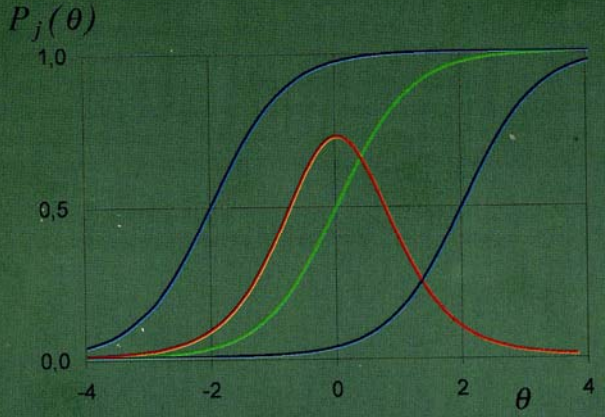
Ким В.С.

ТЕСТИРОВАНИЕ УЧЕБНЫХ ДОСТИЖЕНИЙ



Министерство образования и науки Российской Федерации
Федеральное агентство по образованию
Государственное образовательное учреждение
Высшего профессионального образования
«Уссурийский государственный педагогический институт»

В.С. КИМ

Тестирование учебных достижений

Монография

Уссурийск - 2007

ББК 74.04
К 40

Печатается по решению редакционно-издательского Совета Уссурийского государственного педагогического института.

Рецензенты:
доктор педагогических наук, профессор **Аванесов В.С**.
доктор химических наук, профессор **Вовна В.И.**

Ким В.С.
К 40 Тестирование учебных достижений. Монография. -Уссурийск: Издательство УГПИ, 2007. - 214 с.: ил.

ISBN 978-5-86733-165-8

Монография посвящена теоретическим и практическим проблемам тестирования учебных достижений. Приведен краткий обзор развития тестирования в России и за рубежом. Рассмотрены понятия надежности и валидности теста, структуры и формы тестового задания. Обсуждаются вопросы обработки результатов тестирования, как на основе классической теории тестов, так и с применением Item Response Theory к анализу качества тестовых заданий.

Монография предназначается преподавателям, учителям, аспирантам и всем, кто интересуется тестированием учебных достижений.

ББК 74.04

ISBN 978-5-86733-165-8

ВВЕДЕНИЕ

Тестирование учебных достижений является важной составной частью учебного процесса. Управление учебным процессом, как известно, является одним из определяющих факторов повышения его эффективности.

Процесс обучения, согласно Н.Ф.Талызиной[1], как один из видов управления требует циклического (замкнутого) управления, осуществляемого по принципу «белого ящика». Замкнутость системы управления обусловлена наличием цепи обратной связи. Н.Ф.Талызина рассматривает коррекцию процесса усвоения, за счет действия обратной связи как самостоятельный и важный этап управления.

Коррекция возможна, если есть возможность получения достоверной, объективной информации о состоянии системы, в том числе педагогической. В этой связи необходимо отметить важность диагностичности целей и задач, решаемых системой. Только наличие диагностичных целей позволяет контролировать состояние процесса обучения, а следовательно, его коррекцию и оптимизацию.

Согласно В.П.Беспалько[2] вариативными характеристиками, определяющими качество обучения, являются уровень усвоения деятельности и степень усвоения (автоматизации) деятельности.

Эти величины можно контролировать, то есть достижение высокого качества обучения является диагностичной целью. Наличие диагностичных целей позволяет организовать реально действующий процесс управления обучением.

Достижение высокого качества обучения возможно только при наличии объективных методов диагностики. К сожалению,

традиционная форма оценивания уровня знаний в форме опроса, экзамена, проводимого человеком, весьма субъективна.

По мнению К.Ингенкампа[3] при использовании пятибалльной шкалы преподаватель выставляет оценки с разбросом плюс, минус 1 балл, то есть с точностью 20%. Из этого следует, что за одни и те же знания, испытуемый может быть оценен разными экзаменаторами на «2», на «3» и на «4». Более того, К.Ингенкамп указывает, что один и тот же экзаменатор в разные моменты времени, например с интервалом в 1 месяц, также по разному оценивает один и тот же ответ (на экспериментах использовались видеозаписи ответов испытуемых).

Ясно, что столь неточный «измерительный прибор», каковым является человек, существенно снижает эффективность диагностики учебного процесса. По этой причине, в качестве контрольно-измерительного мероприятия выбирается тестирование. Сам процесс тестирования учебных достижений разбивается на три процесса: 1) разработка теста; 2) процедура тестирования; 3) обработка и интерпретация результатов тестирования.

При обработке результатов используется либо классическая теория тестирования, либо IRT (Item Response Theory), позволяющая измерять уровень достижений испытуемого в специальных единицах измерения – логитах.. Итерационные процедуры оптимизации тестовых заданий позволяют создавать надежные и валидные тесты. Особо следует отметить тот факт, что IRT позволяет получить числовые значения уровня достижений испытуемого в логитах на интервальной шкале. Наличие интервальной шкалы позволяет использовать мощный аппарат математической статистики для интерпретации полученных результатов.

Напомним, что оценки, выставляемые человеком-экзаменатором, размещены на порядковой шкале, что сильно ограничивает возможности математической обработки результатов контроля. Давно критикуемая теоретически и, тем не менее, широко используемая на практике идея расчета среднего балла как среднего арифметического не имеет под собой методологических оснований. По оценкам, например, из школьного классного журнала можно определить моду или медиану. Полученную медиану можно, если угодно, трактовать как средний балл, но надо ясно отдавать себе отчет в том, что это не среднее арифметическое всех оценок, выставленных в журнале.

Тестирование же лишено подобных недостатков, поскольку, при правильном применении, дает результаты на интервальной шкале. Помимо достоверности тесты обладают и высокой степенью

объективности. В практике любого преподавателя есть конфликтные случаи недовольства учащегося (студента) экзаменационной оценкой, в то же время подобные конфликты практически исключены при тестировании.

Отдавая должное объективности тестирования, необходимо еще раз подчеркнуть, что тесты должны быть надежными и валидными. Талызина Н.Ф.[4] приводит пример неудачного использования контролирующих устройств (тестеров). Программа контроля (тест) для этих устройств разрабатывалась различными преподавателями. Последовательная проверка одного и того же контингента учащихся по одной и той же теме, но по разным тестам, показала различные уровни достижений. Это говорит о том, что тесты были невалидными, а возможно и ненадежными. Разумеется, объективность контроля в этом случае низкая и такое тестирование использовать нельзя.

Важность тестов в учебном процессе давно осознана за рубежом. Там теория и практика тестирования развиваются уже сотню лет. В России (в Советском Союзе) интерес к этому виду контроля знаний возродился в 60-х годах прошлого века в связи с развитием программированного обучения. Несмотря на востребованность, ситуация с обеспеченностью тестологической литературой пока еще далека от идеальной. Появление новых учебных пособий, монографий, справочников по тестированию можно только приветствовать.

В данной монографии рассмотрены вопросы теории и практики тестирования учебных достижений.

Первая глава посвящена основным понятиям, определениям и терминам теории тестов, а также содержит краткие сведения о развитии тестирования в России и за рубежом.

Во второй главе рассмотрены формы тестовых заданий. Важность тщательного соблюдения формы задания далеко не сразу осознается теми, кто только приступает к разработке собственных тестов. Казалось бы, что тут может быть неясного? Составить 30-40 вопросов, придумать к ним ответы - вот и вся работа. Это очень глубокое заблуждение. Это настоящее искусство - создание хорошего задания в тестовой форме.

Можно испытать истинное эстетическое наслаждение, наблюдая как, поначалу неуклюжее, многословное, какое-то кургузое словесное сооружение превращается в ясное, прозрачное, предельно лаконичное задание, из которого невозможно убрать ни единого слова, ни единой запятой! У каждого тестового задания есть цель и все должно работать на достижение этой цели. Начертание и размер

шрифта, взаимное расположение элементов задания, место для ответов, графическое и цветовое оформление - все должно содействовать легкому и быстрому восприятию задания.

Третья глава посвящена статистической обработке результатов тестирования. Даже если разработчику удалось создать хорошие, отличные задания в тестовой форме, это еще не означает, что созданы тестовые задания. Только после испытания в реальных условиях становится ясно, работают задания теста или нет. Статистическая обработка, анализ результатов тестирования позволяет выявить неблагополучные задания, наметить пути их совершенствования. После внесенных исправлений тест вновь проверяется и вновь исправляется. Этот процесс повторяется неоднократно. Создание надежного, валидного теста с устойчивыми характеристиками - очень сложное и трудоемкое дело. Статистическая обработка результатов позволяет его облегчить.

В четвертой главе рассмотрены некоторые вопросы тестирования учебных достижений, важные для практического применения. Особое внимание уделено «человеческому фактору» в системе тестирования. Человек не машина, его поступки плохо формализуемы и трудно предсказуемы. Однако есть общие факторы, примерно одинаково влияющие на поступки людей. В их числе - мотивация человеческой деятельности. Рокуэлл Кент приводил пример «истины» - чем больше платят денег, тем больше человек работает. Эскимосы же, после обеда бросали топоры и пускались в разговоры. Они не желали работать весь день. На американский вопрос «Почему?» они отвечали - «Не интересно!». Это очень важно - учитывать мотивы поведения, в том числе и в тестировании.

Пятая глава посвящена очень злободневному вопросу применения «современной» теории тестов - Item Response Theory (IRT) и особенно модели Раша (Rasch Measurement). Слово «современная» взято в кавычки, потому, что она развивается уже около полувека, но до сих пор еще не вошла в широкую практику тестирования. Классической теории тестов в этом повезло больше. Главное достоинство IRT это то, что она позволяет получить результаты на интервальной шкале. Любая наука начинается с измерений. Если нельзя повторить эффект, измерить его, то нельзя ни подтвердить, ни опровергнуть исследовательскую гипотезу. Такая картина пока еще, к сожалению, характерна для гуманитарных наук. IRT делает революционный шаг вперед, давая исследователям мощный инструмент для подлинного измерения латентных (скрытых) свойств человека.

Модель Раша крайне необычна, она противоречит стандартной парадигме научных исследований. Как рассуждает исследователь? Если теория плохо описывает эмпирические данные, то ее надо улучшать. Георг Раш считает иначе, если эмпирические данные противоречат его теории, то эти данные следует отбросить, они недостоверны! Иногда говорят, что теория Раша это однопараметрический вариант IRT. Формально это так, но по самой сути, на концептуальном уровне, это совершенно другая, отдельная теория. Ведь когда А.Бирнбаум вводил в IRT второй и третий параметры, он пытался улучшить теорию, с тем чтобы она точнее описывала экспериментальные данные. Прадигма Rasch Measurement совершено иная - надо улучшать не теорию, а данные. С этим непросто согласиться, особенно исследователям в области естественных наук, но в тестировании учебных достижений это так.

IRT требует очень больших объемов статистических расчетов, причем в итерационных циклах. Без вычислительной техники ее практическое использование невозможно. В конце пятой главе кратко описано применение прикладного программного средства RUMM (Rasch Unidimensional Measurement Model), разработанное под руководством профессора Дэвида Эндрича (D.Andrich). Это программное обеспечение ЭВМ позволяет довольно легко и быстро осуществлять IRT-анализ результатов тестирования.

В заключение хочу выразить надежду, что данная монография окажется полезной всем тем, кто использует тестирование учебных достижений в своей деятельности.

ГЛАВА 1. ПЕДАГОГИЧЕСКОЕ ТЕСТИРОВАНИЕ

В данной главе мы рассмотрим некоторые общие вопросы педагогического тестирования и наиболее важные, для дальнейшего изложения, основные понятия, определения и термины.

1.1. КРАТКАЯ ИСТОРИЯ РАЗВИТИЯ ТЕСТОВ ДОСТИЖЕНИЙ

История тестов учебных достижений насчитывает, по мнению В.Кадневского, по крайней мере, несколько тысячелетий[5]. В.Аванесов указывает на факты, свидетельствующие о еще более древнем применении тестов[6].

Древние вавилоняне знали 400 клинописных знаков, использовали шестидесятеричную систему счета, умели вычислять проценты, измерять площадь и объем различных геометрических фигур. Среди изучаемых предметов были те, которые отвечают современному понятию «профессиональная пригодность». За 2200 лет до н.э. в Китае успешно действовала система проверки способностей и отбора персонала для различных государственных должностей. В частности проверялось умение писать, читать, знать порядок проведения придворных ритуалов и церемоний. В течение последующих 2000 лет в систему отбора чиновников были внесены экзамены по гражданскому праву, военному делу, финансам, сельскому хозяйству, географии[5].

По мнению А.Н.Майорова[7] одним из первых ученых, попытавшихся измерить различия между людьми в области

элементарных психических процессов, был англичанин Френсис Гальтон (Galton F.; 1882-1911).

Гальтон ввел в теорию тестирования три фундаментальных принципа, используемых и по сей день:

1. Применение серии одинаковых испытаний к большому количеству испытуемых.
2. Статистическая обработка результатов.
3. Выделение эталонов оценки.

Все современные тесты построены на основе статистической теории измерений, а идея эталона оценки лежит в основе определения теста как стандартизованного инструмента.

Термин «умственные тесты» ввел Дж. Кеттел (Cattel J., 1860-1944). Дж.Кеттел считал тест средством для проведения научного эксперимента с соответствующими требованиями к чистоте эксперимента. Такими требованиями он определял[7]:

1. одинаковость условий для всех испытуемых;
2. ограничение времени тестирования приблизительно одним часом;
3. в лаборатории, где проводится эксперимент, не должно быть зрителей;
4. оборудование должно быть хорошим и располагать людей к тестированию;
5. одинаковые инструкции и четкое понимание испытуемыми, что нужно делать;
6. результаты тестирования подвергаются статистическому анализу, находят минимальный, максимальный и средний результат, рассчитывают среднее арифметическое и среднее отклонение.

Эти идеи, выдвинутые Дж.Кеттелом, составляют основу для современной тестологии. Одинаковость условий для всех испытуемых, одинаковые инструкции и четкое их понимание испытуемыми – фундаментальные принципы, положенные в основу стандартизации процедуры проведения тестирования; ограничение времени, в настоящее время, после дополнительных исследований, устанавливается в зависимости от возраста испытуемых и особенностей применяемого инструментария; идеи статистической обработки результатов реализованы в достаточно сложных методах статистического анализа и моделирования[7].

Большой вклад в развитие тестов интеллекта внес французский психолог Альфред Бине (Binet A., 1857-1911). Совместно с Теодором

Симоном (Simon T., 1873-1961) он разработал тест, позволяющий дифференцировать нормальных и умственно отсталых детей.

В 1911-1912 годах американские психологи Л.Термен и Х.Чальдс дополнили тест Бине - Симона четырьмя новыми[8]:

1) «Образец обобщения», то есть пояснения сущности или морали басни.
2) Постепенный дополнительный тест по методу Эббингауза.
3) Тест для испытания запаса слов (из 100 слов).
4) Испытание «Практическое суждение» (тест мяча и поля).

Тест на практическое суждение показан на рис.1.1.1.

Задание теста формулировалось в графической форме (рис.1.1.1). Изображался круг, обозначавший поле, заросшее густой травой. Где-то в поле лежит мяч, увидеть который можно, только если подойти к нему не более чем на 10 шагов. Испытуемому нужно выбрать варианты таких траекторий передвижения, чтобы время поиска мяча было наименьшим.

Л.Термен и Х.Чальдс лучшими считали ответы «д» и «е». Интересно отметить, что это не совсем верно. Варианты «д» и «е» обеспечивают успешный поиск мяча, но, сравнение времен поиска для обоих вариантов, показывает, что они не равноценны. Предположив, что скорость перемещения во всех случаях одинакова, мы можем сравнивать не время, а длину пути (траектории) поиска. Из рисунка видно, что длина траектории поиска в случае «д» почти в 3 раза превышает длину в случае «е». Таким образом, верный ответ – «е».

Если первоначально развивалось психологическое тестирование, то в последующем, Маккол В.А. предложил различать тесты психологические - тесты умственного развития (Intelligence Test) и педагогические - тесты учебных достижений (Educational Test)[7].

Основоположником педагогических измерений считается Эдуард Ли Торндайк. Именно Торндайком были созданы первые научно обоснованные педагогические тесты, снабженные нормами.

Большой вклад в развитие теории тестирования внесли Spearman C.E.[9], Gulliksen H.[10], Guttman L.[11] Lord F.M. & Novick M.[12], Kuder G.F. & Richardson M.W. (теория надежности тестов)[13], Crocker Linda & Algina James[14] (современная классическая теория тестов).

