффективность теста и тестовых заданий»[6]. Добавление критерия эффективности теста стало тогда заметным теоретическим продвижением. Были даны определения понятий эффективности теста и тестовых заданий, проведён сравнительный анализ критериев эффективности.

В данной статье эта работа продолжена. Хотя в то время уже возникали некоторые противоречия в понятийном аппарате теории педагогических измерений. Особенно это касалось вопроса определения качества тестовых заданий. Например, естественным образом предполагалось, что «надёжный тест» должен состоять из «надёжных заданий», как и «валидный тест» из «валидных заданий», но так почти никто из числа ведущих западных авторов в последние годы не говорил и не писал. В России эта устаревшая лексика практически неискоренима в короткие сроки, поскольку нет

---

[4] Gulliksen H. Theory of Mental Tests. N - Y. Wiley, 1950 - 486 p.

[5] Аванесов В.С. Проблема качества педагогических измерений. ПИ №2, 2004 г. С. 3–31.

[6] Аванесов В.С. Методологические и теоретические основы тестового педагогического контроля. Дисс…докт. пед. наук. С - Петербургский гос. ун-ет, 1994г.

культурно функционирующего тестового процесса. А потому такого рода лексику приходится слушать и терпеть.

Если надёжность тестовых результатов определяется десятком методов, то определение «надёжности задания» не вполне ясно. Аналогичные трудности вставали при попытках определить понятие «валидность задания». Хотя простые решения были и есть. Например, «валидность задания» очень часто ассоциировалась с мерой корреляции ответов на задание с суммой баллов испытуемых.

Теперь уже не принято писать и говорить о «надёжности и валидности заданий». Эти понятия и критерии обычно относятся только к тестовым результатам. В таком обновлённом истолковании надёжность и валидность продолжают оставаться главными для теории и практики педагогических измерений.

Сейчас понятие "валидность задания" часто заменяется понятиями дифференцирующей или, что лучше, различающей способности задания. Это лучше хотя бы потому, что содержание понятий валидность и надежность гораздо шире; последние относятся ко всему тесту и потому их использование для оценки заданий как объектов частных и элементарных несопоставимо, как несопоставимо единичное и общее в философии.

Однако по мере становления новых теорий педагогических измерений и расширения практики тестирования стала пониматься необходимость увеличения также и числа критериев оценки результатов.

Лексический каркас данной статьи образуют два ключевых понятия теории педагогических измерений: это «*эффективность*» и «*педагогические измерения*». Как уже упоминалось, другие ключевые понятия педагогических измерений – это *надёжность, валидность и объективность*. В философии существует традиция называть ключевые понятия науки категориями. Вокруг них можно и нужно строить весь понятийный аппарат теории педагогических измерений. В мировой тестовой литературе слово «категории» пока не используется.

Зато в практике российской практики тестирования и ЕГЭведения используются ряд надуманных и, нередко, бессмысленных названий, вроде «калибровка тестов», «КИМы», «АПИМЫ», «предтестовые задания» и т.п., претендующие на научность, без элементарной теоретизации.

Именно поэтому они изрядно засорили понятийный аппарат педагогический теории измерений.

Другой вопрос - можно ли считать валидными результаты теста, по которому испытуемые получают много низких или высоких баллов? В классической теории педагогических измерений уже давно сложилась традиция считать результаты тестирования с асимметричным распределением баллов как относительно невалидные по двум возможным причинам: либо трудность большинства заданий не соответствует уровню подготовленности большинства испытуемых, либо наоборот, уровень подготовленности большинства испытуемых не соответствует уровню трудности большинства заданий. И то, и другое снижает качество измерений и практическую полезность. Значит, здесь нужен другой подход и другой критерий.

Интерес к эффективности теста возник и усилился в связи с внедрением в практику математической теории педагогических измерений (Item Response Theory, IRT[7]). Там используется информационная функция, которая вполне может использоваться для оценки эффективности теста и тестовых заданий. Возможность такого истолкования информационной функции была исследована в упоминавшейся диссертации автора этой статьи.

Таким образом, в дополнение к трём уже известным критериям качества педагогических измерений – надежности, валидности и объективности, возникла возможность введения в научный оборот еще одного понятия и, одновременно, критерия эффективности. Потребность в таком критерии стала особенно заметной в системе Rasch Measurement (RM).

[7] См. напр. статьи автора по IRT. Первую – «Item Response Theory: основные понятия и положения»; ПИ № 2, 2007г. Вторая статья – «Истоки и основные понятия математической теории измерений», опубликована в ПИ, №3 2007 г.

Там разработан и математико-статистический аппарат, позволяющий количественно оценить вклад каждого задания в тест как систему измерения.

Проблема эффективности педагогических тестов является частью общей проблемы эффективности форм и методов педагогической деятельности. Естественно поставить вопрос - почему тестирование относятся к эффективной форме организации контроля знаний, а сам тест считается эффективным и объективным методом диагностики уровня и структуры знаний? Краткий ответ на этот вопрос заключается в том, что настоящий тест разрабатывается на научной основе, он технологичен, не только легко поддается автоматизации, но и является, в сущности, основным средством автоматизации контроля. Он экономичен, потому что не требует тех больших затрат живого труда преподавателей, которые сейчас имеют место.

Тест объективен в той мере, в какой удается отграничить процесс тестирования от субъективизма, а порой от произвола и коррупции. Отграничение достигается за счет предоставления одинакового времени, одинаковых условий и правил оценки для всех испытуемых, без исключения. И, наконец, тест рефлексивен в смысле возможностей оценки качества тестовых результатов: без оценки погрешности измерения и адекватности тестовых данных поставленной задачи результаты не признаются как тестовые, заслуживающие доверия.

Практически не исследован в литературе формальный аспект эффективности тестов, если под этим понимать вопрос зависимости эффективности результатов от формы тестовых заданий. Здесь понятие "эффективность" может включать в себя такой понятийный индикатор как "формальная чистота", способствующий лучшему восприятию смысла задания, чёткой оценке и безошибочности учёта тестовых баллов. Нарушение тестовой формы всегда - а это хотелось бы подчеркнуть -

приводит к худшему выражению содержания и к худшему пониманию смысла заданий учащимися и студентами.