В настоящее время за рубежом и в нашей стране широкое применение находит современная теория тестирования - Item Response Theory (IRT). Однопараметрический вариант IRT предложен Георгом Рашем (G.Rasch)[15]. Развитие IRT основывалось на появлении двух и трехпараметрических моделей - Birnbaum A.[16] . Обширная

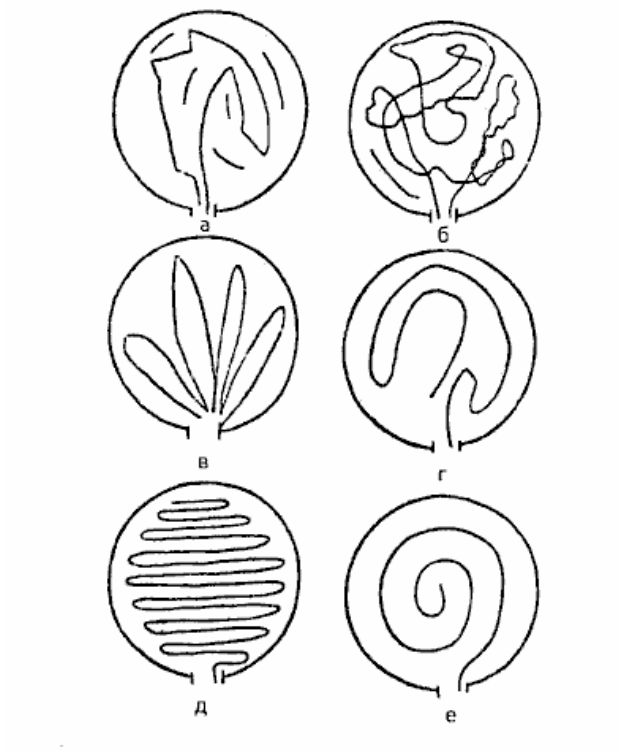


Рис.1.1.1. «Тест мяча и поля» Л.Термена и Х.Чальдса.

деятельность по развитию IRT осуществляется Д.Эндричем (D.Andrich)[17], Б.Райтом (B.Wright)[18]

Фундаментальный труд Анны Анастази «Психологическое тестирование»[19] представляет собой классическую работу, вобравшую в себя все достижения западной тестологии. В 2006 году вышло 7-е издание совместно с С.Урбиной, куда были добавлены главы по современным методам тестирования.

В своей «Педагогической диагностике» Карлхайнц Ингенкамп[3] утверждает, что традиционные способы оценки, существующие в системе образования, срочно нуждаются в систематическом дополнении объективными методами. При этом необходимо найти научное обоснование методики оценок. Без разумного использования информативных тестов добиться существенного улучшения в

оценочной практике учителей невозможно. Это, безусловно верное утверждение, справедливое не только для Германии, но и для России.

Наряду со всеобщим распространением тестов, нарастала и их критика. Дж.Равен[20] указывает на научные и этические аспекты экспансии тестологии в сфере образования. Он называет «безнравственным» пренебрежение ущербом, который практика тестирования наносит судьбам детей и интересам общества. Дж. Равен считает, что традиционные тесты достижений не могут должным образом оценить результаты педагогического процесса, в частности, они не подходят для выявления одаренности учащихся.

Отмечая справедливость его критики, следует все же признать, что развитие тестологии, дающей в руки педагоги качественный измерительный инструмент, явление нужное и прогрессивное. Правда, это должен быть не единственный измеритель, определяющий ход учебного процесса.

РАЗВИТИЕ ТЕСТОЛОГИИ В РОССИИ

Среди российских исследователей занимавшихся тестированием, можно назвать П.П.Блонского, Г.И.Залкинда, М.С.Бернштейна и др. К сожалению, в 1936 году вышло постановление ЦК ВКП(б) «О педологических извращениях в системе Наркомпросов». Тестирование было признано противоречащим советской идеологии со всеми вытекающими последствиями.

В послевоенные годы, работы в области тестирования начали возрождаться, а в 70-80-х годах прошлого столетия педагогическое тестирование стало усиленно развиваться в рамках технологии программированного обучения.

Важную роль в становлении отечественной тестологии сыграли работы Беспалько В.П.[21, 22] и Талызиной Н.Ф.[23, 24, 25] Согласно В.П.Беспалько процесс обучения должен быть технологичным и диагностичным. Если нет достоверной диагностики, то нет и учебного процесса. Н.Ф.Талызина, рассматривая вопросы управления процессом учения, анализирует проблемы педагогической оправданности применения тестов различного типа.

Работы отечественных и зарубежных тестологов были в основном изолированы от друг от друга. В СССР фундаментальные труды зарубежных тестологов были практически неизвестны.

В этой связи следует отметить трудно переоценимую деятельность В.С.Аванесова по применению и развитию передовых

идей и достижений зарубежной тестологии в отечественной теории и практике педагогического тестирования.

Под руководством В.С.Аванесова в 1985 году на базе Московского института стали и сплавов был организован Исследовательский центр по проблемам управления качеством подготовки специалистов. В этом центре началась планомерная переподготовка руководящих работников и преподавателей в системе высшего образования СССР. В Москву на краткосрочные (1 месяц) курсы съезжались преподаватели, доценты, профессора со всех регионов страны - от Дальнего Востока, до Прибалтики и Средней Азии. Именно этот период, видимо следует считать началом широкомасштабного применения современных научных методов в педагогическом тестировании.

В 1989 году В.С.Аванесовым было выпущено учебное пособие «Основы научной организации педагогического контроля в высшей школе»[26], внесшее большой вклад в развитие теории и практики отечественной тестологии. В.С.Аванесов является приверженцем современных технологий в образовании, большую просветительскую деятельность он осуществляет в глобальной сети Интернет. Его сайт www.testolog.narod.ru содержит массу постоянно обновляемой информации весьма полезной для тестологов. Журнал «Педагогические измерения», главным редактором которого является В.С.Аванесов, является авторитетным изданием, где публикуются работы отечественных и зарубежных тестологов. А.Н.Майоров называет Вадима Сергеевича Аванесова классиком отечественной тестологии и с этим трудно не согласиться.

Вместе с В.С.Аванесовым в Исследовательском центре начала свою деятельность М.Б.Челышкова. К этому моменту она защитила кандидатскую диссертацию и увлеченно читала в Исследовательском центре лекции по Item Response Theory (IRT). Это были очень актуальные лекции, следует отметить, что значимость IRT в тестологии возрастает с каждым годом. Ныне, профессор Марина Борисовна Челышкова широко известна в кругах тестологов, а ее учебное пособие «Теория и практика конструирования педагогических тестов»[27], вышедшее в 2002 году пользуется всеобщим признанием.

Очень важные вопросы стандартизации педагогических тестовых материалов подняты в работе Б.У.Родионова, А.О.Татура[28]. Педагогический тест является измерительным инструментом и это должен быть качественный инструмент, позволяющий получать достоверные результаты. В создании качественных педагогических

тестов чрезвычайно велика роль стандартов, которым должны соответствовать педагогические тестовые материалы.

Полный комплекс работ по составлению и использованию тестов школьных достижений представлен в работах А.Н.Майорова[7,29]. В нашей стране остро стоит проблема подготовки кадров для системы тестирования. А.Н.Майоров отмечает, что существует «миф о том, что для составления тестового инструмента нет необходимости иметь специальные знаний. В последние годы, особенно в связи с введением ЕГЭ, появилось множество книг с описанием тестов по любым школьным предметам. Следует понимать, что это не тесты, а некоторые совокупности сырых заготовок, которые следует еще переработать в задания в тестовой форме, а затем, если удастся – в тестовые задания. Только после этого можно говорить, что создан тест для той или иной предметной области.

Если работы А.Н.Майорова имеют больше практическую направленность, то работы Ю.М.Неймана и В.А.Хлебникова носят больше теоретический характер[30, 31, 32]. Им принадлежит русскоязычная трактовка названия теории Раша (G.Rasch) - «Теория моделирования и параметризации педагогических тестов» (ТМППТ). Эти работы имеют большое значение для развития теоретических основ отечественной тестологии. Традиционные контрольные процедуры страдают субъективизмом и неопределенностью оценок. В этой связи Ю.М.Нейман и В.А.Хлебников отмечают, что принципиально изменить ситуацию можно лишь в том случае, если подходить к оцениванию знаний как к процессу объективного измерения, а результаты таких измерений обрабатывать стандартными математическими методами и сопровождать стандартными характеристиками точности. Ими указывается, что педагогический тест, в отличие от, например, контрольной работы, можно рассматривать как своеобразный измерительный инструмент определенной разрешающей силы и точности.

Информационные и телекоммуникационные технологии оказывают сильное воздействие, как на организационные формы, так и на обработку результатов тестирования. В работе В.И.Нардюжева и И.В.Нардюжева[33] рассмотрены вопросы построения системы компьютерного тестирования. Программные разработки этих авторов использовались для организации абитуриентского компьютерного тестирования Федеральным центром тестирования Минобразования РФ (ЦТ МО РФ). Прикладные программные средства Tester - для проведения тестирования, Operator - для конфиденциальной передачи результатов тестирования в ЦТ МО РФ, StatInfo - для статистической

обработки результатов тестирования, показали себя как надежные и удобные программные продукты.

Термин «Дидактическая тестология» вводит Е.А.Михайлычев[34]. Если исходить из того, что дидактика - это теория обучения, а педагогика - теория и обучения и воспитания, то термин, предложенный Е.А.Михйлычевым представляется более точным, нежели термин «педагогическое тестирование». Однако следует отметить, что в научной терминологии уже устоялся термин «педагогическое тестирование». Е.А.Михайлычевым очень обстоятельно описаны проблемы валидизации теста и пути их решения.

Применение модели G.Rasch (Раш) к изучению латентных переменных в образовании в социально-экономических системах развивается в работах А.А.Маслака[35]. Следует отметить вклад А.А.Маслака в разработку конструктов, содержащих индикаторные переменные для социально-экономических систем, анализ точности педагогических измерений на основе модели Раша. В качестве эффективного инструмента в исследованиях А.А.Маслака используется программное средство RUMM (Rasch Unidimensional Measurement Model), разработанное под руководством профессора Д.Эндрича[17].

В монографии В.Ю.Переверзева[36] рассматриваются характеристики критериально-ориентированных тестов и их сравнение с нормативно-ориентированными тестами, описываются методики определения оптимального количества заданий в тесте. В справочном руководстве[37] приводится обширный справочный материал по разработке тестовых заданий, как для бланкового, так и для компьютерного тестирования.

Вопросы применения тестовых технологий для гуманитарных и экономических специальностей рассмотрены в учебном пособии Войтова А.Г.[38]. Приведена методика применения прикладного программного средства «СУБД Системное тестирование» для компьютерного тестирования.

На Дальнем Востоке большой вклад в развитие тестологии внес И.А.Морев[39]. В Тихоокеанском институте дистанционных образовательных технологий (ныне Открытый университет ДВГУ), руководимым профессором В.И.Вовной, И.А.Морев теоретически обосновал и реализовал на практике технологию «мягкого, непрямого» тестирования (зарубежные аналоги - «grading» и др.) в форме деловых компьютерных игр. Под руководством И.А.Морева был разработан ряд компьютерных программ-тестеров, среди которых следует отметить

программные пакеты STEACHER, DIALOG, PHRACON и DIDACTOR[40]. Большой статистический материал (несколько десятков тысяч испытуемых) позволил И.А.Мореву обнаружить важные закономерности, взглянуть на тестирование не только как на измерительный, но и как на полноценный дидактический инструмент. И.А.Моревым показано, что тестирование, при определенных технологических условиях, способно успешно стимулировать рост мотивации учащихся к учебе, рост показателей их обученности и обучаемости[40,41].

Проблемы использования IRT в учебном процессе вуза исследуются К.Т.Кузовлевой[42] в Дальрыбвтузе. В Тихоокеанском военно-морском институте В.В.Черненко проводит интересные исследования как в области технологии применения тестов достижений, так и в области интерпретации полученных результатов. Работы К.Г.Кречетникова[43,44] посвящены вопросам организация контроля и корректировочных действий в информационной образовательной среде вуза.

Педагогическое тестирование развивалось и в Уссурийском государственном педагогическом институте. С 1994 года в УГПИ разрабатывались тестовые задания по школьному и вузовскому курсам физики, информатики. Выполнялась статистическая обработка результатов тестирования, создавались компьютерные программы, как для тестирования, так и для обработки полученных результатов[45, 46, 47, 48, 49]. Технология «мягкого, непрямого» тестирования разрабатывается О.Н.Фалалеевой[50]. Для организации абитуриентского тестирования был создан региональный межвузовский центр тестирования.

Из приведенного, очень краткого и неполного обзора следует, что тестирование учебных достижений широко используется за рубежом и довольно высокими темпами развивается в России.

1.2. ОСНОВНЫЕ ПОНЯТИЯ И ТЕРМИНЫ ТЕСТИРОАНИЯ УЧЕБНЫХ ДОСТИЖЕНИЙ

Рассмотрим основные понятия и термины тестологии, которые нам потребуются в дальнейшем.

Для теории тестов, педагогических измерений, квалиметрии, можно выстроить следующую иерархическую структуру[51]:

Теория тестов ⊂ Теория педагогических измерений ⊂ Общая теория педагогического оценивания ⊂ Квалиметрия

Эта схема позволяет уяснить место и роль теории педагогических измерений в системе наук вообще и в педагогическом оценивании в частности.

Рассмотрим основные понятия и термины, необходимые для однозначного понимания дальнейших утверждений, суждений и выводов.

ИЗМЕРЕНИЕ — операция для определения отношения одной (измеряемой) величины к другой однородной величине, которая берётся за единицу. Получившееся значение будет численным значением измеряемой величины. Наука, предметом изучения которой являются все аспекты измерений, называется МЕТРОЛОГИЕЙ.

ПЕДАГОГИЧЕСКОЕ ИЗМЕРЕНИЕ - это процесс установления соответствия между оцениваемыми характеристиками обучаемых и точками эмпирической шкалы, в которой отношения между различными оценками характеристик выражены свойствами числового ряда[27].

Существует много определений ТЕСТА, довольно заметно отличающихся друг от друга. Приведем некоторые из них.

Согласно словарю ЕГЭ[52], ТЕСТ - это измерительная процедура, включающая инструкцию и набор заданий, прошедшая широкую апробацию и стандартизацию.

Рубинштейн С.Л.[53] дал следующее определение: ТЕСТ — это испытание, которое ставит своей целью градуирование, определение рангового места личности в группе или коллективе, установление её уровня.

Это определение сформулировано только с точки зрения достижения цели, не оговаривая, как эта цель достигается, а главное, не определяет тест как измерительный инструмент.

К.Ингенкамп[3] - ТЕСТИРОВАНИЕ - это метод педагогической диагностики, с помощью которого выборка поведения, репрезентирующая предпосылки или результаты учебного процесса, должна максимально отвечать принципам сопоставимости, объективности, надежности и валидности измерений, должна пройти обработку и интерпретацию и быть готовой к использованию в педагогической практике.

В определении К.Ингенкампа рассматривается метод, а не средство педагогической диагностики и никак не характеризуются задания теста.

В.С.Аванесов определяет ПЕДАГОГИЧЕСКИЙ ТЕСТ как систему параллельных заданий равномерно возрастающей трудности, специфической формы, позволяющую качественно и эффективно измерить уровень и оценить структуру подготовленности учащихся.

В одной из последних работ[54] В.С.Аванесов (2005) несколько смягчил формулировку:

ПЕДАГОГИЧЕСКИЙ ТЕСТ определяется как система параллельных заданий возрастающей трудности, специфической формы, которая позволяет качественно и эффективно измерить уровень и структуру подготовленности испытуемых.

Сравнение обоих определений показывает, что произошло исключение требования равномерности возрастания трудности заданий. Обусловлено это тем, что обеспечить возрастание трудности заданий достаточно легко. Чтобы достичь этого составитель тестовых заданий ориентируется на различную сложность элементов предметной области. Для каждого элемента составляются задания и затем эмпирически проверяются, что действительно получены задания различной трудности. В самом тесте задания располагаются в порядке возрастания трудности.

Требование же равномерности возрастания трудности задания чрезвычайно сложно реализовать на практике. Хотя такой тест обеспечил бы линейную шкалу трудностей, что снизило бы ошибку измерения.

Исключение требования равномерности возрастания трудности задания заметно упрощает создание теста. Отметим, однако, что в этом случае, шкала трудностей получается нелинейной, с неравномерным покрытием заданного диапазона трудности заданий

теста. Это, естественно, снижает точность педагогического теста как измерительного инструмента.

М.Б.Челышкова отмечает, что понятийный аппарат теории педагогических измерений еще полностью не сформирован. В частности не существует общепризнанного определения теста. Каждый исследователь отражает в определении теста свое видение проблемы педагогического тестирования.

Согласно М.Б.Челышковой итоговый НОРМАТИВНО-ОРИЕНТИРОВАННЫЙ ТЕСТ – это система тестовых заданий, упорядоченных в рамках определенной стратегии предъявления и обеспечивающих информативность оценок уровня и качества подготовки испытуемых[27]. М.Б.Челышкова очень осторожно подходит к формулировке теста, намеренно ограничивая его сферу применения и тип. Приведенное определение вполне подходит для тестов, предназначенных для ранжирования испытуемых.

А.Н.Майоров приводит следующее определение теста, разработанное в 1997-1998 гг. группой авторов при разработке понятийного аппарата тестологии:

ТЕСТ – это инструмент, состоящий из квалиметрически выверенной системы тестовых заданий, стандартизованной процедуры проведения и заранее спроектированной технологии обработки и анализа результатов, предназначенный для измерения качества и свойств личности, изменение которых возможно в процессе систематического обучения[7].