Вот почему в заданиях с выбором нескольких правильных ответов из числа предлагаемых на выбор можно говорить о зависимости эффективности задания от правильности формы, от числа ответов к каждому заданию и от соотношения числа правильных ответов к общему числу ответов.

Исследование ключевых вопросов данной статьи полезно начать с определения используемых здесь понятий.

Педагогические измерения

К настоящему времени накопилось много определений педагогических измерений. Часть их уже была приведена в предыдущих работах автора[8]. Самое простое и общее определение было дано в зарубежном учебном пособии по математической психологии[9]. Оно оказалось одинаково применимо для любой общественной науки, куда входят педагогика и педагогические измерения. В том пособии измерение определялось предельно просто, как процесс, посредством которого интересующие латентные свойства личности выражаются числами. В наше время к этому определению полезно добавить, что числа получаются в результате трансформации счётных данных в значения интервальной шкалы. Получится новое определение. *Педагогические измерения - это процесс, посредством которого интересующие латентные свойства личности выражаются числами, которые получаются в результате трансформации счётных данных в значения интервальной шкалы.*

---

[8] Аванесов В.С. Проблема педагогической теории измерений. ПИ №1, 2004. С.15-21.

[9] Cooms, Clyde H., Dawes Robert M., Tversky Amos. Mathematical Psychology. An Elementary Introduction. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1970. p.7

Например, знания, интеллект и уровень подготовленности испытуемых являются латентными свойствами личности. Слово «латентные» означают, что эти свойства внешне невидимы, но их можно оценить по проявлениям, фиксируемыми посредством индикаторов. В качестве индикаторов лучше всего использовать качественные тестовые задания.

К числу концептуально насыщенных можно отнести определение, данное B.D.Wright & J.M.Linacre[10]. Измерение, в их интерпретации, рассматривается как следствие специально организованного процесса сравнения результатов испытуемых на шкале натуральных логарифмов. В этом определении главное – процесс трансформации исходных тестовых баллов испытуемых в значения шкалы натуральных логарифмов, после чего, собственно, и появляется педагогическое измерение уровня подготовленности. До процесса логарифмического преобразования этими авторами исходные баллы не рассматриваются как педагогические измерения. Симметрично проводится шкалирование мер трудности заданий на той же шкале. На английском языке этот процесс называется Item calibration. На русском языке оба эти процесса автор данной статьи определяет как *измерение уровня подготовленности испытуемых* и *шкалирование заданий по уровню их трудности*.

Основные требования к педагогическим измерениям1. Свойство линейности шкалы измерения, что допускает удобства в применении математических аксиом и операций.

2. Параметры заданий и испытуемых не должны быть взаимно зависимы. Это главное научное достижение G.Rasch. Вся технология RM вытекает из свойства независимости параметров испытуемых от параметров заданий, и наоборот.

---

[10] Wright B.D., Linacre J.M. Rasch model derived from objectivity. Rasch Measurement Transactions 1:1 p.5. 1987. *A measurement is the quantification of a specifically defined comparison.* Фамилия второго автора читается Линека, с ударением на первом слог.

3. Метод измерения должен быть сравнительно легким, компьютеризованным, полностью, по возможности, технологичным. Это требование позволяет привлечь к проведению измерений большое число профессорско-преподавательского состава.

4. Одномерность измеряемого свойства для начинающих исследователей и лиц, упомянутых в п.3. Продвинутые исследователи, имеющие подходящее математическое и статистическое образование, тяготеют к многомерным моделям измерения, которые всегда интереснее одномерных моделей.

5. Монотонность отображения свойства в числовую шкалу. Смысл этого требования прост: испытуемые, имеющие более высокий уровень подготовленности, должны получать и более высокий балл в RM.

Измерения по теории Rasch отвечают всем этим требованиям.

Для узко понимаемой педагогики педагогические измерения не были нужны. Не случайно в соответствующих учебниках они много лет даже не упоминались. И только недавно в России стала понемногу пониматься идея, что эффективная образовательная деятельность без качественных и эффективных педагогических измерений невозможна.

Основным методом педагогических измерений является тест.

Педагогический тест в новой формулировке определяется так: это система вариативных заданий, равномерно возрастающей трудности, позволяющая качественно оценить структуру и эффективно измерить уровень подготовленности испытуемых по одной или нескольким учебным дисциплинам. Смысл словосочетания «система вариативных заданий» означает, что каждое задание теста имеет свои параллельные варианты. Смыслы всех остальных терминов этого определения читатель найдёт в работе автора этой статьи[11].

---

[11] Подробнее см. Аванесов В.С, Основы теории педагогических измерений. ПИ №1, 2004г. С.15-21.

B.D.Wright и M. Stone, вслед за Л.Л. Гуттманом,обратили внимание на важный системный фактор распределения заданий теста по уровню трудности. Трудность рядом стоящих заданий теста не должна отличаться более чем на 0,5 логита[12]. Иначе на шкале образуются провалы. Расстояние в 0,5 логита – это довольно либеральное требование. Лучше, когда расстояние между заданиями бывает не более чем 0,25 логита трудности. Это требование можно назвать условием достаточной плотности расположения числа заданий на шкале.

Например, представленные на рис.1, три задания располагаются на шкале трудности с разрывом между собой в один логит трудности. Поскольку это расстояние больше, чем упомянутые выше 0,5 и 0,25 логита, то, очевидно, эта часть теста имеет дефект, связанный с подбором заданий по уровню трудности. Условие достаточной плотности заданий на шкале здесь не выполняется, а потому результаты всего теста будут неточными в данной окрестности, также как невалидными и неэффективными в целом.

На рисунке 1 по оси абсцисс представлена мера трудности заданий, по оси ординат – вероятность правильного ответа на задания, в зависимости от уровня подготовленности испытуемых. Ось абсцисс представляет единую (общую) логарифмическую шкалу уровня подготовленности испытуемых и уровня трудности заданий[13].

---

[12] Исходное значение логита трудности задания находится из выражения $\ln q_j / p_j$, где q_j является долей неправильных ответов испытуемых на задании теста под номером j, а p_j - это доля правильных ответов испытуемых на то же самое задание под номером j.

[13] На рис.1 мера трудности заданий изображена символом b вместо принятого в нашем журнале символа β. Так бывает часто: b здесь представляет выборочные значения мер трудности заданий, а β – значения параметров заданий в генеральной совокупности.

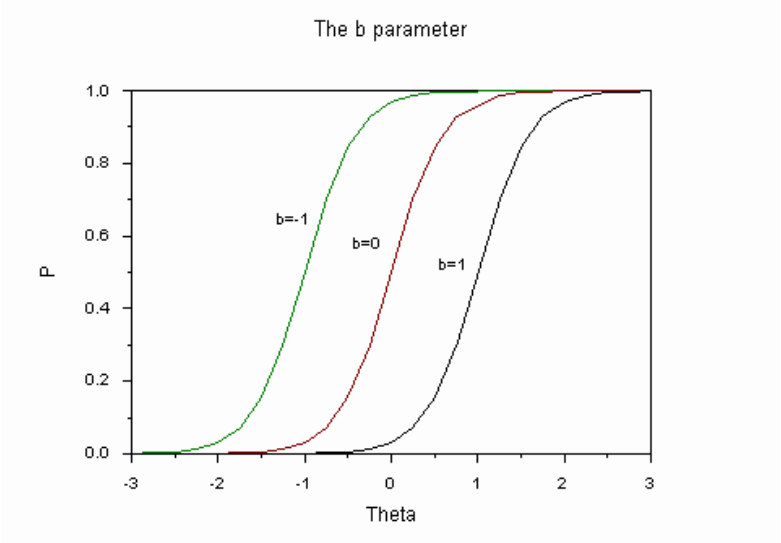

Рис.1. Пример заданий, расположенных реже допустимого уровня

Эффективность теста

Понятие «эффективность» имеет две основы. Первую основу образует смысл слова «эффект». Эффект – это результат какого-либо действия или деятельности. Он может быть как положительным, так и отрицательным. Английский аналог – слово effect, что можно перевести как воздействие, влияние. Можно исследовать, например, положительный или отрицательный эффект от удаления части заданий в разрабатываемом тесте. Если удаление некоторых заданий не снижает надёжности и валидности тестовых результатов, то эффект, очевидно положительный, в смысле уменьшения числа заданий теста без потери его полезных свойств. Вторую основу понятия «эффективность» образует идея деятельности, нацеленной на получение искомого эффекта с наименьшими затратами.

Эффективность – относительный показатель результативности (эффекта). Эффективность определяется как понятие, производное от полученного результата, делённого на расходы, время и др.

Идея эффективности интересна для оценки содержания теста и тестовых заданий. Можно создать сколько угодных вариантов одного и того

же теста, имеющих сходное или иное содержание. Если сходно, то мы можем думать о параллельном варианте теста. Если иное содержание теста, нацеленного на измерение одного и того же свойства, то проблема переносится в плоскость другого важного критерия – валидности результатов, получаемых тем или иным тестом, измеряющим общее свойство.

Вопрос эффективности теста ранее уже получил некоторое освещение в работе автора. Эффективным там назывался тест, если он лучше, чем другие тесты, измеряет знания студентов интересующего уровня подготовленности, с меньшим числом заданий, качественнее, быстрее, дешевле, и все это - по возможности, в комплексе. Если из какого-либо теста с большим числом заданий сделать оптимальный выбор меньшего числа, то может образоваться система, не уступающая заметно по своим свойствам тесту со сравнительно большим числом заданий. Тест с меньшим числом заданий в таком случае можно называть сравнительно более эффективным[14].

Два ключевых фактора тестирования – это число заданий теста и уровень подготовленности студентов.

Эффективность теста можно попытаться оценить с точки зрения его дифференцирующей способности; последняя тем выше, чем лучше видны различия между тестовыми баллами студентов. В качестве одного из возможных показателей дифференцирующей способности теста обычно используется дисперсия исходных тестовых баллов.

Если, например, имеются два теста по одной и той же учебной дисциплине и один из них имеет большую дисперсию, чем второй (в той же самой группе), то при прочих равных условиях тест с большей дисперсией можно считать эффективней, чем тест с меньшей дисперсией. Отношение большей дисперсии к меньшей при одинаковом, например, числе

[14] Аванесов В.С. Проблема качества педагогических измерений. ПИ, №2, 2004г. С.3-27.

заданий, с последующим умножением на сто, может служить в качестве одного из показателей сравнительной эффективности теста с позиции дифференцирующей (различающей) способности.

Другой важное требование к качественным тестовым результатам является близость логарифмических оценок средних арифметических баллов множества испытуемых и среднего арифметического балла меры трудности множества заданий. При условии, что обе упомянутые средние арифметические выражаются в одной и той же шкале натуральных логарифмов. Средняя трудность заданий не должна отличаться от среднего уровня подготовленности испытуемых более чем на 0,5 логита.

Пример неэффективного тестирования, в котором уровень трудности теста оказался заметно ниже уровня подготовленности испытуемых представлен на рис. 2. Графическую информацию о неэффективности проводимого измерения дают две совмещённые на одной оси гистограммы распределения результатов испытуемых (сверху) и мер трудности заданий (снизу). Рис. 2. получен посредством применения математико-статистического пакета RUMM – 2020 в процессе его апробации.