В результате анализа приведенных определений теста мы склоняемся к выводу, что приемлемым может оказаться следующее определение:

ПЕДАГОГИЧЕСКИЙ ТЕСТ - это система тестовых заданий различной трудности, которая позволяет качественно и эффективно измерить уровень и структуру подготовленности испытуемых.

Это достаточно лаконичное и полное определение основано на определении В.С.Аванесова с некоторыми изменениями. Рассмотрим эти отличия.

1. Вместо слова «задание» использован термин «тестовое задание». Это позволило исключить требование «специфической формы», поскольку оно содержится в понятии «задание в тестовой форме» и, следовательно, в понятии «тестовое задание».

2. Исключено требование «параллельности» заданий. Это требование введено В.С.Аванесовым для повышения «живучести» теста, с тем, чтобы обеспечить возможность многократного использования теста, за счет варьирования в нем параллельных заданий. С этой точки зрения это вполне обоснованное требование. Однако, если мы определяем тест как таковой, отвлекаясь от привлекательной для практики его применения свойства «непотопляемости», то требование параллельности можно исключить.

3. Требование «возрастающей трудности» заменено требованием «различной трудности». Дело в том, что если мы располагаем тестовыми заданиями различной, известной трудности, то, формируя тест, легко можем расположить их в любом порядке, в частности, в порядке возрастания трудности.

Некоторые авторы предлагают размещать задания в порядке уменьшения трудности, аргументируя это оптимальным распределением умственного напряжения тестируемых во времени. К концу тестирования, когда испытуемые утомлены, целесообразно предъявлять им более легкие задания[62].

Иногда предлагается дать возможность выбора задания самим испытуемым, которые будет соизмерять свои возможности с теми усилиями, которые им понадобятся при прохождении теста. Это позволит им показать наилучший результат.

При компьютерном тестировании зачастую используется случайный порядок предъявления заданий, при этом сам тест формируется «на лету». Тестовые задания автоматически извлекаются из банка заданий в соответствии с той или иной процедурой, заданной разработчиком теста[45]. Если порядок предъявления одинаков для всех испытуемых, то, находясь в одном компьютерном классе, за соседними компьютерами, они могли бы подглядывать за ответами других испытуемых. При случайном порядке предъявления заданий уменьшается вероятность подобного нарушения процедуры тестирования.

Все вышеприведенные доводы представляются обоснованными, но для окончательного вывода требуется убедительная экспериментальная проверка их справедливости.

Во всех приведенных определениях фигурирует понятие тестового задания, которое в свою очередь, требует определения.

Создание тестового задания - это трудоемкий процесс, включающий создание некоторой заготовки тестового задания и

собственно тестового задания. Приведем определения этих типов заданий.

М.Челышкова различает предтестовые и тестовые задания.

ПРЕДТЕСТОВОЕ ЗАДАНИЕ – это единица контрольного материала, содержание, логическая структура и форма представления которого удовлетворяют ряду специфических требований и обеспечивают однозначность оценок результатов испытуемых в выбранной шкале[27].

Предтестовое задание называется ТЕСТОВЫМ, если апостериорные количественные оценки его характеристик удовлетворяют определенным критериям, нацеленным на проверку качества содержания, формы и на выявление системообразующих свойств заданий теста[27].

В.Аванесов различает «задание в тестовой форме» и «тестовое задание».

ЗАДАНИЕ В ТЕСТОВОЙ ФОРМЕ - одно из основных понятий педагогической теории измерений. В.Аванесов определяет задание в тестовой форме как педагогическое средство, отвечающее следующим требованиям[54]:

1) цель;
2) краткость;
3) технологичность;
4) логическая форма высказывания;
5) определенность места для ответов;
6) одинаковость правил оценки ответов;
7) правильность расположения элементов задания;
8) одинаковость инструкции для всех испытуемых;
9) адекватность инструкции форме и содержанию задания.

Эти требования позволяют отличить задания в тестовой форме от остальных.

Сравнивая требования В.Аванесова с требованиями Дж.Кеттела к тесту (как системе заданий) приведенными выше, можно прийти к выводу, что они частично перекрываются, что соответствует логике тестирования как применения измерительного инструмента.

Рассмотрим более детально, перечисленные требования.

ЦЕЛЬ. Разработчик должен отчетливо представлять себе какова цель предъявления тестового задания испытуемому. Цель создания тестового задания зависит, в частности, от типа теста (критериально-ориентированный или нормативно-ориентированный), в который оно будет включено, от степени подготовленности

испытуемых, от вида тестирования (итоговое, текущее, для самоконтроля) и так далее.

Пожалуй, самым важным фактором, определяющим цель задания является получение информации о степени усвоения испытуемым той или иной единицы учебного материала. Формулированию цели заданий способствует создание технологической матрицы (тестовой решетки), представляющей собой перечень учебных тем с относительным распределением тестовых заданий по этим темам[7]. Распределение заданий в технологической матрице определяется важностью, объемом и количеством учебного времени.

Не менее важным, представляется определение вида умения, выявляемого тестовым заданием. В частности, в таксономии Б.Блума[55] выделяются следующие уровни усвоения учебного материала - «знание», «понимание», «применение», «анализ», «синтез», «оценка». Предлагаются и другие таксономии (В.П.Беспалько[56], М.Н.Скаткин[57], В.П.Симонов[58], М.В.Кларин[59]). Задания, посвященные одной и той же учебной единице, но ориентированные на тот или иной уровень усвоения, имеют разные цели и, по этой причине, будут иметь разное содержание.

КРАТКОСТЬ. Формулировка задания должна быть предельно лаконичной, но не в ущерб пониманию сути задания. Необходимо исключить повторы слов и, тем более, целых фраз. Чем лаконичнее задание, тем лучше оно воспринимается.

Рассмотрим пример. Пусть мы поставили целью выяснить, знает ли испытуемый, что объем вещества в жидком состоянии неизменен (первый уровень в таксономии Блума). Для достижения поставленной цели, сформулируем задание следующим образом:

1. ВЕЩЕСТВА, НАХОДЯЩИЕСЯ В ЖИДКОМ СОСТОЯНИИ, ХАРАКТЕРИЗУЮТСЯ СЛЕДУЮЩИМИ СВОЙСТВАМИ
+1) сохраняют неизменный объем
2) сохраняют неизменную форму

Теперь переформулируем задание, следуя требованию краткости. Во-первых, перенесем повторяющиеся фрагменты из блока ответов в основную часть задания.

2. ВЕЩЕСТВА, НАХОДЯЩИЕСЯ В ЖИДКОМ СОСТОЯНИИ, ХАРАКТЕРИЗУЮТСЯ СЛЕДУЮЩИМИ СВОЙСТВАМИ - СОХРАНЯЮТ НЕИЗМЕННЫМИ

+1) объем
2) форму

Во-вторых, попытаемся сократить основу задания.

3. ЖИДКОСТИ СОХРАНЯЮТ
+1) объем
2) форму

Во всех приведенных вариантах задания, поставленная цель достигается, но в третьем примере суть задания воспринимается гораздо легче. От испытуемого требуется меньше усилий, чтобы воспринять содержание задание. Экономя усилия испытуемого на технической процедуре восприятия задания, мы позволяем ему в большей степени сосредоточиться на выполнении задания.

ТЕХНОЛОГИЧНОСТЬ. Это требование порождено использованием технических, компьютерных средств в обучении. Если задания технологичны, то тестирование проводится быстро, экономично, объективно. Особенно это важно для компьютерного тестирования.

Рассмотрим пример. По техническим причинам легче создать компьютерную программу-тестер, которая работает в текстовом режиме. Технологичность в этом случае означает, что в тестовом задании нельзя использовать рисунки, чертежи, формулы. Это вполне допустимо для гуманитарных дисциплин и здесь текстовый режим успешно используется.

Совершенно другая ситуация складывается, например, для таких дисциплин как физика или математика. Здесь формулы, рисунки и чертежи очень важны. В 70 - 80-х годах прошлого века в основном создавались программы, работавшие именно в текстовом режиме. Разработчики тестов по физике, были вынуждены формулировать задания, используя самое примитивное построение формул, какое только позволял текстовый режим. О рисунках не было и речи. Поскольку это ограничивало невербальное сопровождение тестового задания, то в последующем стали разрабатываться программы-тестеры, работающие в графическом режиме, хотя это и потребовало

усложнения программ. В результате содержание требования технологичности изменилось.

ЛОГИЧЕСКАЯ ФОРМА ВЫСКАЗЫВАНИЯ. Согласно В.С.Аванесову, это средство упорядочения и эффективной организации содержания теста. Эта форма во многих случаях заменяет вопросы. Логическое преимущество задания в тестовой форме заключается в возможности естественного превращения утверждения, после ответа испытуемого, в форму истинного или ложного высказывания[54]. Этого же мнения придерживается и М.Б.Челышкова.

Следует отметить, что не все разделяют это мнение. А.Н.Майоров считает, что хорошо сформулированное задание в вопросительной форме ничем не уступает хорошо сформулированному вопросу в форме утверждения. Если попытаться в утвердительной форме вопроса поставить два отрицания, то такое задание становится совершенно непонятным[7]. Голландский институт CITO дает рекомендацию: «Используйте прямые вопросы. Предпочтительнее применять прямые вопросы, представляющие собой полное предложение с вопросительным знаком в конце»[60] (из книги А.Н.Майорова). В.Ю.Переверзев[37] также считает, что основа тестового задания может формулироваться в вопросительной форме, что позволяет заданиям соответствовать более высоким познавательным уровням таксономии.

По этому поводу можно сказать следующее. Наш личный опыт показывает, что всегда удается сформулировать задание с выборов в форме утверждения. Иногда возникают ситуации, когда, казалось бы, задание формулируется только в виде прямого вопроса. Однако после тщательного анализа цели задания, его содержания, все же удается подобрать утвердительную форму задания. В результате задание получается, по крайней мере, не хуже, чем задание в форме прямого вопроса. В частности, все примеры, где используется прямой вопрос, из книги А.Майорова, можно преобразовать в утвердительную форму. Рассмотрим, например, задание №45 у А.Майорова[7].

КТО АВТОР ПАМЯТНИКА ПУШКИНУ НА ПЛОЩАДИ ИСКУССТВ?

1) Аникушин
2) Орловский
3) Козловский
4) Трубецкой
5) Опекушин

Теперь преобразуем его в утвердительную форму.

АВТОРОМ ПАМЯТНИКА ПУШКИНУ НА ПЛОЩАДИ ИСКУССТВ ЯВЛЯЕТСЯ

1) Аникушин
2) Орловский
3) Козловский
4) Трубецкой
5) Опекушин

Рассмотрим пример 3.6. для когнитивного уровня «Оценка» в таксономии Блума из книги В.Переверзева[37].

Если $0 < a < 1$ и $b > 1$, то, какое из следующих выражений имеет наибольшее значение?

1) $(b/a)^2$
2) b/a
3) $(a/b)^2$
4) a/b
5) Нельзя определить из данных задачи

Переформулируем это задание.

Если $0 < a < 1$ и $b > 1$, то наибольшее значение имеет выражение

1) $(b/a)^2$
2) b/a
3) $(a/b)^2$
4) a/b

При переформулировании задания нами использованы принципы импликации и однородности (отброшен пятый вариант ответа), описанные в главе 3, и требование логической формы высказывания.

Из этих примеров видно, что переформулировать задание в утвердительную форму можно. Таким образом, требование логической формы высказывания, по нашему мнению, следует считать обоснованным.

Остановимся на вопросе размещения ответа в содержательной части (основе) задания. Желательно, чтобы ответ размещался в конце задания, как показано в вышеприведенных примерах. Однако в

некоторых случаях допустимо встраивать ответ внутрь основы задания.

ЗВУКОВЫЕ ВОЛНЫ В ____ РАСПРОСТРАНЯЮТСЯ СО СКОРОСТЬЮ 330 м/с.

1) воде
+2) воздухе
3) стекле

В месте прочерка вставляется выбранный ответ, и утверждение превращается в истинное высказывание. Отметим, что все же надо пытаться размещать ответ в конце задания.

СО СКОРОСТЬЮ 330 м/с ЗВУКОВЫЕ ВОЛНЫ РАСПРОСТРАНЯЮТСЯ В

1) воде
+2) воздухе
3) стекле

Во втором примере задание читается несколько хуже, не так «гладко», как в первом.

ОПРЕДЕЛЕННОСТЬ МЕСТА ДЛЯ ОТВЕТА является одним из внешних и сущностных признаков задания в тестовой форме[54]. Важным требованием к заданию, является «экономность» в его оформлении. На бланке (дисплее компьютера) не должно быть ничего лишнего. Форма задания должна всячески минимизировать усилия испытуемых на восприятие его содержания. Определенность места для ответа – один из приемов такой минимизации. Представим себе, что место для ответа меняется от задания к заданию (в нижнем правом углу, в левом верхнем, посередине и т.д.), насколько бы это усложнило работу испытуемого. Любые ненужные усилия испытуемых, увеличивают вероятность ошибочного ответа, что приводит к снижению результата тестирования, по причинам, не связанным с его уровнем знаний. Ясно, что такой тест, как измерительный инструмент, будет некачественным.

ОДИНАКОВОСТЬ ПРАВИЛ ОЦЕНКИ ОТВЕТОВ заключается в том, что все испытуемые находятся в равном положении. Это необходимое условие использования теста в качестве средства измерения.

ПРАВИЛЬНОСТЬ РАСПОЛОЖЕНИЯ ЭЛЕМЕНТОВ ЗАДАНИЯ. Это требование заключается в строгом соблюдении размещения элементов (блоков) задания согласно его структуре. О структуре задания будет сказано далее.

ОДИНАКОВОСТЬ ИНСТРУКЦИИ ДЛЯ ВСЕХ ИСПЫТУЕМЫХ. Это очевидное требование, в противном случае появится большая систематическая погрешность измерения.

АДЕКВАТНОСТЬ ИНСТРУКЦИИ ФОРМЕ И СОДЕРЖАНИЮ ЗАДАНИЯ означает взаимное соответствие компонентов, что необходимо для выполнения заданием своей функции. Несоответствие формы содержанию, и, наоборот, содержания форме, вызывает ошибку понимания смысла задания[54].

Следует различать задания в тестовой форме и тестовые задания. Только после статистической проверки задание в тестовой форме может стать тестовым заданием.

Тест состоит не из заданий в тестовой форме, не из вопросов и задач, а только из тестовых заданий[54].

ТЕСТОВОЕ ЗАДАНИЕ - это составная единица теста, отвечающая требованиям к заданиям в тестовой форме и, кроме того, статистическим требованиям:

1) известной трудности;

2) дифференцирующей способности (достаточной вариации тестовых баллов);

3) положительной корреляции баллов задания с баллами по всему тесту, а также другим математико-статистическим требованиям[54].

Отметим также, что тестовые задания должны удовлетворять условию локальной независимости[61]. В.Переверзев[37] отмечает, что ответ испытуемого на каждое тестовое задание не подвергается влиянию и статистически независим от ответа на любое другое тестовое задание. Локальная независимость предполагает, что испытуемый, отвечая на задание, не может использовать добавленное знание, полученное из ответа на любое другое тестовое задание.

Из локальной независимости, в частности, следует, что компьютерные тестирующие программы не должны информировать испытуемого об успешности выполнения очередного задания, например, выводить на дисплей 0 или 1 баллов (для дихотомического случая). Допустим, что испытуемый не знает правильного ответа на какое-либо задание. В этом случае он может попытаться угадать

верный ответ. Если попытка окажется успешной, о чем ему сообщит компьютерная программа-тестер, то испытуемый получит добавленное знание. Это может помочь ему справиться с каким-либо другим заданием, то есть условие локальной независимости заданий будет нарушено.

Рассмотрим пример нарушения условия локальной независимости тестовых заданий. Пусть испытуемый, при выполнении нижеприведенного задания, угадал правильный ответ (третий)

СО СКОРОСТЬЮ 330 м/с ЗВУК РАСПРОСТРАНЯЕТСЯ В
1) воде
2) стекле
+3) воздухе

В результате выполнения задания он получил новое знание – оказывается скорость звука в воздухе равна 330 м/с. Пусть затем, ему попалось такое задание:

СО СКОРОСТЬЮ 1500 м/с ЗВУК РАСПРОСТРАНЯЕТСЯ В
+1) воде
2) стекле
3) воздухе

Анализируя ответы, испытуемый уже знает, что третий – неверный. В результате задание утратило один дистрактор и превратилось в задание с двумя ответами. Если бы задания изначально имели два ответа, то из-за нарушения условия локальной независимости, второе задание вообще бы перестало быть тестовым.

1.3. ВРЕМЯ ТЕСТИРОВАНИЯ

Выполнение тестового задания требует определенного времени. Общее время тестирования определяется количеством и сложностью заданий. Должно ли это время быть ограниченным или не ограниченным - определяется конкретной ситуацией, в которой применяется тест.

А.Майоров указывает, что каждый тест имеет оптимальное время тестирования, уменьшение или превышение которого снижает качественные показатели теста. В.Аванесов считает время выполнения системообразующим фактором при разработке и использовании теста.