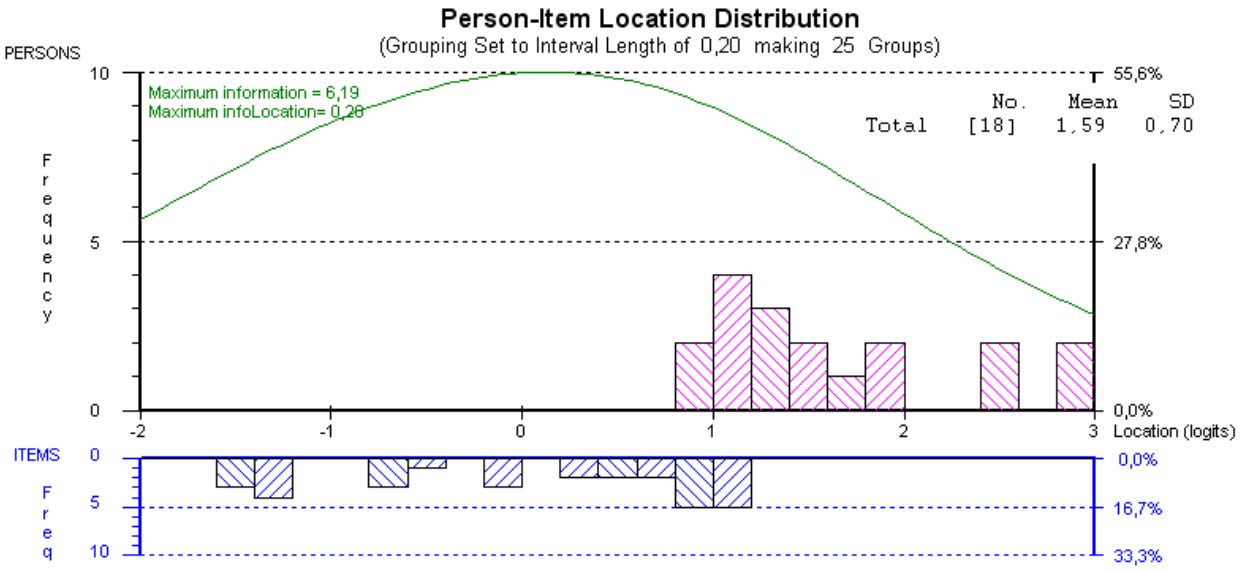

Рис. 2. Убедительный пример несоответствия уровня подготовленности испытуемых уровню трудности заданий.

Можно также проверить влияние формы и содержания заданий, а также ответов на качество теста. Добавление в задание с выбором одного или нескольких правильных ответов дополнительного числа дистракторов (ответов правдоподобных, но неправильных), имеет своим эффектом уменьшение вероятности угадывания правильных ответов. Аналогично можно думать о влиянии (об эффекте) увеличения числа заданий в тесте на показатель надёжности тестовых результатов. Классический пример такого рода дают широко известные формулы Spearman-Braun.

Можно говорить и о влиянии добавления в тест заданий определённого уровня трудности на распределении исходных результатов испытуемых. Если добавляется примерно десяток трудных заданий, то можно видеть относительное снижение баллов у большого числа испытуемых. Если в проектируемый тест добавить примерно такое же количество сравнительно лёгких заданий, то эффектом становится относительное повышение исходных тестовых баллов у большинства испытуемых. Здесь вполне можно видеть изменения в распределении баллов в зависимости от состава заданий теста[15].

В диссертационной работе автора было дано следующее определение эффективности теста: тест называется эффективным для измерения знаний студентов с уровнем, соответствующим точке оси θ_i, если он обеспечивает в этой точке максимум информации о значении уровня подготовленности при минимуме числа заданий. Эффективность измерений достигается за счет дифференцированного подбора заданий требуемого уровня трудности для каждого студента, имеющего уровень знаний θ_i.

---

[15] См примеры асимметричных распределений в ЕГЭ 2016 и 2017 г. по русскому языку.

Приводятся в Докладе на Международной научно-практической конференции «Обучение русскому языку как иностранному в специальных целях: теория и практика». на факультете русского языка как иностранного российского университета Дружбы народов, 25 октября 2018г. Представлен на сайте вуза и на сайте viperson.ru

http://viperson.ru/articles/optimalnoe-vremya-pedagogicheskogo-testirovaniya

Графические образы эффективных и неэффективных тестовых заданий

Решение задачи поиска самых эффективных заданий теста нередко бывает полезно бывает начинать, наоборот, с попытки определения самых неэффективных заданий. Что позволяет уменьшить размер обрабатываемых матриц, улучшить интерпретируемость результатов и быстрее, т.е. эффективнее, решить эту задачу. Хороший материал для выбраковки неэффективных заданий дают методы множественного корреляционного, регрессионного и факторного анализа. Первый и второй позволяют оценить так называемый в статистике чистый вклад каждого задания в общую вариацию тестовых баллов, в то время как различные варианты факторного анализа являются классическими методами проверки гомогенности тестов. Применение всех методов в разработке и обосновании качества тестов представляет предмет отдельного учебного курса.

Трудно найти задания с одинаковой эффективностью получаемых оценок. И уже совсем редко находятся задания, которые имеют минимум погрешностей оценивания. На рис. 3 представлен графический образ идеального тестового задания, имеющего высокий уровень дифференцирующей способности, на очень узком интервале измеряемого свойства.

Из графика этого задания видно, что в группах испытуемых, набравших, по всему тесту, меньше семи баллов ни один испытуемый не смог правильно его выполнить. Соответственно, доля правильных ответов по данному заданию, для этих балльных групп, равна нулю. Зато те испытуемые, кто набрал больше семи баллов все, до единого человека, дают правильный ответ на данное задание. Это и есть видимый (в точном смысле этого слова, как на рентгеновском аппарате), признак высокой дифференцирующей способности данного, действительно образцового тестового задания.

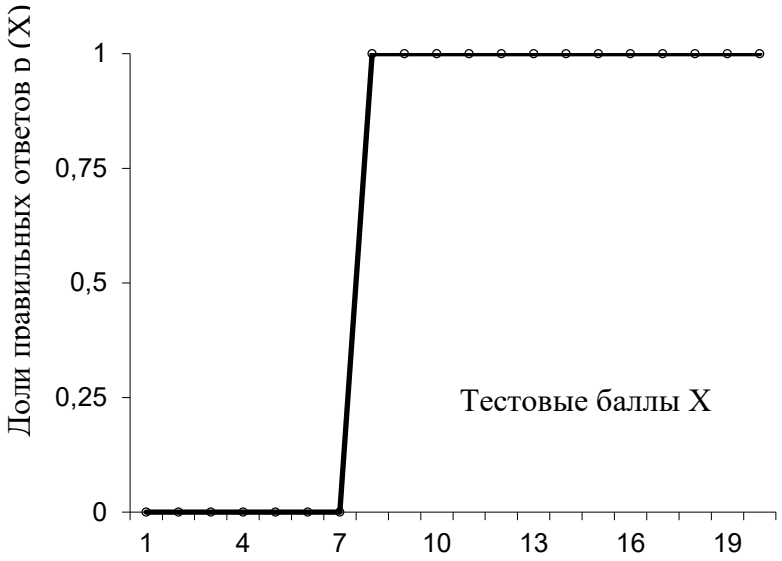

Рис. 3 Графический образ задания, отлично дифференцирующего испытуемых на две группы: от одного до семи и от восьми и более.