Такое внимание этому вопросу уделяется по той причине, что неверно установленное время тестирования не позволяет тестовым заданиям достичь своей цели - проверить знает ли испытуемый тот или иной элемент, проверяемой дидактической единицы.

К чему приведет, например, слишком малое время тестирования? Слабые учащиеся не справятся с тестом потому, что имеют слабую подготовку, а сильные - потому, что не имели достаточно времени на выполнение заданий. У всех испытуемых будут примерно одинаково низкие индивидуальные баллы, то есть произойдет уменьшение дифференцирующей способности теста. Результаты такого теста не будут объективно отражать уровень подготовленности учащихся[*].

Так же неблагоприятно влияет на тестирование и слишком большого времени выполнения теста. В этом случае мы также получим негативное воздействие на измерительные качества теста. В частности, сильные учащиеся, досрочно завершив тестирование, в оставшееся время начнут шуметь, отвлекать тех, кто еще не закончил тестирование, подсказывать им и т.д. (нарушение процедуры тестирования). Другие испытуемые, будут долго сидеть над заданиями, не решаясь выбрать ответ. Это вызовет у них утомление, снижение концентрации внимания, расслабление, что также снижает точность тестирования. Утомление обусловлено чувством усталости, которое проявляется процессами торможений в клетках коры головного мозга. В состоянии утомления, испытуемый способен показать лишь малую долю своих истинных способностей. Тестировать его в этом случае бесполезно, так что мы не добьемся цели тестирования.

[*] Вспомним, как резко меняется сила игры шахматиста в зависимости от того, играет он двухчасовую партию или пятиминутный блиц.

А.Майоров приводит следующие эффекты проявления утомления:

1) на поведенческом уровне - приводит к уменьшению скорости и точности работы;

2) на физиологическом уровне - приводит к повышению инерции в динамике нервных процессов;

3) на психологическом уровне, ведет к нарушению качеств внимания, процессов памяти, степени адекватности функционирования интеллектуальных процессов;

4) происходят сдвиги в эмоционально - мотивационной сфере.

Как же определить оптимальное время тестирования? Вопрос не простой. Начнем с практических рекомендаций. На выполнение одного задания обычно отводится 30-60 секунд. Если задания соответствуют простому «узнаванию» (первый уровень таксономии Блума), то, как показывает наш опыт, вполне достаточно 5 - 10 секунд. По мере продвижения на верхние уровни таксономии Блума, это время должно увеличиваться в десятки раз. Имея опыт, еще на этапе разработки тестового задания можно грубо оценить время его выполнения. Суммарное время по всем заданиям даст общее время тестирования.

Длина теста (количество заданий) и время тестирования - тесно связанные и, в определенном смысле эквивалентные характеристики, но определяющим является все же именно время тестирования, поскольку оно задает порог утомления, за которым тест начинает терять свои измерительные свойства.

Теоретически рассчитать это время невозможно, поэтому рекомендуется использовать эмпирические данные по результатам первичной апробации теста.

Оптимальное время тестирования - это время от начала процедуры тестирования до момента наступления утомления[7]. Как определить момент начала утомления? В.Аванесов предлагает отслеживать момент достижения максимума дисперсии тестовых результатов (рис.1.3.1.).

А.Майоров считает, что оптимальное время тестирования соответствует не максимуму дисперсии, а моменту точке начала ее увеличения[7].

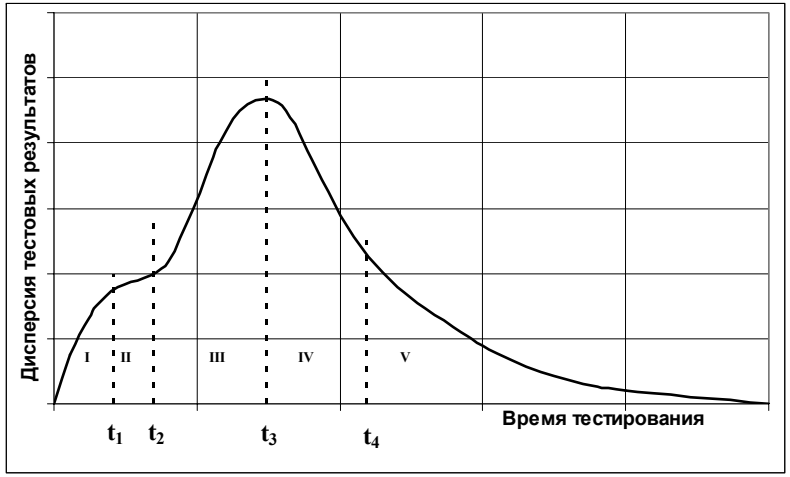


Рис.1.3.1. Время тестирования.

Рассмотрим детальнее эти разногласия. На рисунке показана гипотетическая зависимость дисперсии тестовых результатов от времени тестирования.

Предположим, что все испытуемые одновременно начинают и одновременно заканчивают сеанс тестирования (досрочное завершение невозможно). Очевидно, что при очень малом времени тестирования, все испытуемые одинаково не выполнят ни одного задания, то есть дисперсия должна отсутствовать. При очень больших значениях времени тестирования (большая длина теста) ввиду высокой степени утомления, все испытуемые также не смогут выполнить тест, то есть дисперсия снова будет близка к нулю. При оптимальном времени тестирования (согласно В.Аванесову это t_3) дисперсия будет максимальной. А.Майоров, считает, что оптимальное время тестирования соответствует точке t_2.

Весь временной интервал разбивается на пять характерных областей I, II, III, IV и V. В области I (очень малые времена тестирования) дисперсия быстро растет в связи с тем, что время реакции у испытуемых разное и, поэтому, часть испытуемых начнет успевать справляться с некоторыми заданиями теста. Произойдет

дифференциация испытуемых, что и обусловит быстрый рост дисперсии на начальном участке.

Далее, во второй области рост дисперсии замедляется, так как теперь испытуемые с замедленной реакцией тоже начнут успевать выполнять задания. Темп увеличения дифференциация испытуемых уменьшится, то есть, замедлится рост дисперсии тестовых баллов.

В третьей области III скорость возрастания дисперсии снова увеличится. Это происходит по причине того, что теперь время тестирования достаточно велико и большинство испытуемых успевают полноценно проанализировать задания. Здесь начинает работать другой механизм - дифференциация испытуемых происходит не за счет различия во времени реакции, а за счет различия в уровне подготовленности.

В точке t_3 дисперсия достигает максимума и далее, в области IV, начнет снижаться. Уменьшение дисперсии обусловлено усилением утомления испытуемых. В области V утомление становится настолько сильным, что дисперсия тестовых баллов падает практически до нуля.

В области III, утомление испытуемых, появившись, начинает воздействовать на дисперсию тестовых баллов, а в точке t_3 становится насколько сильным, что начинает снижать дисперсию.

Из приведенного анализа следует, что точка зрения В.Аванесова предпочтительнее, во всяком случае, для нормативно-ориентированного тестирования. Основным доводом в пользу этого является то, что важнейшей задачей теста является дифференциация испытуемых. В точке t_3, эта дифференциация будет в основном обеспечена именно различием в уровне подготовленности испытуемых.

Таким образом, для эмпирического определения оптимального времени тестирования, необходимо провести серию сеансов различной длительности. Эти серии сеансов неоднократно повторить на выборках испытуемых, как можно более близких по своим характеристикам. После обработки собранного статистического материала, необходимо построить функцию, как показано на рисунке 1.3.1, и определить значение момента времени t_3. Это и будет оптимальное время тестирования.

Согласно рекомендации С.Отиса, в качестве оптимального времени тестирования приближенно можно принять время, в течение которого с тестом справляются не более пяти процентов испытуемых (А.Майоров,2001).

До сих пор мы обсуждали время тестирования как таковое, абстрагируясь от личности испытуемого. Зависимость на рис.1.3.1 –

гипотетическая, основанная на умозрительных предположениях. В.В.Черненко[62] приводит экспериментальные данные по временной зависимости степени утомления (рис.1.3.2)[63].

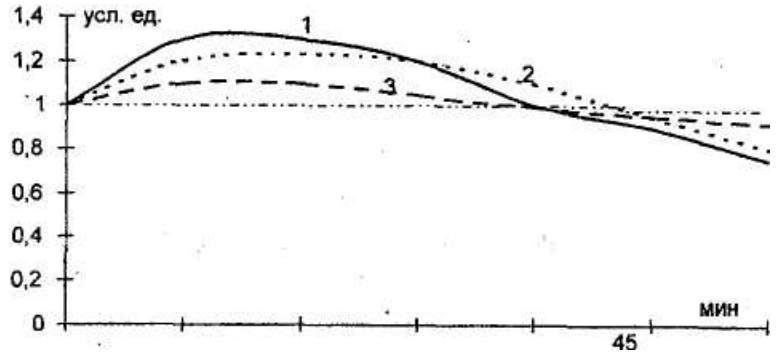


Рис.1.3.2. Изменение относительного объёма воспринимаемой информации по зрительному (1), слуховому (2) каналам в течение занятия.

Согласно В.В.Черненко, в первые 9 минут эксперимента объем информации воспринимаемой осознанно как по зрительному, так и по слуховому каналам достигает своего максимума. Далее, в течение последующих 18 минут плавно, но незначительно снижается, а в последующие 9 минут для зрительного канала достигает первоначального значения, для слухового канала достигает первоначально значения в течение 18 минут. По истечении 45 минут относительный объем осознанно воспринимаемой информации довольно значительно падает[62].

Если предположить, что занятия по теоретическим дисциплинам и тестирование требуют равного интеллектуального напряжения, то из приведенной зависимости следует, что утомляемость испытуемых начинает заметно проявляться через время t_y равное 36 минутам после начала тестирования. Оптимальное время тестирования t_3 соответствует моменту, когда положительный эффект обусловленный увеличением времени тестирования будет компенсирован отрицательным воздействием утомления испытуемых. Время t_3 должно быть немного больше времени t_y - запаздывание обеспечивает попадание в область максимума дисперсии тестовых результатов. Тогда из этих данных следует, что t_3 примерно равно 40 -

45 мин. Эти значения находятся в удовлетворительном согласии с рекомендациями, ограничивать длину теста 50-60 заданиями. Если на одно задание отводить 30-60 секунд, то общее время тестирования составит примерно 50 минут. А.И.Буравлев и В.Ю.Переверзев[64] показали, что критериально-ориентированный тест из 50 заданий может обеспечить надежность равную 0,9. Иными словами, тест из 50-60 заданий с одной стороны обеспечивает достаточно высокую надежность, а с другой – эффект утомления для такого теста еще слабо влияет на результаты.

В.В.Черненко отмечает, что источником систематической погрешности может стать пренебрежение суточным и недельным распределением момента начала тестирования.

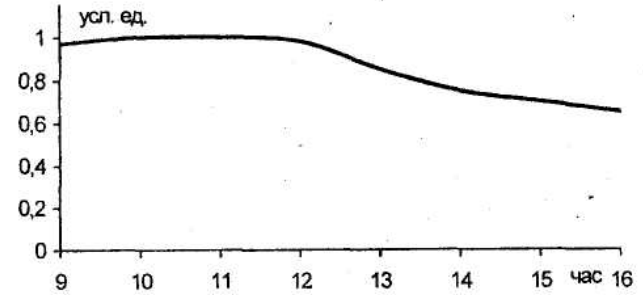


Рис.1.3.3. Изменение относительного объема воспринимаемой информации по часам учебного дня

Зависимость, показанная на рис.1.3.3, свидетельствует о том, что проведение тестирования в послеобеденное время приведет к снижению результатов по сравнению с утренним тестированием[62]. Пренебрежение этим эффектом может крайне негативно сказаться на результатах нормативно-ориентированного тестирования. Получается, что ранг испытуемого зависит от того, когда его тестировали – утром или вечером.

Аналогичный эффект проявляется при проведении тестирования в различные дни недели. Результаты исследований[63],

представленные на рис. 1.3.4, показывают, что день недели может вносить систематическую погрешность в результаты тестирования.

Исследования В.В.Черненко показывают, что указание оптимального времени является необходимым, но недостаточным параметром теста. При определении выборки стандартизации необходимо указывать день недели и часы тестирования. Это особенно важно для нормативно-ориентированных тестов.

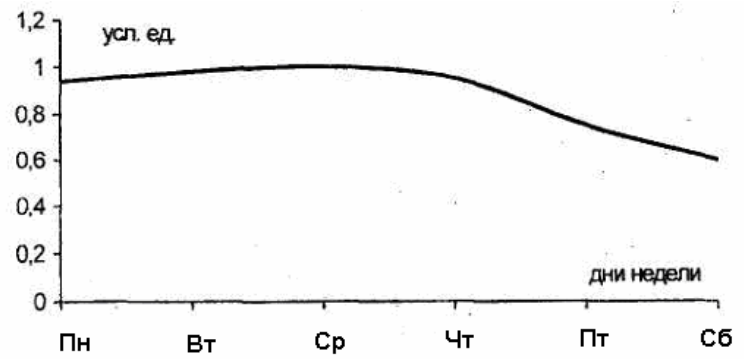


Рис.1.3.4. Относительные объемы воспринимаемой информации по дням недели.

По данным А.Анастази[19] результаты теста интеллекта для одного и того же испытуемого в начале недели могут дать показатель равный 110, а в конце недели – 80. Вполне возможно, что это связано со снижением работоспособности к концу недели.

Таким образом, при определении времени тестирования необходимо учитывать следующие рекомендации:

1) время тестирования определяется по расположению максимума дисперсии тестовых результатов и не должно превышать 60 минут;

2) длина теста не должна превышать 60-70 заданий, в предположении, что на выполнение одного задания требуется не более одной минуты;

3) тестирование необходимо проводить в первой половине дня;

4) тестирование желательно проводить в середине недели.

1.4. НОРМАТИВНО-ОРИЕНТИРОВАННЫЕ И КРИТЕРИАЛЬНО-ОРИЕНТИРОВАННЫЕ ТЕСТЫ

Педагогическое тестирование широко применяется для контроля знаний учащихся в различных целях. По целям применения педагогических тестов их можно разделить на два больших класса - нормативно - ориентированные и критериально - ориентированные[65, 66, 67, 68].

НОРМАТИВНО-ОРИЕНТИРОВАННЫЙ тест (norm-referenced test) позволяет ранжировать испытуемых по уровню знаний. Такой тест позволяет сравнивать учебные достижения испытуемых друг с другом.

Целью нормативно-ориентированного теста является упорядочение испытуемых по уровню их подготовленности. В результате может оказаться, что все испытуемые плохо справились с тестом - получили низкие индивидуальные баллы. Тем не менее, и в этом случае можно ранжировать испытуемых - кто-то получил низкий балл, а кто-то еще ниже. Возможны случаи, когда какое-то задание не дифференцирует испытуемых, например, задание легкое и все успешно на него ответили. И наоборот, очень трудное задания и все на него не ответили. Такие задания не позволяют провести ранжирование и, поэтому, должны быть удалены из теста. Если все испытуемые не ответили ни на одно задание, или верно ответили на все задания, то нормативно-ориентированный тест не работает, так как не позволяет достичь поставленной цели и подлежит дальнейшей переработке. Отметим, что, возможно, этот тест неплохо будет работать как критериально-ориентированный.

КРИТЕРИАЛЬНО-ОРИЕНТИРОВАННЫЙ тест (criterion-referenced test) позволяет выявить степень усвоения испытуемым определенного раздела в заданной предметной области. Эти тесты появились в 60-х годах прошлого века, то есть значительно позже нормативно-ориентированных. Критериально-ориентированные тесты в свою очередь делятся на domain-referenced test (ориентированные на предметную область) и mastery-tests (квалификационные тесты). Целью критериально-ориентированного теста является выяснение - знает ли испытуемый стандартный учебный материал (предмет, раздел, тему). В результате тестирования может оказаться, что все испытуемые успешно выполнили все задания. Это означает, что они освоили учебный материал. Если все испытуемые не справились с заданиями теста, то это означает, что учебный материал не усвоен. В обоих случаях тест выполнил свою задачу.

В дальнейшем нам потребуются следующие определения:

ОБЛАСТЬЮ СОДЕРЖАНИЯ теста называется тот полный объем знаний, умений и навыков, который должен быть усвоен учащимися в результате определенного курса обучения и овладение которым измеряется критериально-ориентированным тестом[65].

КРИТЕРИАЛЬНО-ОРИЕНТИРОВАННЫЙ педагогический тест представляет собой систему заданий, позволяющую измерить уровень учебных достижений относительно полного объема знаний, умений и навыков, которые должки быть усвоены учащимися[66].

Теперь перейдем к сравнению нормативно-ориентированных и критериально-ориентированных тестов.

Внешне оба типа тестов имеют много общего - в них используются тестовые задания сходные по форме, эти задания сопровождаются похожими инструкциями, выполняются задания одинаковым образом. Но, несмотря на внешнюю схожесть, это совершенно разные тесты. Они имеют следующие различия[65].

п.1. ЦЕЛЬ СОЗДАНИЯ ТЕСТА. Нормативно-ориентированные тесты создаются специально для того, чтобы сравнить испытуемых в той области содержания, для которой тест предназначен. Эти тесты можно использовать, например, для отбора абитуриентов при поступлении в вузы. В тех случаях, когда конкурс составляет несколько человек на одно место, возникает проблема ранжирования испытуемых с тем, чтобы выбрать наилучших.

Критериально-ориентированные тесты нужны для аттестации испытуемых в определенной области содержания. Такие тесты используются в итоговом тестировании, например по завершении обучения в среднем общеобразовательном учреждении. Здесь важно выяснить - усвоена ли в надлежащем объеме школьная программа. Вопросы ранжирования тут не играют большой роли.

Если критериально-ориентированные тесты использовать в качестве нормативно-ориентированных, то ввиду малой дисперсии тестовых результатов, эти результаты будут отличаться низкой надежностью. Справедливо и обратное - применение нормативно-ориентированных тестов в критериально-ориентированном тестировании также даст малонадежные результаты. Это обусловлено сильной вариацией тестовых заданий по трудности в нормативно-ориентированном тесте.

п.2. УРОВЕНЬ ДЕТАЛИЗАЦИИ ОБЛАСТИ СОДЕРЖАНИЯ.

Разработка теста начинается с создания его спецификации и эти спецификации для обоих типов тестов сильно отличаются.

Спецификации критериально-ориентированных тестов гораздо детальнее описывают элементы области содержания, поскольку это позволит адекватно интерпретировать результаты тестировании. Для нормативно-ориентированных тестов уровень детализации области содержания гораздо ниже. Для этих тестов гораздо важнее получить вариативные тестовые задания.

п.3. СТАТИСТИЧЕСКАЯ ОБРАБОТКА результатов тестирования. Шкалированные баллы нормативно-ориентированного тестирования основываются на тестовых нормах, полученных на «выборках стандартизации». При критериально-ориентированном тестировании тестовые баллы не связаны с какой-либо нормативной группой испытуемых. Обычно тестовый балл отражает долю правильно выполненных заданий и выражается в процентах.

п.4. АНАЛИЗ И ОТБОР ТЕСТОВЫХ ЗАДАНИЙ.

Для нормативно-ориентированных тестов большое значение имеют статистические характеристики - уровень трудности задания, его дифференцирующая способность. Если задание имеет средний уровень трудности и высокую дифференцирующую способность, то оно считается хорошим для нормативно-ориентированного теста.

Эти статистические характеристики не имеют большого значения для критериально-ориентированного теста. Здесь главным критерием для включения задания в тест является соответствие специфике и элементу области содержания.

п.5. РАСПРЕДЕЛЕНИЕ ИСПЫТУЕМЫХ по индивидуальным баллам имеет различный характер для обоих видов тестов. Для нормативно-ориентированного теста кривая распределения симметрична и близка к гауссовой кривой (рис.1.4.1). В случае критериально-ориентированного теста эта кривая несимметрична и обычно сдвинута в область высоких индивидуальных баллов (рис.1.4.2).

п.6. ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ ТЕСТИРОВАНИЯ.

Поскольку цели нормативно-ориентированного и критериально-ориентированного тестирования различны, то и интерпретация полученных данных будет различной.

Результаты нормативно-ориентированного тестирования интерпретируются на основе статистически обоснованных тестовых норм. При этом имеется возможность определить положение испытуемого относительно нормативной группы. Информации же о том, какие как усвоены те или иные разделы, элементы области содержания, нормативно-ориентированный тест дает мало.

Рис.1.4.1. Нормативно-ориентированный тест.

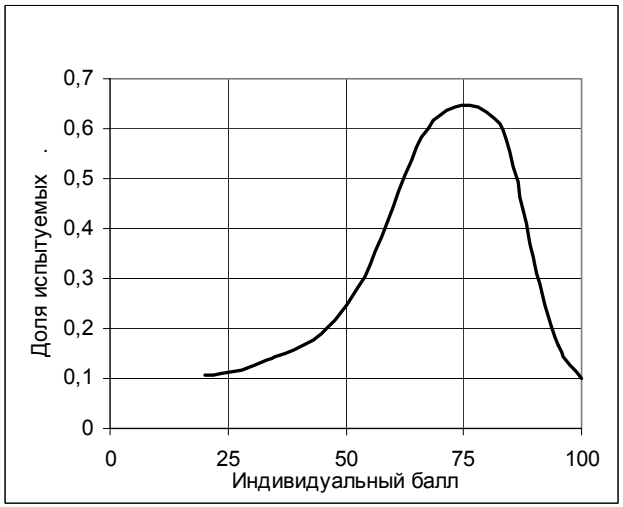


Рис.1.4.2. Критериально-ориентированный тест.

Результаты критериально-ориентированного тестирования интерпретируются с точки зрения полноты освоения области содержания, согласно детальной спецификации теста.

Ниже приведена таблица, содержащая сравнение характеристик обоих типов тестов.

Таблица 1.4.1. Сравнительные характеристики нормативно-ориентированных и критериально-ориентированных тестов по Д.Вилфорду[65]

Нормативно - ориентированные тесты	Критериально - ориентированные тесты
1. Цель тестирования: возможность сравнения уровня подготовки испытуемых друг с другом в той области содержания, для которой тест предназначен. Пример использования: конкурсный отбор кандидатов на обучение.	1. Цель тестирования: возможность аттестации испытуемого в соответствии с его уровнем усвоения определенной области содержания. Пример использования: итоговая аттестация уровня обученности студентов, уровня профессиональной подготовки кадров.
2. Используемые шкалы: нормативные (или стандартные) шкалы. Необходимо указание среднего значения и стандартного отклонения в выбранной шкале.	2. Используемая шкала - в основном, шкала процентов с выбранным одним (или несколькими) критериальным баллом (баллами). Особое внимание уделяется методики оптимального выбора критериального балла (или баллов).
3. Распределение индивидуальных баллов: близко к нормальному, в большинстве случаев имеет симметричный вид (рис.1.4.1)	3. Распределение индивидуальных баллов: произвольное, в большинстве случаев асимметричное и имеет несимметричный вид (рис.1.4.2)
4. Уровень детализации области содержания - несущественен. Авторы теста выбирают наиболее значимые элементы содержания.	4. Уровень детализации области содержания - подробный. Авторы теста разрабатывают спецификацию (план) теста, включающую все элементы содержания. Затем по этой спецификации разрабатываются задания.
5. Нормативная группа испытуемых обязательна. Обработанные (или шкалированные) баллы по результатам нормативно-ориентированного тестирования базируются на статистических данных нормативной группы, то есть специфической достаточно большой выборке испытуемых. В большинстве случаев применяются специальные нормативные таблицы, где каждый индивидуальный балл для данного теста имеет однозначное соответствие с процентильным эквивалентом, определенным на нормативной группе.	5. Нормативная группа испытуемых не является необходимой. Индивидуальный балл испытуемого интерпретируется по отношению к доле учебного материала успешно им освоенного. Чаще всего балл студента отражает процент правильно выполненных заданий и выражается шкале процентов.

6. Статистический анализ и отбор тестовых заданий. Статистические показатели тестовых заданий (в основном это уровень трудности и различающая способность) играют важную роль в отборе заданий. Выбираются задания со средним уровнем трудности (от 0,3 до 0,7) и высокой различающей способностью (большей 0,3). Существуют ряд других важных статистических показателей качества заданий.	6. Статистический анализ и отбор тестовых заданий. Уровень трудности и различающая способность заданий не является существенными факторами включения в состав теста, или наоборот исключения из него. Главное условие отбора заданий - это их соответствие (их конгруэнтность) спецификации и элементу содержания. Статистические характеристики тестовых заданий используются для составления параллельных форм (вариантов) теста и для выбора оптимального критериального балла.
7. Надежность теста. Оценивается либо путем нахождения корреляциимежду результатами двух тестирований, либо методом расщепления теста на две половины при однократном тестировании.	7. Надежность теста. Оценивается степенью постоянства принятия решения «зачет – незачет» при двукратном тестировании.
8. Валидность. Наряду с содержательной валидностью для тестов конкурсного отбора учащихся особое внимание уделяется высоким показателям прогностической валидности.	8. Валидность. Особое внимание уделяется содержательной валидности. В случае принятия важных решений по результатам тестирования исследуются критериальная и конструктная валидность.

Подытоживая, отметим, что нормативно-ориентированные и критериально-ориентированные тесты сильно отличаются друг от друга. При использовании тестов необходимо придерживаться следующих правил:

1) нельзя использовать критериально-ориентированный тест в качестве нормативно-ориентированного и наоборот;

2) нельзя использовать один и тот же тест и в качестве нормативно-ориентированного и в качестве критериально-ориентированного.

Нарушение этих правил приводит к получению тестовых результатов, обладающих низкой надежностью и большой ошибкой измерения.

1.5. НАДЕЖНОСТЬ И ВАЛИДНОСТЬ ТЕСТА

Под надежностью, или релиабильностью, измерения понимается степень надежности, или точности, с какой может быть измерен тот или иной конкретный признак[19]. Надежность теста характеризует воспроизводимость его результатов. Отметим, что определяя надежность теста, следует иметь в виду, что измерение не может быть стабильнее измеряемой латентной переменной. Если переменная очень лабильна, то ее измерение в принципе не может характеризоваться высокой повторяемостью.

Научно обоснованный тест - это метод, соответствующий установленным стандартам надежности и валидности[26]. Если тест имеет низкие надежность и валидность, то использовать его нельзя.

Надежность характеризуется коэффициентом надежности. Коэффициент надежности, это корреляционный коэффициент, показывающий степень совпадения результатов тестирования осуществленного в одинаковых условиях одним и тем же тестом.

Другая важнейшая характеристика теста – валидность. Валидность характеризует пригодность теста для измерения определенной величины. Следует отметить, что нельзя говорить о валидности теста, не указав условий его применения[6].

Можно привести такой наглядный пример. Два стрелка стреляют по мишени. Первый набрал 60 очков, а второй 90 из 100. Какой стрелок лучше? На первый взгляд кажется, что второй. Но при уточнении условий задачи оказалось, что второй стрелок поразил чужую мишень. Поэтому, несмотря на высокую надежность стрельбы, второй стрелок является «не валидным», он не может достигнуть цели, которая перед ним ставилась. Ясно, что первый стрелок предпочтительнее.

Тест может иметь высокую надежность, но низкую валидность. Тест с высокой валидностью обязательно имеет высокую надежность. Если тест имеет низкую валидность, то применять его нельзя, даже если он имеет высокую надежность.

Понятия надежности и валидности педагогического теста чрезвычайно важны, поскольку именно они характеризуют тест как измерительный инструмент. Тест с неизвестными надежностью и валидностью непригоден для измерения. Когда преподаватель, разработав тест, проводит тестирование, то полученные результаты следует интерпретировать (например, для ранжирования испытуемых) очень осторожно, так как неизвестны надежность и валидность вновь составленного теста. Эти, крайне важные понятия более подробно будут рассмотрены в третьей главе.

1.6. ИЗМЕРИТЕЛЬНЫЕ ШКАЛЫ

Поскольку мы рассматриваем тест как средство педагогического измерения, то, как и во всех измерениях, нам следует сначала рассмотреть вопрос об измерительных шкалах. В измерениях используется много различных шкал, мы рассмотрим четыре основные шкалы[69,70,71,72].

Согласно Пфанцаглю И.[69], шкала задается группой допустимых преобразований. Номинальная шкала (шкала наименований) задается группой всех взаимнооднозначных преобразований, шкала порядка - группой всех строго возрастающих преобразований.

ШКАЛА НАИМЕНОВАНИЙ используется для идентификации элементов множества. На этой шкале определены две операции - «равно» и «не равно». Номинальная шкала допускает те преобразования, которые, у одинаковых объектов оставляет одинаковые имена (идентификаторы). Это могут быть имена собственные, названия городов и т.д. Рассмотрим пример трех множеств из пяти элементов. Первое множество образуют фамилии людей, второе - знаки зодиака, третье - номера комнат. Элементы этих множеств приведены в таблице.

Элементы номинальной шкалы.

№	Множество 1 «фамилии»	Множество - 2 «знаки зодиака»	Множество - 3 «номера комнат»
1	Иванов	♋	27
2	Сидоров	♑	81
3	Петров	♍	108
4	Алексеев	♐	312
5	Яковлев	♌	105

Значения на номинальной шкале всего лишь дают возможность отличить один объект от другого. Эти значения не могут быть упорядочены и рассматриваются изолированно друг от друга.

Специально отметим, что числа, приведенные в последнем столбце (Множество - 3), числами не являются. Это «имена» комнат. С ними нельзя, например, выполнить действие сложения: 27+81=108. Тем более, на номинальной шкале нельзя выполнять арифметические операции умножения и деления.

ШКАЛА ПОРЯДКА, как и шкала наименований, является качественной, но позволяет не только именовать, но и ранжировать элементы множества. Порядковая шкала допускает только монотонные преобразования, то есть такие, которые не нарушают порядок следования значений измеряемых величин. Самый яркий пример порядковой шкалы - это шкала Мооса для твердости минералов.

Минерал	Твердость по Моосу
Тальк	1
Гипс	2
Кальцит	3
Флюорит	4
Апатит	5
Ортоклаз	6
Кварц	7
Топаз	8
Корунд	9
Алмаз	10

При построении шкалы твердости рассуждали следующим образом: тальк - самый мягкий минерал, им ничего нельзя поцарапать, поэтому ему присвоена самая низкая твердость. Гипс царапает тальк, следовательно, он тверже и ему присваивается твердость, равная двум. В свою очередь, кальцит царапает гипс, значит, он еще тверже и ему приписывается твердость 3. Самым твердым оказывается алмаз, который царапает все минералы и ни один минерал не царапает его.

Отличительной особенностью порядковой шкалы является то, что значения по этой шкале упорядочены. В рассмотренном примере минералы строго упорядочены по своей твердости. Пусть мы хотим определить твердость неизвестного минерала. Проведем серию испытаний, пытаясь поцарапать известные минералы. Допустим, оказалось, что мы можем поцарапать кварц, но не можем корунд. Значит наш минерал тверже кварца, но мягче корунда. Следовательно, твердость нашего минерала равна 8. Отметим, что мы не знаем *насколько* наш минерал тверже кварца, такую информацию порядковая шкала не содержит.

Другой пример - это школьные отметки.

Уровень знаний	Отметка
Совершенно неудовлетворительно	1
Неудовлетворительно	2
Удовлетворительно	3
Хорошо	4
Отлично	5

Отметки имеют свои имена (1, 2, 3, 4, 5) и упорядочены. Нам известно, что 4 означает более высокий уровень знаний, чем 3, но не известно насколько. С отметками нельзя выполнять арифметические операции: 5-4=1, 3-2=1, 5-4=3-2 и т.д.. Ясно, что различие в знаниях между отличником и хорошистом не такое же, как между троечником и двоечником. Это общеизвестный факт. С другой стороны в образовательных учреждениях широко практикуется средний балл. Для определения среднего балла складывают, например, все отметки за год и делят на их количество. Это недопустимо. Ни складывать, ни делить отметки нельзя, так как они расположены на порядковой шкале[*].

ШКАЛА ИНТЕРВАЛОВ, в отличие от шкалы порядка, позволяет не только ранжировать элементы множества, но и задает известные интервалы между элементами. Интервальная шкала допускает линейные преобразования вида:

$$y = a \cdot x + b$$

где a - положительное число, b - положительное или отрицательное число.

Изменение a приводит к изменению масштаба шкалы, изменение b вызывает сдвиг по шкале, то есть положение нуля на интервальной шкале не определено. Интервальные шкалы используются, например, для измерения температуры. При этом температурные интервалы равны, а положение нуля зависит от вида температурной шкалы, например по Цельсию, или по Фаренгейту.

[*] Говорят, Лев Ландау ввел 10-ти балльную шкалу для оценивания женской красоты. Если это так, то его шкала должна была быть порядковой.

Если это неизвестно, то для описания закономерностей следует использовать отношение интервалов:

$$\frac{y_1 - y_2}{y_3 - y_4} = \frac{(ax_1 + b) - (ax_2 + b)}{(ax_3 + b) - (ax_4 + b)}$$

ШКАЛА ОТНОШЕНИЙ допускает линейные преобразования вида:

$$y = a \cdot x$$

Шкала отношений, в отличие от интервальной шкалы, обладает точкой нулевого отсчета. Этот тип шкал используется для измерения массы тела, его длины и так далее. Например, длина может измеряться в метрах, футах, парсеках - это определяется масштабным множителем $\boldsymbol{a}$. Если нам неизвестны единицы измерения, то для описания закономерностей следует использовать *отношение* величин, которое является инвариантом для шкалы отношений.

Литература к главе 1

[1] Талызина Н.Ф. Управление процессом усвоения знаний. -М.: МГУ, 1975. -343 с.

[2] Беспалько В.П. Программированное обучение. Дидактические основы. –М., 1970. -300 с.

[3] . Ингенкамп К. Педагогическая диагностика. -М.: Педагогика, 1991. -240 с.

[4] Талызина Н.Ф. Формирование познавательной деятельности младших школьников. –М: Просвещение, 1988. -175 с.

[5] Кадневский В.М. История тестов. Монография. –М.: Народное образование, 2004. -464 с.