Если бы на каждом интервале измеряемого свойства – от 0 до 1, от 1 до 2, от 2 до 3 и т.д. – были бы только такие задания, то тест, состоящий из них, можно было бы назвать идеальным, настолько, что в практическое существование таких заданий верится с трудом. Остаётся только добавить, что в литературе тест, состоящий из таких совершенных заданий, называют шкалой L.L.Guttman. Парадоксально, но в системе Rasch Measurement такого рода идеальные задания в некоторых случаях таковыми не считаются[16].

На рис. 4 представлен графический образ неэффективного задания. График этого задания имеет малую крутизну, что означает его довольно низкую дифференцирующую способность. Ранее уже отмечалось - чем выше крутизна графика, тем лучше задание оценивает на данном интервале измерения. Но в случае с заданием на рис.4 наблюдается противоположная картина; интервал измерения для него – вся шкала, от нуля до 20

[16] Masters, G. (1988) "Item discrimination: when more is worse", Journal of Educational Measurement, 25:1, 15-29, and www.rasch.org/rmt/rmt72f.htm - RMT 7:2, 289. А так же: My best items don't fit! Rasch Measurement Transactions, 2004, 18:3 p. 992. http://www.rasch.org/rmt/rmt183n.htm

баллов. На каждом балльном уровне оно «работает» с дефектом, плохо различия знающих испытуемых от незнающих испытуемых.

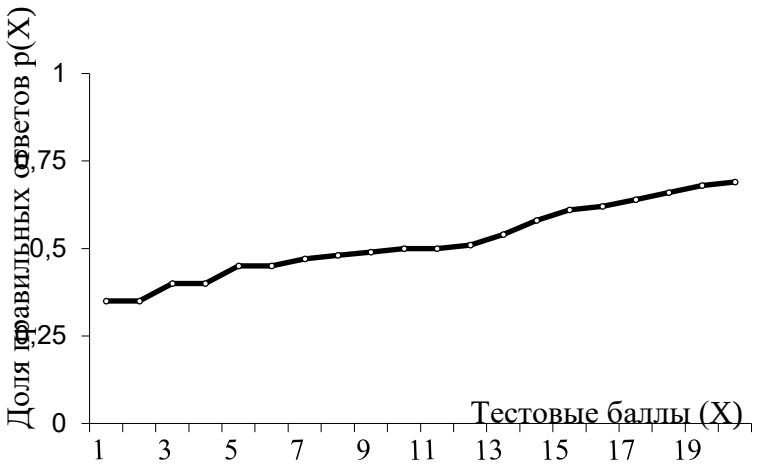

Рис.4. Графический образ задания, плохо дифференцирующего испытуемых любого уровня подготовленности.

Можно задать уточняющий вопрос: а почему задание на рис. 4 отнесено к числу неэффективных? Потому что, во-первых, оно сравнительно лёгкое для самых незнающих; 30% слабо подготовленных испытуемых справляются с ним. Во-вторых, оно оказывается довольно трудным для части хорошо подготовленных испытуемых. Это значит, что оно имеет дефект формулировки, такой, что даже отлично подготовленные испытуемые не понимают содержание задания.

Напомним, что на оси X отложены значения тестовых баллов испытуемых, а на оси ординат – доли правильных ответов (p), полученных в каждой балльной группе испытуемых. Произведение (p • 100) и даёт отмеченный процент.

Столь противоречивая сущность данного задания выражается и на графике. Там обращает на себя внимание слабый - а можно сказать и чуть эмоциональнее - вялый прирост доли (или процента) правильных ответов, в зависимости от уровня подготовленности испытуемых. Различающая способность оказалась низкой на всех значениях континуума измерения.

И даже в группе самых подготовленных испытуемых доля правильных ответов не превышает 65 процентов.

Педагогическая интерпретация подобных заданий примерно такова. Скорее всего, это задание с двумя или тремя ответами, с уровнем угадывания правильного ответа не менее 30%. На нём ошибаются и слабые, и хорошо подготовленные испытуемые. Задание требует переработки в направлении достижения большей ясности его смысла испытуемым всех уровней подготовленности. Тогда его станут лучше понимать и соответственно, правильнее на него отвечать. В первую очередь те, кто лучше подготовлен.

Задания нужно формулировать так, чтобы их смысл был понятен всем испытуемым. Знают ответ не многие, но понимать смысл задания должны все, или почти все. Это случается, если разработчики сумеют доступно и четко изложить содержание задания. Давно сказано - кто ясно мыслит, тот ясно излагает.

Тестологическая интерпретация такого случая довольна проста. Ответы на задание слабо коррелируют с суммой баллов. Именно об этом свидетельствует низкая крутизна графика. В нем проявляет себя высокая вероятность угадывания, а, следовательно, высока и погрешность измерения. Вот почему такому заданию в тесте места нет. Хотя задание может быть в тестовой форме, оно не тестовое. Отсюда легко понять плодотворность лексики, разработанной к книге автора этой статьи[17]

В общем случае задания с большей крутизной графика оказываются и более эффективными для разделения испытуемых на группы подготовленных и неподготовленных. Из трёх заданий одинакового уровня трудности, но разного уровня крутизны рис. 5 более эффективным для изме-

[17] Аванесов В.С. Форма тестовых заданий. М.: Центр тестирования, 2006 г. См. в бесплатном доступе по адресу: http://viperson.ru/articles/forma-testovyh-zadaniy

рения оказывается задание №1. Его и надо включать в тест. При этом задания под №№ 2 и 3 в тест не попадут, потому что в одном тесте иметь задания одинаковой трудности нет необходимости.

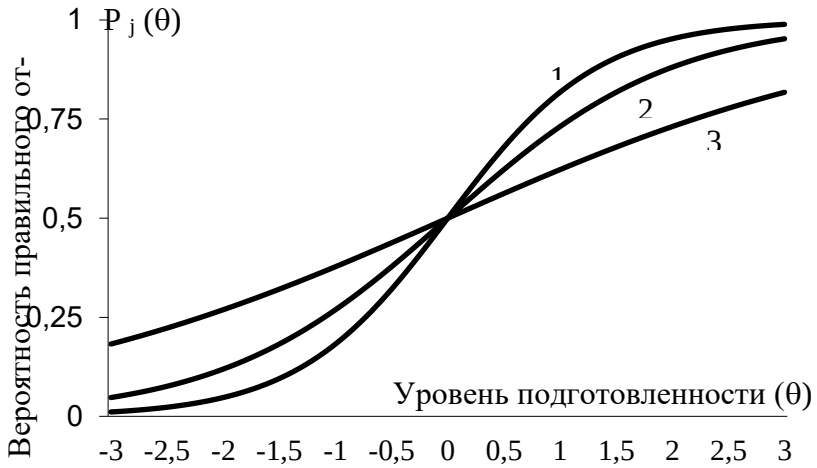

Рис.5. Графики заданий со значениями a_1 =1,5; a_2 =1,0; a_3=0,5.

Однако в системе Rasch Measurement используется иная логика. Там в тест предпочтительно включаются задания, близкие к одному общему среднему уровню крутизны графиков.

В общем случае у более эффективного задания короче диапазон действия и больше крутизна графика. Это явление можно видеть на рис.6, где представлены сравнительные графики двух заданий различной эффективности.

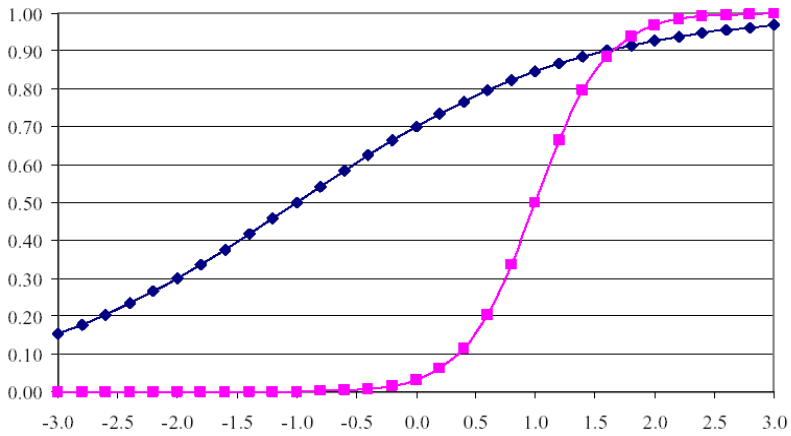

Рис.6. Два задания с разной различающей способностью.

Авторы книги Best Test Design[18] ввели ряд относительно новых понятий, так или иначе связанных с валидностью и эффективностью результатов. Например, какие результаты более валидны с точки зрения измерения – баллы по тесту, имеющему задания в широком диапазоне трудности (wide test), или баллы по тесту, имеющем задания на узком диапазоне измеряемой величины (narrow test)? Эти же авторы ввели ещё одно интересное понятие, которое по-английски пишется как operating level of the test, что можно перевести как меру эффективности теста для измерения уровня подготовленности испытуемых определённого диапазона уровня подготовленности испытуемых.

Понятие «operating level of the test» ассоциировано с понятием дифференциальной эффективности теста, в соответствии с которым тест эффективен только в той точке континуума знаний, для которой более всего он подходит по уровню трудности содержащихся в нём заданий.

Тест не может быть эффективным вообще, на всем диапазоне подготовленности студентов. Он может быть более эффективен на одном уровне знаний и менее - на другом. Именно этот смысл вкладывается в понятие дифференциальной эффективности теста в процессе раскрытия идеи критерия эффективности.

Три правила отбора заданий

Можно сформулировать три правила отбора заданий для создания эффективного педагогического теста:

1. В тесте нужны задания, измеряющие преимущественно интересующее содержание учебной дисциплины. Это правило основывается на принципе гомогенности содержания теста. Практически трудно найти за-

[18] Wright B.D., Stone M.H. Best Test Design. Chicago. MESA Press, 1979. - P.7.

дание, которое измеряет только интересующее свойство, и ничего другого. Посредством факторного анализа удаётся определить меру адекватности каждого задания измеряемому свойству личности.

2. В тесте нужны задания более или менее равномерно возрастающей трудности. Это вытекает из данного выше определения теста и из принципа соответствия уровня трудности заданий уровню подготовленности испытуемых.

3. Задания, имеющие сходные параметры, избыточны для эффективного теста.

Сравнение критериев эффективности, надёжности и валидности результатов измерений

Уже упоминалось в этой и других работах автора, что сейчас уже не говорят «надёжный тест» или «надёжное задание». «валидный тест» или «валидное задание». Вместо этого зарубежные классики советуют говорить о надежных и валидных результатах педагогических измерений. И объясняют, что один и тот же тест, будучи применённым в разных ситуациях измерения и в группах с разной подготовкой испытуемых может давать весьма различающиеся показатели качества измерения, в смысле точности и адекватности результатов поставленным целям измерения[19]. В такой постановке вопроса смысл общего понятия надёжный или валидный «тест» действительно размывается. Возникает фактор ситуативности, зависимости качества измерений от ситуации измерения.

Постепенно стала осознаваться необходимость рассмотрения не только эффективности теста и тестовых заданий, но также эффективность теории измерений, интерпретаций и принимаемых решений в процессе

[19] Thompson, B. & Vacha-Haase, T. (2000). Psychometrics *is* datametrics: The test is not reliable. *Educational and Psychological Measurement, 60,* 174-195.