[6] Аванесов В.С. Тесты: история и теория // Управление школой, 1999, №12.

[7] Майоров А.Н. – Теория и практика создания тестов для системы образования. – М.: «Интеллект-центр», 2001. -296 с.

[8] Мейман Э. Лекции по экспериментальной педагогике. Ч.2. -М., 1917. -С.163-198.

[9] Spearman C. Correlation calculated from faulty data //British Journal of Psychology, 1910, Vol.3, N2. -P.271-295.

[10] Gulliksen H. Theory of Mental Tests. -New-York, Wiley, 1950. -486 p.

[11] Guttman, L. A special review of Harold Gulliksen, Theory of mental tests. Psychometrika, 1953, 18, 123-130.

[12] Lord F.M. & Novick M. Statistical Theories of Mental Test Scores. -Addison-Wesley Publ. Co. Reading, Mass. 1968. -560 p.

[13] Kuder G.F., Richardson M.W. The Theory of the estimation of test reliability // Psychometrika, 1937, Vol.2, N3.

[14] Crocker Linda, Algina James. Introduction to Classical and Modern Test Theory. –New-York: Harcourt Brace Jovanovich, 1986.

[15] Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen, 1960, Danish Institute of Educational Research. (Expanded edition, Chicago, 1980, The University of Chicago Press).

[16] Birnbaum A. Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In F.M. Lord and M.R.Novick. Statistical Theories of Mental Test Scores. Readinf Mass.: Addison-Wesly, 1968. Ch.17-20. -p.397-479.

[17] Andrich, D., Sheridan, B., Lyne, A. & Luo, G. RUMM: A windows-based item analysis program employing Rasch unidimensional measurement models (Perth: Murdoch University), 2000.

[18] Wright B.D. & Stone M.H. Best Test Design. -Chicago, MESA PRESS, 1979. -222 p.

[19] Анастази А., Урбина С. Психологическое тестирование. –Спб.: Питер, 2006. -688 с.

[20] Равен Джон. Педагогическое тестирование: Проблемы, заблуждения, перспективы / Пер. с англ. -М.: «Когито-Центр», 1999. -144с.

[21] Беспалько В.П. Программированное обучение. Дидактические основы. –М., 1970. -300 с.

[22] Беспалько В.П., Татур Ю.Г. Системно-методическое обеспечение учебно-воспитательного процесса подготовки специалистов: Учеб-метод.пособие. –М., 1989. -144 с.

[23] Талызина Н.Ф. Управление процессом усвоения знаний. –М.:МГУ, 1975. – 343 с.

[24] Талызина Н.Ф. Теоретические проблемы программированного обучения. –М.:МГУ, 1969. - 134 с.

[25] Талызина Н.Ф. Теоретические основы контроля в учебном процессе. -М.: Знание, 1983.

[26] Аванесов В.С. Основы научной организации педагогического контроля в высшей школе. -М., 1989. -167 с.

[27] Челышкова М.Б. Теория и практика конструирования педагогических тестов: Учебное пособие. –М.: Логос, 2002. -432 с.

[28] Родионов Б.У., Татур А.О. Стандарты и тесты в образовании. -М.: МИФИ, 1995.

[29] Майоров А.Н. Тесты школьных достижений: конструирование, проведение, использование. Издание второе - СПб.:Образование и культура,1997.-304с.

[30] Нейман Ю.М., Хлебников В.А. Введение в теорию моделирования и параметризации педагогических тестов. -М., 2000. - 168 с.

[31] Нейман Ю.М., Хлебников В.А. Педагогическое тестирование как измерение. – Москва, Центр тестирования МО РФ, 2002. - 67 с.

[32] Нейман Ю.М. Об оценивании уровня подготовленности учащихся по результатам единого государственного экзамена. -М.: Poligraph, 2002. -30 с.

[33] Нардюжев В.И., Нардюжев И.В. Модели и алгоритмы информационно-вычислительной системы компьютерного тестирования Монография –М.: Прометей, 2000 -148 с

[34] Михайлычев Е.А. Дидактическая тестология. -М.: Народное образование, 2001. -432 с.

[35] Маслак А.А. Измерение латентных переменных в социально-экономических системах: Монография. -Славянск-на-Кубани: Изд.центр СГПИ, 2006, -333 с.

[36] Переверзев В.Ю. Критериально-ориентированные педагогические тесты для итоговой аттестации студентов. -М.: НМЦ СПО Минобразования РФ, 1999. - 152 с.
[37] Переверзев В.Ю. Технология разработки тестовых заданий: справочное руководство. –М.: Е-Медиа, 2005. -265 с.
[38] Войтов А.Г. Учебное тестирование для гуманитарных и экономических специальностей: Теория и практика. -2-е изд., перераб. –М.: Издательско-торговая корпорация «Дашков и К», 2005. -402 с.
[39] Морев И. А. Образовательные информационные технологии. Часть 2.Педагогические измерения: Учебное пособие. – Владивосток: Изд-во Дальневост. ун-та, 2004. -174 с.
[40] Морев И.А. Образовательные информационные технологии. Часть 5. Методическая система стимулирования обучаемости средствами дидактического тестирования. Монография. – Владивосток: Изд-во Дальневост. ун-та, 2004. -120 с.
[41] Морев И.А. Образовательные информационные технологии. Часть 4. Развивающий измерительный процесс в вузе. Монография. – Владивосток: Изд-во Дальневост. ун-та, 2004. -148 с.
[42] Кузовлева К.Т., Яхонтов С.В., Сиськов В.В. Программа для обработки тестовых результатов в рамках классической теории тестов и Item Response Theory (IRT) // Тез.докладов Всерос. конф. «Развитие системы тестирования в России», 25-26 ноября 1999 г., ч.3, -М., 1999. –С.66-67.
[43] Кречетников К.Г. Проектирование креативной образовательной образовательной среды на основе информационных технологий в вузе. Моногр. -М.: Изд-во Госкоорцентр, 2002. -296 с.
[44] Кречетников К.Г. "Задания в тестовой форме и методика их разработки" (Уч.-метод. пособие). - Владивосток: ДВГУ, 2002. – 40 с.
[45] Ким В.С. Компьютерная поддержка дисциплины “Общая электротехника” // Новые информационные технологии в педагогическом образовании. – тезисы докладов XII Республиканская научно-практическая конференция, 24-26 апреля, 1995, Магнитогорск -Магнитогорск, изд-во МГПИ, 1995, С.81-82.
[46] Ким В.С. Формирователь бинарных и корреляционных матриц // Тезисы межвузовской научно-методической конференции “Наука и учебный процесс” 17-19 декабря,1996, Ч.2, -Владивосток, 1996. - С.95-96.

[47] Ким В.С. Анализ результатов тестирования в процессе Rasch measurement //Педагогические измерения, N4, 2005. –С.39-45.
[48] Ким В.С. Развивающая функция тестовых заданий // Педагогические измерения, 2007, № 1. -С.77-84.
[49] Ким В.С. Компьютерное тестирование, как элемент управления учебным процессом // Вестник МГОУ. Серия "Педагогика", 2007, том 2. -С. 94-98.
[50] Фалалеева О.Н. Оценивание учебных достижений методом мягкого тестирования. Вестн. МГОУ. Серия "Открытое образование". - 2(33). Том 2. - 2006. - М.: Изд-во МГОУ. - С. 126-130.
[51] Аванесов В.С. От редактора //Педагогические измерения, 2004, №1. –С.3-7.
[52] Термины Единого Государственного экзамена.. http://www.ege.ru/dict/dict2.htm.
[53] Рубинштейн С.Л. К критике метода тестов / Против педологических извращений в педагогике. Л., 1938.
[54] Аванесов В.С. Форма тестовых заданий. -М.: Центр тестирования, 2005. -156 с.
[55] Bloom B.S., Hasting J.T., & Madaus G.F. Handbook on Formative and Summative Evaluation of Student Learning. New-York: McGraw-Hill, 1971. -923 p.
[56] Беспалько В.П. Слагаемые педагогической технологии. -М., 1989.
[57] Качество знаний учащихся и пути его совершенствования /Под ред. М.Н.Скаткина, В.В.Краевского. -М.: Педагогика, 1978. -208 с.
[58] Симонов В.П. Педагогический менеджмент: Учебное пособие. -М.: РПА, 1997. -160 с.
[59] Кларин М.В. Инновационные модели обучения в зарубежных педагогических поисках. -М.: Арена, 1994. - 223 с.
[60] Роберт Ван Криген, Стивен Баккер. Подготовка и проведение экзаменов. Руководство для организации и разработки централизованных экзаменов. CITO, Национальный институт по оценке достижений в области образования. -Амхем, Нидерланды, 1995.
[61] Hambleton R.K. Aplication of Item Response Theory. -Vancouver: Educ.Res. Inst. B.C., 1983.
[62] Черненко В.В., Котенкова Н.А., Лобанова И.В. Пряженникова О.А. О механизме возникновения систематической погрешности при тестировании уровня интеллектуальных способностей // Мат. Всерос. НТК посвященной 300-летию военного, военно-морского и высшего профессионального образования в России. Т.1. Военно-

исторические,военно-педагогические, гуманитарные и социально-экономические вопросы. ТОВМИ им С.О. Макарова. - Владивосток, 2000. -С. 156-158.

[63] Колдаева В.Б., Колдаев В.М. К вопросу о планировании занятий по теоретическим дисциплинам // Гуманитарные и социально-экономические аспекты обучения и воспитания кадров ВМФ. Сб.научных статей. Вып.3. – Владивосток, ТОВМИ, 2000.

[64] Буравлев А.И., Переверзев В.Ю. Выбор оптимальной длины педагогического теста и оценка надежности его результатов. http://www.e-joe.ru/sod/99/2_99/st160.html.

[65] Вилфорд Д. Современная типология педагогических тестов. Информационный бюллетень «Тесты в образовании», 1999, вып.1.

[66] Berk R.A. Criterion-referenced measurement: The state of art. Baltimor, MD: Johns Hopkins University Press, 1980.

[67] Keeves J.P. (Ed.) Educational Research, Metodology and Measurement: An International Handbook. - Oxford: Pergamon press, 1988.

[68] Gronlund N.E. How To Construct Achievement Test. -N.J.: Prentice Hall, 1998.

[69] Пфанцагль И. Теория измерений. - М.:Мир, 1976. - 165 с.

[70] Суппес П., Зинес Дж. Психологические измерения. -М.: Мир, 1967.

[71] Загоруйко Н.Г., Савельев Л.Я. Относительная мощность измерительных шкал. //Структурный анализ символьных последовательностей (Вычислительные системы, вып.101). - Новосибирск, 1984. -С.111-129.

[72] Орлов А.И. Статистика объектов нечисловой природы (Обзор). – Журнал «Заводская лаборатория». 1990. Т.56. No.3. С.76-83.

ГЛАВА 2. ФОРМЫ ТЕСТОВЫХ ЗАДАНИЙ

Тестовые задания имеют специфическую форму, что отражено даже в определении тестового задания. Задание, имеющее правильную форму, позволяет точно выразить содержание, понятно всем испытуемым, исключает возможность появления ошибочных ответов по формальным признакам[1,2].

2.1. ФОРМА ТЕСТОВЫХ ЗАДАНИЙ

Согласно В.С.Аванесову ФОРМА ТЕСТОВЫХ ЗАДАНИЙ – это способ организации, упорядочения и существования содержания теста. Соединившись с содержанием, форма придает заданию конкретный облик, или иначе, содержание принимает определенную форму. Форма может рассматриваться как инвариант. В тестах по разным учебным дисциплинам может использоваться одна и та же форма заданий[2].

Умение правильно, ясно и лаконично оформлять задания есть необходимое, но недостаточное условие для создания теста. Можно в совершенстве овладеть искусством правильного оформления заданий, но так и не получить полноценного теста. Это обусловлено тем, что даже хорошо сформулированное задание еще не является тестовым. Необходима еще достаточно трудоемкая процедура эмпирической проверки задания, статистическая обработка результатов его применения. Иногда возможно чисто визуально, на экспертном уровне, определить будет ли задание тестовым. Однако нередко бывает, что задание в тестовой форме не выдерживает статистической проверки, то есть не становится тестовым заданием.

Почему экспертный анализ не позволяет еще на этапе конструирования задания выявить его несоответствие статистическим требованиям? Возможно, это обусловлено трудно предсказуемым результатом восприятия вербальной и невербальной информации таким сложным элементом социальных систем как человек.

На практике почти невозможно определить, почему прекрасно оформленное задание не проходит статистическую проверку. Рациональный путь заключается в выбраковке такого задания, без малопродуктивных попыток умозрительного анализа его несостоятельности. Задание переформулируют и заново подвергают статистической проверке и, возможно, не один раз.

Если задания неудачно сформулированы, допускают многозначное толкование своего содержания, то результаты тестирования будут искажены. Индивидуальные баллы испытуемых не будут соответствовать действительности, так как будут занижены вследствие неверного понимания испытуемыми содержания задания. В результате будет получена большая ошибка измерения и тем больше, чем сильнее отклонение от правильной формы. Если обратиться к Дж.Кеттелу, то речь идет о соответствии задания первому и пятому требованиям (см. главу 1).

Мало того, если даже нет явных ошибок в содержательной части задания, сама форма представления его содержания может сказаться, например, на восприятии, на трудности задания.

Рассмотрим теперь, примеры формулировок задания.

Пример № 1.

К ДВУПЛЕЧЕМУ РЫЧАГУ ПОДВЕШЕНЫ ДВА ГРУЗА МАССОЙ m_1 = 400 г, и m_2 = 200 г. ПЛЕЧО l_1 = 10 см, l_2 = 20 см. ЭТОТ РЫЧАГ В РАВНОВЕСИИ

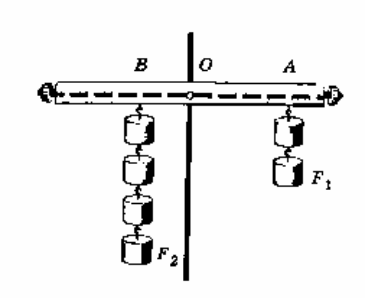
1) находится
2) не находится

Рисунок для примера взят из учебника А.В.Перышкина[3].

Пример № 2.

РЫЧАГ В РАВНОВЕСИИ

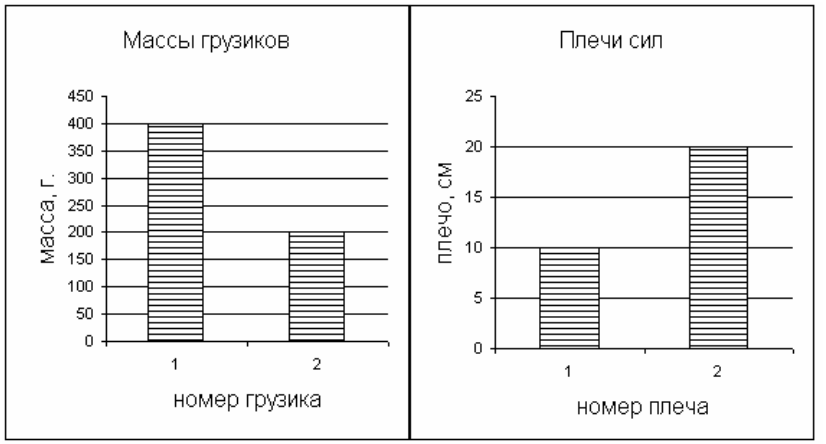
1) находится
2) не находится



Пример № 3.

НА ДИАГРАММАХ ПОКАЗАНЫ МАССЫ ГРУЗИКОВ, ПОДВЕШЕННЫХ К РЫЧАГУ И ПЛЕЧИ СИЛ, ДЕЙСТВУЮЩИХ НА РЫЧАГ. ЭТОТ РЫЧАГ В РАВНОВЕСИИ

1) находится
2) не находится



Сравнивая между собой приведенные примеры заданий, можно прийти к выводу, что восприятие содержания задания во всех случаях совершенно разное. Наиболее предпочтительным представляется задание, сформулированное в примере №2. Цель задания заключалась в том, чтобы выяснить - знает ли испытуемый условие равновесия рычага. Во всех трех примерах эта цель достигается, но в примере №2 это достигается наиболее экономным способом. В примере №2 испытуемый выполняет минимальные умственные действия, чтобы понять задание (условие задачи). Поэтому большую часть усилий он направит на собственно решение задачи.

2.2. КЛАССИФИКАЦИЯ ТЕСТОВЫХ ЗАДАНИЙ

Формы заданий в тестовой форме могут быть весьма разнооб4разными, в частности, Распопов В.М.[4] предлагает 24 формы тестовых заданий. В этой связи классификация тестовых заданий представляется полезной.

С точки зрения формы тестового задания можно ввести следующую их классификацию (рис.2.2.1).

Рис.2.2.1. Классификация тестовых заданий.

Все задания разбиваются на две большие группы - задания в открытой форме и задания в закрытой форме. В основу классификации положено наличие или отсутствие ввода дополнительной информации испытуемым. Если дополнительная информация нужна, то это задание в открытой форме. Если не нужна, то это задание в закрытой форме.