тестирования, эффективность формы, содержания, методов, технологий и даже дистракторов к заданиям. И всего остального, что не охватывается тремя уже известными критериями.

Расширилась и сфера применения критериев. В процессе тестирования приходится принимать во внимание затраты времени и средств на измерение интересующего свойства испытуемых, искать возможности применения кратковременных тестов вместо длительных методов педагогического оценивания или трёх-четырёхчасового некачественного государственного экзамена. Фактор времени – один из главных признаков различия между тестом и экзаменом[20].

Качественно разработанный тест всегда короче по времени, точнее, адекватнее для совокупности испытуемых, технологичнее, имеет меньшую погрешность измерения, объективнее, справедливее и экономнее, чем любая другая форма проверки знаний. Отсюда видно, что к числу других понятий, сопряжённых с эффективностью, можно назвать надёжность и валидность педагогических измерений.

Эффективность теорий педагогических измерений

В последние годы не было недостатка в утверждениях о преимуществах IRT по сравнению с классической теорией. И действительно, в некоторых отношениях IRT имеет больше возможностей, чем классическая теория. Но означает ли это, что IRT лучше, а классическая хуже? Сравнительные исследования этих двух теорий[21] убеждают, что на самом деле обе эти теории сейчас используются, они работоспособны, обе техноло-

[20] Аванесов В.С. Оптимальное время педагогического тестирования. http://viperson.ru/articles/optimalnoe-vremya-pedagogicheskogo-testirovaniya

[21] Courville T.G. An empirical comparison of item response theory and classical test theory item/person statistics. A Dissertation Submitted to Texas A&M, August 2004.

гичны, взаимно дополняемы. Каждая имеет свои преимущества и недостатки. А потому нет смысла в их противопоставлении. Однако есть смысл в интеграции возможностей каждой теории при разработке теста.

Классическая теория эффективна на начальном этапе выбраковки некачественных заданий по критериям трудности, вариации и корреляции. Здесь легко, быстро и технологично выполняются расчёты, легко понимаются решающие правила выбраковки неэффективных заданий, легко определяется коэффициент надёжности тестовых результатов типа коэффициента альфа L.J.Cronbach. Например, с одного взгляда на табл. 1 становится понятно, что задания №1 и №10 имеют очень мало шансов попасть в тест в силу низкого значения коэффициента корреляции ответов испытуемых на задания (X_j), где индекс j – номер задания, с суммой баллов, полученных всеми испытуемыми. А задание № 20 является безнадёжным, требующим переработки, или удаления из проектируемого теста. Предел допустимости заданий в педагогический тест обычно принимается $r < 0,30$.Кроме того, во внимание приходится принимать значения меры трудности и содержание задания.

Таблица 1. коэффициентов корреляции между заданиями проектируемого теста (Var j[22]) и суммой баллов испытуемых.

Var1	0,19	**Var15**	0,27
Var2	0,27	**Var16**	0,28
Var3	0,61	**Var17**	0,34
Var4	0,62	**Var18**	0,43
Var5	0,63	**Var19**	0,53
Var6	0,50	**Var20**	0,11
Var7	0,56	**Var21**	0,56
Var8	0,45	**Var22**	0,25
Var9	0,42	**Var23**	0,24
Var10	0,22	**Var24**	0,56
Var11	0,35	**Var25**	0,29

[22] От англ. variable – переменная величина.

Var12	0,40	**Var26**	0,50
Var13	0,59	**Var27**	0,46
Var14	0,47	**Var28**	0,50

Эффективность тестовой формы

Автор этой статьи считает форму заданий не менее важной, чем содержание. Традиционно при тестировании чаще других используются задания с выбором одного правильного ответа из четырёх-пяти ответов, представляемых на выбор. При этом, однако, высока вероятность угадывания правильного ответа теми испытуемыми, которые не знают его. Таким образом, не менее пятой части всех получаемых правильных ответов априорно считается ошибочной, недостоверной информацией.

Эффективность заданий с выбором одного правильного ответа снижается из-за тенденции ставить правильные ответы не в начале и не в конце, а посредине. Таким образом, разработчики как бы прячут правильные ответы в середину. Если это имеет место, что видно по примеру первых заданий, опытные испытуемые, при прочих равных условиях, при угадывании ответов на трудные задания избегают выбор первого и последнего ответа. Тем самым повышают исходный тестовый балл[23].

Для преодоления этого недостатка делаются попытки применить задания открытой формы, и оценивать ответы по данной форме удвоением баллов. Но тогда возникает потребность в огромном количестве дорогостоящих сканеров, в ручной коррекции нечётко считанных символов операторами, резко снижается понимаемость заданий испытуемыми, повышается число возможных правильных ответов. Эффективность измерений резко снижается.

[23] Yigal Attali & Maya Bar-Hillel. Guess Where: The Position of Correct Answers in Multiple-Choice Test Items as a Psychometric Variable (June 2001) Journal of Educational Measurement, 40 (2003), 109-128

В общем, задания открытой формы однозначно проигрывают по критерию технологичности заданиям с выбором одного правильного ответа. В качестве выхода из этой ситуации автор данной работы уже много лет внедряет в практику задания с выбором нескольких правильных ответом. В таких заданиях полностью сохраняется технологический потенциал, резко снижается вероятность угадывания всех правильных ответов, повышается дисперсия результатов испытуемых и появляются другие преимущества, необходимые для создания качественных тестов.

Инструкция для испытуемых может иметь такое содержание. Вашему вниманию предлагаются задания, в которых могут быть один, два, три и большее число правильных ответов. Нажимайте на клавиши с номерами всех правильных ответов:

1. КОНСТИТУЦИЯ ХАРАКТЕРИЗУЕТ РОССИЮ КАК ГОСУДАРСТВО

1) светское	8) либеральное
2) унитарное	9) парламентское
3) социальное	10) олигархическое
4) федеративное	11) демократическое
5) общенародное	12) социалистическое
6) республиканское	13) капиталистическое
7) конфедеративное	14) народно-демократическое

Задания с выбором нескольких правильных ответов усилиями автор внедрены в медицинском образовании. Пример:

2/ АНТИБАКТЕРИАЛЬНЫЕ ПРЕПАРАТЫ

1) делагил	6) ацикловир
2) амиксин	7) цефтриаксон
3) амикацин	8) пенициллин
4) реаферон	9) левомицетин
5) цефазолин	10) сульфасалазин
6) нистатин	12. амоксициллин

Применение IRT для оценки эффективности дистракторов в заданиях в тестовой форме

Для решения прикладных образовательных задач в данной статье рассматривается вариант использования IRT для проведения анализа эффективности не только тестов и заданий, но и отдельных дистракторов заданий в тестовой форме. Напомним, что дистракторами называют неправильные, но правдоподобные ответы в заданиях с выбором одного или нескольких правильных ответов. Это название происходит от английского глагола to distract, что можно перевести словом «отвлекать», имея в виду, что дистрактор предназначен для испытуемых, не знающих правильные ответы на задания, а потому ищет правдоподобные.

Необходимость проведения дистракторного анализа вытекает из логики организации эффективного процесса педагогического тестирования. Не бывает качественных измерений без проведения дистракторного анализа и без публикации информации о качестве используемых в тесте заданий и их дистракторов.

С точки зрения психологии, дистракторы выполняют интерферентную функцию. То есть, они вторгаются в процесс мышления испытуемых, внося в него помехи, побуждают его сравнивать разные ответы к одному и тому же заданию, анализировать их, искать аргументы в пользу правильности (или неправильности) каждого ответа, выбирать правильный или, иногда, самый правильный ответ, если это написано в инструкции для испытуемых. Число дистракторов в каждом задании варьирует, в зависимости от содержания задания.

Качество дистракторов обычно оценивается двумя основными методами. Первый - умозрительным анализом корректности его содержания, предположительным процентом выбора испытуемыми целевой группы. Второй метод – эмпирическим подсчётом процента испытуемых, которые выбирают его при ответе на задание теста в процессе эмпирической апробации задания. Чем выше процент испытуемых, выбравших

данный неправильный ответ, тем выше его привлекательность для незнающих студентов. А значит выше и ценность (или качество) самого дистрактора. Эта прагматическая точка зрения не всегда верна.

Есть и третий метод, который рассмотрим на упрощённом примере задания об определении площади круга радиусом три сантиметра, имеющемся в литературе. Это задание с выбором одного правильного ответа.

Нажимайте на клавишу с номером правильного ответа:

1. ПЛОЩАДЬ КРУГА С РАДИУСОМ 3 см. РАВНА[24]

1) 9.00 cm^2

2) 18.85 cm^2

3) 28.27 cm^2

Поскольку каждый испытуемый может выбрать только один ответ из трёх предлагаемых взаимоисключающих случаев, то применима теорема сложения вероятностей: сумма вероятностей выбора любого одного из трёх ответов равна 1.

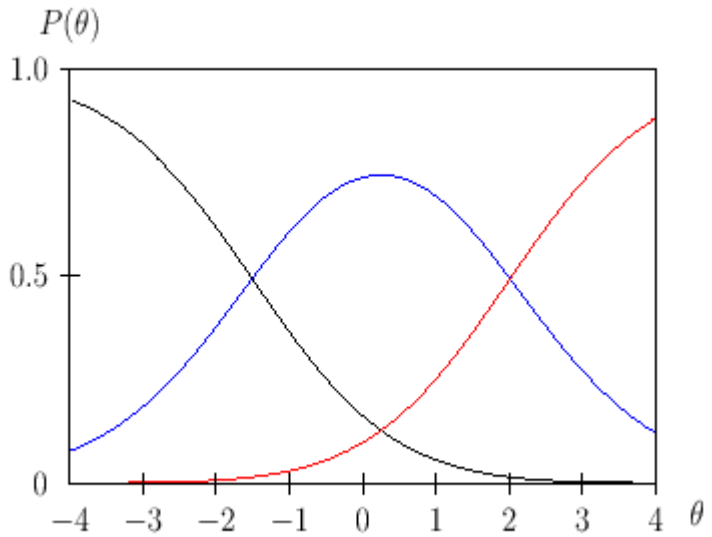

Рис.1. Вероятности выбора правильного ответа испытуемых на каждый из трёх ответов задания.

[24] Partchev, Ivailo. A visual guide to item response theory. Friedrich-Schiller-Universit®at Jena, 2004.

На рис. 1 приведены три графика[25], отражающих идею вероятностной функции выбора каждого из трёх ответов, из которых первый и второй неправильны, а третий – правильный. Первый и второй дистракторы имеют неодинаковую содержательную ценность. Эти кривые получены И. Партчевым посредством программы RUMM 2020.

Соответственно, первый слева график рис. 1 показывает, что по мере роста подготовленности испытуемых вероятность выбора первого ответа заметно уменьшается. Иначе говоря, выбор первого ответа в этом задании свидетельствует о полном незнании данного фрагмента учебного предмета.

График вероятности выбора второго ответа можно объяснить интерференцией – испытуемые путают формулы определения площади (третий, правильный ответ) и длины окружности (второй ответ). Он свидетельствует о знаниях хотя бы одной из двух формул, применённой, однако, неправильно. Из этого графика видно – как меняется вероятность выбора второго ответа в зависимости от уровня подготовленности испытуемых. Максимум вероятности выбора такого ответа приходится на испытуемых среднего уровня подготовленности.

И, наконец, график вероятности выбора третьего ответа может свидетельствовать о знании и умении применить формулу для определения площади круга. Этот график указывает на педагогически осмысленное поведение испытуемых с хорошей подготовкой: чем выше уровень подготовленности испытуемых, тем выше у них вероятность правильного ответа

Полученный математический факт вполне согласуется с педагогической логикой. Чем выше уровень подготовленности испытуемых, тем меньше следует ожидать выбора неправильного ответа. Таким образом,

[25] Там же.

применение IRT помогает лучше понять педагогическую ценность не только каждого задания теста, но и каждого ответа.