Задания в открытой форме подразделяются на задания с дополнением и в виде свободного изложения. В первом случае испытуемому необходимо дополнить содержание задания своей информацией. В результате задание должно превратиться в истинное логическое высказывание. Дополнение должно быть кратким - одно, в крайнем случае, два - три слова. При свободном изложении объем водимой информации может быть значительно больше.

В тестировании чаще всего используются задания в закрытой форме. Эти задания характерны тем, что содержат в себе и основу (вопрос, утверждение) и ответы (элементы ответов), из которых испытуемый должен выбрать или составить верный ответ.

В простейшем случае испытуемый просто указывает ответ, который ему кажется правильным - задания с выбором верного ответа. Об этих заданиях более подробно будет сказано далее.

В заданиях с выбором нескольких верных ответов, испытуемый должен указать все верные ответы. Процедура оценивания здесь сложнее, чем в предыдущем случае. Сумма баллов за такое задание может быть больше чем в заданиях с выбором одного верного ответа.

Задания с градуированными ответами содержат ответы, которые возможно все являются правильными в той или иной степени. Ответы имеют градацию по степени правильности. Задача составителя заключается в том, чтобы найти и применить признак, позволяющий осуществить такую градацию. Максимальное количество баллов испытуемый получает, если его градация ответов полностью совпадает с градацией эксперта, например, разработчика задания.

Задания на установление соответствия требуют от испытуемого найти соответствие между элементами двух множеств. Соответствие устанавливается на основании логических умозаключений или использовании смысловых ассоциаций.

В заданиях на установление правильной последовательности испытуемому необходимо не просто выбрать соответствующие элементы ответа, но и расположить их в нужной последовательности. Заданиями такого типа хорошо проверять знание алгоритмов действий, технологических приемов, логики рассуждений и т.п. С помощью этих заданий удобно проверять знание и понимание испытуемыми формулировок определений, понятий, терминов, путем конструирования их из отдельным слов, предложений, символов, графических элементов.

2.3. ЗАДАНИЯ С ВЫБОРОМ ОДНОГО ВЕРНОГО ОТВЕТА

Задания с выбором одного верного ответа (далее задания с выбором) широко используются в практике тестирования. Другое название этой формы - задания в закрытой форме.

В заданиях этого типа испытуемому предлагается несколько вариантов ответа, среди которых один верный, а остальные не верные. Неверные ответы называются дистракторами (distractor)[5]. Подбор хороших, правдоподобных дистракторов представляет собой непростую задачу. Если дистракторы подобраны неудачно, то они перестают работать и тогда задание, например с тремя ответами, превращается в задание с двумя ответами. Дистракторы должны отвечать принципу равной привлекательности и достаточно высокой правдоподобности. Считается, что каждый дистрактор должен выбираться не менее чем 5 процентами испытуемых.

Приведем рекомендации М.Б.Челышковой по созданию дистракторов. Первый метод - предъявление ученикам неоконченного списка вариантов выбора и последующее использование неправильных ответов, предложенных учениками. Второй метод - предъявление группе испытуемых заданий в открытой форме и последующий анализ типичных ошибок учеников в составленных ими ответах. Оба метода дают исходную информацию о типичных ошибках испытуемых, на основании которой можно создать весьма правдоподобные, с точки зрения испытуемых, дистракторы.

Характерной особенностью заданий с выбором является то, что испытуемый выполняет задание только выбором одного из ответов. При этом нет необходимости вписывать верный ответ или его фрагмент в бланк задания. Эта особенность является очень важным достоинством заданий с выбором, поскольку здесь достигается однозначное соответствие замысла разработчика задания и понимание задания испытуемым (в заданиях открытого типа это не так) и эта же особенность порождает и, часто критикуемые, недостатки заданий с выбором. Другим достоинством заданий с выбором, является их технологичность. Задания относительно легко оформляются, результаты выполнения фиксируются довольно просто и однозначно. Далее, отметим, что обработка результатов тестирования легко выполняется на компьютере, а при использовании некоторых приемов (особые бланки с копировальной бумагой)[6], и вручную.

Наряду с достоинствами, задания с выбором обладают и недостатками. Задания с выбором одного верного ответа подвергаются критике по следующим причинам.

1) Испытуемому демонстрируют неверные ответы, которые он может запомнить. В этом случае благодаря проявлению действия обучающей функции теста, испытуемый закрепляет в своем сознании неверные ответы как верные. Происходит искажение, как содержания, так и структуры знаний испытуемого. Эти рассуждения не лишены оснований, но, к настоящему моменту, экспериментальных доказательств их истинности нет.

2) Испытуемому предоставляется возможность случайным или целенаправленным образом, угадать верный ответ.

Происходит завышение индивидуального балла испытуемого (количеств верных ответов испытуемого по тесту в целом). Возникает ошибка измерения, которая тем меньше, чем больше дистракторов в задании.

С этим явлением можно с тем или иным успехом бороться, вводя поправку на угадывание. Вопросы расчета поправок на угадывание рассмотрены в главе 4. Отметим, что с педагогической точки зрения необходимо учитывать мотивацию испытуемых к угадыванию при расчете поправок[7].

Как быть, если испытуемый не знает верного ответа, но не желает угадывать его, считая это неэтичным? В этом случае желательно предоставить испытуемому возможность отказа от ответа[8]. Технически это можно сделать, добавив к дистракторам еще один вариант ответа, например «не знаю». Можно также никак не отмечать номер выбранного ответа, например не обводить кружочком. В компьютерном тестировании необходимо предусмотреть возможность пропуска тестового задания и перехода к следующему.

Чтобы мотивировать испытуемых к отказу от угадывания, можно начислять баллы согласно той или иной специальной процедуре. Например, за верный ответ ставить +1 балл, за неверный ответ -1 балл, а за отказ от ответа 0 баллов.

В целом, учитывая все достоинства и недостатки, следует признать, что тестовые задания с выбором одного верного ответа оправдывают свое назначение и вполне могут применяться в тестировании.

Минимальное количество вариантов ответа в задании равно двум, а максимальное, в принципе, не ограничено. На самом деле, это, конечно не так.

Во-первых, чисто технически очень непросто разработать большое количество правдоподобных дистракторов. Если в задание ввести большое количество слабых дистракторов, то проку от этого

будет мало - дистракторы не будут работать и, как указывалось выше, в действительности мы получим тестовое задание с гораздо меньшим числом ответов.

Во-вторых, поскольку время тестирования ограничено, то испытуемый может просто не успеть проанализировать всю совокупность дистракторов по всему тесту. Может получиться ситуация, когда испытуемый показал низкие результаты не потому, что имел низкий уровень знаний, а потому, что не имел физической возможности для обстоятельного анализа заданий.

Максимальное число вариантов ответов выбирается из вышеприведенных соображений. Рекомендуется не делать его больше пяти - шести[2]. Эти рекомендации не абсолютны, окончательное решение принимает разработчик тестового задания. В главе 4 будет показано, что в некоторых случаях, количество ответов может быть значительно больше пяти.

Под правильным, верным ответом не всегда может подразумеваться некий эквивалент истинного знания. Это относится к заданиям с выбором неверного ответа. Е.А.Михайлычев[9] считает допустимым использование заданий с выбором единственного дистрактора. Рассмотрим пример такого задания. В нижеприведенных примерах знаком «+» помечен правильный ответ.

Обведите кружочком номер неверного ответа.

Пример № 4.

ВЕЩЕСТВО В ЖИДКОМ СОСТОЯНИИ
+1) сохраняет свою форму
2) имеет неизменный объем
3) обладает текучестью

Какова была цель этого задания? Требовалось выяснить, знает ли испытуемый, что жидкости не сохраняют свою форму. При выполнении этого задания, испытуемому сначала надо понять его, привлечь имеющиеся знания и выбрать верный ответ. Их два – 2-й и 3-й. Затем, испытуемый вспоминает, что от него требуется указать неверный ответ и выбирает ответ номер 1. Эти дополнительные усилия по переключению внимания с верных ответов на неверный ответ, ничего не дают с точки зрения достижения цели задания.

Если испытуемый знает, что жидкости не сохраняют свою форму, но упустил, что надо указать неверный ответ, то формально он

с этим заданием не справился. Получается, что задание измеряет не только подготовленность испытуемого, но и еще какие-то другие латентные свойства, например, степень сосредоточенности внимания на выполняемой работе. Это является недостатком задания. Попытаемся переформулировать его.

Обведите кружочком номер верного ответа.

Пример № 5.

ВЕЩЕСТВО В ЖИДКОМ СОСТОЯНИИ
+1) может менять свою форму
2) имеет переменный объем
3) не обладает текучестью

Теперь от испытуемого не требуется дополнительных действий для выполнения задания. Цель задания достигается более экономным способом. А.Майоров считает, что отрицания все же можно использовать, если утвердительный вопрос вызывает слишком много правильных ответов. Для того, чтобы испытуемый обратил внимание на отрицание, необходимо выделять их, используя курсив, жирный шрифт или подчеркивание.

2.4. СТРУКТУРА ЗАДАНИЯ В ТЕСТОВОЙ ФОРМЕ

Рассмотрим структуру задания в тестовой форме.

Задание с выбором состоит из трех блоков:

1) блока инструкции;

2) основного, содержательного блока, сформулированного в виде утверждения или вопроса;

3) блока вариантов ответов

На бланке эти элементы обычно размещаются именно в таком порядке.

В блоке инструкции описывается, как следует выполнять задание, например, «обвести кружочком номер выбранного ответа». Рассмотрим подробнее вопрос о размещении инструкции.

Если тест содержит однотипные задания, то допустимо инструкцию указывать в начале, для всего теста, а не для каждого задания. В тех случаях, когда испытуемый знает, как надо выполнять задание, инструкция вообще не нужна. Обучение испытуемого правилам выполнения задания можно провести посредством тренировочного тестирования. Особенно это касается компьютерного тестирования. Благодаря тренировочному тесту, испытуемые осваивают интерфейс программы-тестера, учатся выбирать ответ с помощью мыши или клавиатуры и избавляются от стресса, вызванного непривычностью процедуры тестирования. В дальнейшем, после запуска основного теста, испытуемые уже не испытывают каких-либо процедурных затруднений при выполнении заданий.

Удаление текста ненужной инструкции из бланка задания повышает его удобочитаемость. Лишняя, ненужная информация не отвлекает испытуемого от самого задания, что способствует концентрации его внимания на содержательном блоке.

Таким образом, в некоторых случаях, блок инструкции в задании может отсутствовать.

Перейдем теперь к описанию оформления задания.

В.С.Аванесов придерживается правила: основной, содержательный блок выводится прописными буквами, а блок вариантов ответов - строчными. Варианты ответов помечаются числом и закрывающей скобкой. Номер тестового задания помечается числом и точкой.

Обвести кружочком номер выбранного ответа.

Пример № 6.

ПЛОВЦА, ПЕРЕПЛЫВШЕГО РЕКУ ШИРИНОЙ 3 км, СНЕСЛО ПО ТЕЧЕНИЮ НА 4 км. МОДУЛЬ ЕГО ПЕРЕМЕЩЕНИЯ РАВЕН

1) 3 км
2) 4 км
+3) 5 км
4) 7 км

Одно из требований к оформлению заданий с выбором заключается в том, что все повторяющиеся фрагменты должны быть перенесены из блока ответов в содержательный блок.

В нашем примере в ответах повторяется слово «км». Перенесем его в содержательный блок.

Пример № 7.

ПЛОВЦА, ПЕРЕПЛЫВШЕГО РЕКУ ШИРИНОЙ 3 км, СНЕСЛО ПО ТЕЧЕНИЮ НА 4 км. МОДУЛЬ ЕГО ПЕРЕМЕЩЕНИЯ В КИЛОМЕТРАХ РАВЕН

1) 3
2) 4
+3) 5
4) 7

Обратим внимание на следующий факт - все ответы являются числами. Принцип указания варианта ответа также с помощью чисел приводит к плохой читаемости ответов и может вызвать ошибку испытуемого. Испытуемый может перепутать номер ответа с самим ответом. Чтобы избежать этого, лучше, в данном примере, помечать ответы буквами.

Пример № 8.

ПЛОВЦА, ПЕРЕПЛЫВШЕГО РЕКУ ШИРИНОЙ 3 км, СНЕСЛО ПО ТЕЧЕНИЮ НА 4 км. МОДУЛЬ ЕГО ПЕРЕМЕЩЕНИЯ В КИЛОМЕТРАХ РАВЕН

а) 3
б) 4
+в) 5
г) 7

Теперь обратимся к правилу оформления содержательного блока прописными буквами. В тех случаях, кода этот блок невелик, содержит 1-2 строки, такое правило позволяет сделать задание внешне очень выразительным, четко отделяя содержательный блок от блока ответом уже самим размером символов. Однако в случаях, когда содержательный блок содержит большее число строк, будучи оформленным прописными буквами, он приобретает громоздкий вид. По этой причине можно считать допустимым оформление содержательного блока строчными буквами.

Пример № 9.

Пловца, переплывшего реку шириной 3 км, снесло по течению на 4 км. Модуль его перемещения в километрах равен

а) 3
б) 4
+в) 5
г) 7

Дизайн задания ухудшился, поскольку содержательный блок и блок ответов плохо отличимы друг от друга, затруднилось восприятие задания.

Для устранения этого недостатка М.Челышкова[10] и В.Переверзев[11] используют выделение содержательного блока жирным начертанием шрифта.

Пример № 10.

Пловца, переплывшего реку шириной 3 км, снесло по течению на 4 км. Модуль его перемещения в километрах равен

а) 3
б) 4
+в) 5
г) 7

А.Майоров[12] использует словесное выделение блоков задания.

Пример № 11.

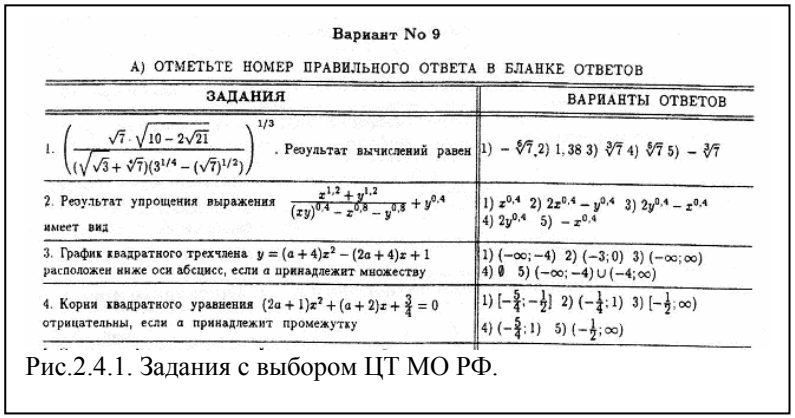
Вопрос: Пловца, переплывшего реку шириной 3 км, снесло по течению на 4 км. Модуль его перемещения в километрах равен

Варианты ответа:

а) 3
б) 4
+в) 5
г) 7

Недостатком этого способа является загромождение бланка задания словами «**Инструкция**», «**Вопрос**», «**Варианты ответов**», которые нужны не для уяснения сути задания, а только для построения его структуры. В примерах В.Аванесова и М.Челышковой цель структурирования задания достигается простым изменением шрифта, без введения дополнительной информации.

На рис.2.4.1 показан пример оформления заданий с выборов в бланках Федерального Центра тестирования Минобразования РФ. Здесь использован табличный способ структурирования задания.



Вариант No 9

А) ОТМЕТЬТЕ НОМЕР ПРАВИЛЬНОГО ОТВЕТА В БЛАНКЕ ОТВЕТОВ

ЗАДАНИЯ	ВАРИАНТЫ ОТВЕТОВ
1. $\left(\frac{\sqrt{7}\cdot\sqrt{10-2\sqrt{21}}}{(\sqrt{\sqrt{3}+\sqrt{7}})(3^{1/4}-(\sqrt{7})^{1/2})}\right)^{1/3}$. Результат вычислений равен	1) $-\sqrt[6]{7}$ 2) 1,38 3) $\sqrt[6]{7}$ 4) $\sqrt[3]{7}$ 5) $-\sqrt[3]{7}$
2. Результат упрощения выражения $\frac{x^{1,2}+y^{1,2}}{(xy)^{0,4}-x^{0,8}-y^{0,8}}+y^{0,4}$ имеет вид	1) $x^{0,4}$ 2) $2x^{0,4}-y^{0,4}$ 3) $2y^{0,4}-x^{0,4}$ 4) $2y^{0,4}$ 5) $-x^{0,4}$
3. График квадратного трехчлена $y=(a+4)x^2-(2a+4)x+1$ расположен ниже оси абсцисс, если a принадлежит множеству	1) $(-\infty;-4)$ 2) $(-3;0)$ 3) $(-\infty;\infty)$ 4) $\emptyset$ 5) $(-\infty;-4)\cup(-4;\infty)$
4. Корни квадратного уравнения $(2a+1)x^2+(a+2)x+\frac{3}{4}=0$ отрицательны, если a принадлежит промежутку	1) $[-\frac{5}{4};-\frac{1}{2}]$ 2) $(-\frac{1}{4};1)$ 3) $[-\frac{1}{2};\infty)$ 4) $(-\frac{5}{4};1)$ 5) $(-\frac{1}{2};\infty)$

Рис.2.4.1. Задания с выбором ЦТ МО РФ.

Замечания по этой форме такие же приведены выше (Пример № 11.).

Анализ приведенных примеров оформления заданий с выбором показывает, что наиболее предпочтительными являются примеры №8 и №10.

Ясно, что стиль оформления заданий должен быть единым по всему тесту. Если решено для содержательного блока использовать строчные буквы, то этому правилу должны подчиняться все задания независимо от количества строк в блоке.

При компьютерном тестировании текстовых и графических

возможностей для построения структуры задания гораздо больше. Однако здесь следует предостеречь от необоснованного использования ярких цветовых гамм как для текстовой (содержательной) части задания, так и для фона, наличия анимированных объектов. Считается, что если программа предназначена для младших школьников, то наличие в диалоговых окнах различных анимированных объектов (подмигивающие рожицы, уморительные кошечки, бегающие по буквам текста и т.п.) повышают дружественность интерфейса. Это очень спорный вопрос и требуются специальные исследования для выявления роли диалоговых интерфейсов компьютерных программ в вербальном восприятии информации испытуемыми.

2.5 ПРИНЦИПЫ ФОРМУЛИРОВАНИЯ ЗАДАНИЙ С ВЫБОРОМ

Разработка заданий с выбором, несмотря на их кажущуюся простоту, является непростым делом. Более того, это настоящее искусство – создание удачного задания в тестовой форме[2],[13]. Созданию качественных заданий в тестовой форме способствует следование принципам их разработки.

В.Аванесов[2], выделяет две группы принципов формулирования заданий. Одна группа используется при подборе ответов к заданиям, другая – при разработке содержания заданий.

Подбор ответов к заданиям можно осуществлять на основе следующих принципов:

1) противоречивости;
2) противоположности;
3) однородности;
4) кумуляции;
5) сочетания;
6) градуирования;
7) удвоенного противопоставления.

Содержание заданий формулируется на основе следующих принципов:

8) фасетности;
9) импликации.

Приведенные принципы оказывают существенную помощь разработчику в создании качественных заданий в тестовой форме. Каждый из перечисленных принципов, будет рассмотрен далее на конкретных примерах.

2.6. ЗАДАНИЯ С ДВУМЯ ОТВЕТАМИ

Задания этого вида содержат всего два варианта ответа – один верный ответ и один дистрактор. Таким образом, это задание альтернативных ответов – «Истина» - «Ложь». В качестве вариантов обычно используются ответы «Да» - Нет», «Верно» - «Не верно», «Имеет» - «Не имеет» и т.п. Задание формулируется в виде утверждения, которое после выбора ответа превращается в истинное или ложное высказывание.

Задания с двумя ответами характеризуются высокой вероятностью угадывания правильного ответа – 50% и могут использоваться для быстрого и грубого отсева испытуемых по принципу «зачет – не зачет».

Пример № 12.

ПРИ ПРОХОЖДЕНИИ ТОКА ЭЛЕКТРИЧЕСКИЙ ЗАРЯД
+1) переносится
2) не переносится

Проблему угадывания можно решить, вводя поправку на угадывание (глава 5) или используя серии заданий с альтернативами для одного элемента знаний. В этом случае задание состоит из серии - нескольких заданий с выбором. Задание считается выполненным, если на все задания в серии получен верный ответ.

Пример № 13.

Обвести кружочком ответ «да» или «нет»
ПРИ РАВНОМЕРНОМ ДВИЖЕНИИ ТЕЛА НЕИЗМЕННЫ

+да	нет	скорость
+да	нет	ускорение
да	+нет	путь
да	+нет	время
да	+нет	перемещение

Вероятность угадывания в этом случае резко падает (в рассмотренном примере не превышает 2%). А.Майоров, ссылаясь на рекомендации CITO[14], считает, что задания из серий в большей степени подходят для выявления уровня овладения сложными определениями, знаниями достаточно сложных графиков, диаграмм, схем и т.д.

При построении заданий с двумя ответами часто используется

■ принцип ПРОТИВОРЕЧИВОСТИ, который означает, что в

вариантах ответа используются отрицания

Пример № 14.

ПРИ УМЕНЬШЕНИИ РАССТОЯНИЯ МЕЖДУ ЗАРЯДАМИ, КУЛОНОВСКАЯ СИЛА ВЗАИМОДЕЙСТВИЯ
+1) увеличивается
2) не увеличивается

■ Принцип ПРОТИВОПОЛОЖНОСТИ очень похож на принцип противоречивости, но он реализуется с помощью антонимов, а не отрицаний.

Пример № 15.

ПРИ УМЕНЬШЕНИИ РАССТОЯНИЯ МЕЖДУ ЗАРЯДАМИ, КУЛОНОВСКАЯ СИЛА ВЗАИМОДЕЙСТВИЯ
+1) увеличивается
2) уменьшается

3.7. ЗАДАНИЯ С ТРЕМЯ И БОЛЕЕ, ОТВЕТАМИ

Задания с тремя, четырьмя, и т.д., ответами имеют более широкую область применения по сравнению с заданиями с альтернативой, так имеют большее количество дистракторов.

Рассмотрим применение принципа противоположности к созданию заданий с тремя ответами.

Пример № 16.

ВЫСОТА ПОЛЕТА ТЕЛА, БРОШЕННОГО ПОД УГЛОМ К ГОРИЗОНТУ
1) увеличивается
3) уменьшается
+3) сначала увеличивается, затем уменьшается

В этом задании второй ответ противоположен первому, а третий - первому и второму.

Пример № 17.

ПРИ УМЕНЬШЕНИИ РАССТОЯНИЯ МЕЖДУ ЗАРЯДАМИ
И ИХ ВЕЛИЧИНЫ В 2 РАЗА, КУЛОНОВСКАЯ СИЛА

ВЗАИМОДЕЙСТВИЯ
1) увеличивается
+2) не изменяется
3) уменьшается

■ Принцип ОДНОРОДНОСТИ требует использования в ответах членов одного множества, одного гомологического ряда.

Пример № 18.

ПРИНЦИП РАДИОСВЯЗИ ВПЕРВЫЕ ПРОДЕМОНСТРИРОВАН
+1) Поповым А.С.
2) Маркони Г.
3) Эдисоном Т.А.

Здесь в ответах используются элементы множества «фамилии людей, занимавшихся электротехникой».

Попытаемся нарушить принцип однородности.

Пример № 19.

ПРИНЦИП РАДИОСВЯЗИ ВПЕРВЫЕ ПРОДЕМОНСТРИРОВАН
+1) Поповым А.С.
2) Маркони Г.
3) Американским изобретателем

Третий ответ является элементом другого множества - «изобретатели из различных стран». Формально последний пример позволяет достичь поставленной цели - выяснить знает ли испытуемый, кто впервые продемонстрировал принцип радиосвязи. Однако, бросающаяся в глаза неоднородность ответов, позволяет сделать выводы о неравной привлекательности дистракторов, а это уже недопустимо.

■ Принцип КУМУЛЯЦИИ предполагает последовательное включение предыдущего ответа в последующий.

Пример № 20.

ПРИНЦИП РАДИОСВЯЗИ ВПЕРВЫЕ ПРОДЕМОНСТРИРОВАН
+1) Поповым А.С.
2) Поповым А.С. и Маркони Г.
3) Поповым А.С., Маркони Г. и Эдисоном Т.А.

Слабые испытуемые, предполагая, что самый полный ответ - правильный, пытаясь угадать, будут выбирать самый последний из предложенных ответов. Это следует учитывать при конструировании задания - самый полный (длинный) ответ не всегда должен быть верным. Отметим, что, постоянно следуя правилу - самый полный ответ всегда не верный, мы превращаем этот ответ в неработающий дистрактор. Это произойдет, как только испытуемому попадутся несколько подобных заданий, верные ответы на которые, он знает.

Отметим, что первый ответ, являющийся правильным, по принципу кумуляции включен во второй и третий ответы, что делает их частично правильными. Более подробно мы обсудим это при рассмотрении заданий с градуированными заданиями.

■ Принцип СОЧЕТАНИЯ предполагает использование в вариантах ответа сочетаний слов, терминов, чисел, знаков.

Пример № 21.

ДЛЯ РАДИОСВЯЗИ НЕОБХОДИМЫ
+1) передатчик и приемник
2) выпрямитель и усилитель
3) усилитель и детектор

■ Принцип ГРАДУИРОВАНИЯ предполагает использование градаций какой-либо характеристики.

Пример № 22.

ПРИ УВЕЛИЧЕНИИ ОБЪЕМА ТЕЛА ЕГО МАССА
+1) увеличивается
2) не изменяется
3) уменьшается

■ Принцип УДВОЕННОГО ПРОТИВОПОСТАВЛЕНИЯ обычно используется в заданиях с четырьмя ответами.

Пример № 23.

ПРИ НЕРАВНОМЕРНОМ ДВИЖЕНИИ ТЕЛА
+1) меняются скорость и ускорение
2) меняется скорость, но не меняется ускорение
3) не меняется скорость, но меняется ускорение
4) не меняются ни скорость, ни ускорение

При составлении задание можно использовать сочетание принципов.

Пример № 24.

КУЛОНОВСКОЕ ВЗАИМОДЕЙСТВИЕ ЧАСТИЦ ВОЗНИКАЕТ, ЕСЛИ ИХ ЗАРЯД
1) положительный
2) отрицательный
3) положительный или отрицательный

Здесь использованы принципы противоположности и сочетания.

Рассмотрим, теперь, применение принципов формулирования основной, содержательной части задания.

■ Принцип ФАСЕТНОСТИ предполагает использование фасетов.

Фасет (от англ. facet – грань, сторона) это специальная конструкция, состоящая из набора однородных элементов, используемых для формирования различных вариантов содержательной основы задания.

Фасеты позволяют легко создавать параллельные задания.

В.Аванесов дает следующее определение параллельных заданий: "Задания, образованные заменой элементов из фасета, во многих случаях, но не всегда, можно называть параллельными по содержанию".

Рассмотрим пример фасетного задания.

Пример № 25.

ПРИ { равномерном / не равномерном } ДВИЖЕНИИ { скорость / ускорение } ТЕЛА

1) изменяется
2) не изменяется

Использование фасетов позволяет повысить информационную безопасность тестирования. Фасеты позволяют легко организовать банк тестовых заданий для заданной предметной области,

использование, которого очень эффективно при организации компьютерного тестирования. О защите банка тестовых заданий будет сказано в главе 4.

■ Принцип ИМПЛИКАЦИИ предполагает использование логического условия «ЕСЛИ … ТО …»

Пример № 26.

ЕСЛИ ЕМКОСТЬ КОНДЕНСАТОРА КОЛЕБАТЕЛЬНОГО КОНТУРА УВЕЛИЧИВАЕТСЯ, ТО УМЕНЬШАЕТСЯ

+1) резонансная частота
2) период колебаний
3) амплитуда колебаний

Импликация не всегда осуществляется с помощью конструкции «ЕСЛИ … ТО …», которая может быть неявной.

Пример № 27.

УВЕЛИЧЕНИЕ ЕМКОСТИ КОНДЕНСАТОРА КОЛЕБАТЕЛЬНОГО КОНТУРА ВЫЗЫВАЕТ УМЕНЬШЕНИЕ

+1) резонансной частоты
2) периода колебаний
3) амплитуды колебаний

Остановимся на таком вопросе как использование в заданиях ответов типа «все вышеперечисленные», «ни один из вышеперечисленных». В.Аванесов считает, что «закон исключения третьего» требует обязательного наличия верного ответа, налагает логический запрет на использование вышеупомянутых ответов («все» или «ни один»). Этой же точки зрения придерживается А.Майоров. В.Переверзев считает, что в некоторых случаях такие ответы можно использовать. Например, в заданиях, в которых испытуемые могут получить правильный ответ, используя только информацию в основном содержательном блоке задания, перед поиском правильного ответа среди предлагаемых вариантов. Обычно это относится к заданиям на вычисления.

Рассмотрим пример 4.13 у В.Переверзева.

Какую часть от трех четвертых составляет одна десятая?

A) 1/8
+B) 2/15

C) 3/40
D) 15/2
E) Ни один из вышеперечисленных не является правильным

В.Переверзев считает, что это задание можно использовать и с четырьмя ответами (от А до Е), но тогда вероятность угадывания возрастет с 20% (задание с пятью ответами) до 25% (задание с четырьмя ответами). Следовательно, дифференцирующая способность задания уменьшится. Более того, добавление ответа (Е), «ни один из вышеперечисленных», позволяет повысить трудность, в том числе и потому, что ответ (Е) может быть и правильным[11].

Отметим, что проблему повышения вероятности угадывания в этом задании можно решить добавлением еще одного числового дистрактора. Из трех числовых дистракторов два - (С) и (D) являются правдоподобными (их можно получить путем различных манипуляций с числами 3/4 и 1/10). Дистрактор (А) очень слабый, представляющий собой произвольное число. Если разработчик задания решил использовать дистрактор типа (А), то, что ему мешает включить в ответы еще один, такой же слабый дистрактор, например число 3/15? Тогда количество ответов будет доведено до пяти – отпадает довод о возрастании вероятности угадывания.

Несмотря на свое положительное отношение к ответам «все» или «ни один», В.Переверзев предупреждает, что создать хорошее задание с такими ответами трудно. Бездумное употребление таких ответов выглядит как искусственное увеличение количества ответов в задании до нужного значения. Испытуемые достаточно легко это обнаруживают и, в дальнейшем, понимая, что это вероятнее всего дистракторы, даже не будут утруждать себя их анализом (число работающих дистракторов уменьшится). Ясно, что тогда вероятность угадывания повышается.

С нашей точки зрения, применение ответов «все» или «ни один» не оправдано по следующим причинам:

- ухудшение дизайна задания;
- повышение вероятности угадывания, так как чаще всего это будет не работающий дистрактор;
- нарушение закона «исключение третьего» - если в инструкции указывается «указать верный ответ», то в задании он должен быть;
- увеличение ошибки измерения, вызванной запутыванием испытуемого некачественным заданием, а такое вполне возможно ввиду высокой сложности создания корректного задания такого вида.

2.8. ЗАДАНИЯ С ВЫБОРОМ НЕСКОЛЬКИХ ПРАВИЛЬНЫХ ОТВЕТОВ

Задания с выбором одного правильного ответа (одна из разновидностей задания в закрытой форме) справедливо критикуются за довольно высокую вероятность угадывания верного ответа. Этого недостатка лишены задания с выбором нескольких правильных ответов. Такие задания иногда называют заданиями с множественным выбором[11]. В этих заданиях в блоке ответов размещено несколько верных ответов и несколько дистракторов.

Пример № 28.

УВЕЛИЧЕНИЕ ЕМКОСТИ КОНДЕНСАТОРА КОЛЕБАТЕЛЬНОГО КОНТУРА ВЫЗЫВАЕТ

+1) уменьшение резонансной частоты
+2) увеличение периода колебаний
3) увеличение резонансной частоты
4) уменьшение периода колебаний
5) увеличение резонансной частоты и уменьшение периода колебаний
6) увеличение резонансной частоты и увеличение периода колебаний
+7) уменьшение резонансной частоты и увеличение периода колебаний
8) уменьшение и резонансной частоты и периода колебаний

Блок ответов получился громоздким и его можно улучшить, используя предположение, что испытуемым известны обозначения физических величин: ω_0 – резонансная частота, T – период колебаний.

Пример № 29.

УВЕЛИЧЕНИЕ ЕМКОСТИ КОНДЕНСАТОРА КОЛЕБАТЕЛЬНОГО КОНТУРА ВЫЗЫВАЕТ

+1) уменьшение ω_0
+2) увеличение T
3) увеличение ω_0
4) уменьшение T
5) увеличение ω_0 и уменьшение T
6) увеличение ω_0 и увеличение T
+7) уменьшение ω_0 и увеличение T
8) уменьшение и ω_0 и T

Увеличение количества верных ответов приводит к общему увеличению числа ответов. Если считать оптимальным соотношение один верный ответ на два дистрактора (например, в заданиях с тремя ответами), то при трех верных ответах потребуется 6 дистракторов, итого 9 вариантов ответов. Это достаточно трудно.

Оценивание выполнения такого задания сложнее, чем оценивание задания с выбором одного верного ответа. В.Аванесов[2] предлагает за полностью правильное решение дать три балла, за каждую ошибку снимать один балл. Если ошибок больше трех, то давать 0 баллов. Таким образом, максимальное число баллов равно 3, а минимальное – 0, то есть испытуемый может получить за выполнение такого задания 0, 1, 2, 3 балла. Предложенная схема позволяет получать только положительные баллы в предположении, что число верных ответов равно трем и более.

Схема оценивания задания с выбором нескольких верных ответов получается более сложной, чем для заданий с выбором одного ответа, кроме того, вклады в итоговый результат (индивидуальный балл испытуемого) у них разный. В первом случае за одно полностью выполненное задание испытуемый получает 3 балла, а во втором случае – 1 балл. Это может привести к снижению точности измерений такого теста.

М.Челышкова[10] рекомендует за полностью выполненное задание с выбором нескольких верных ответов давать 1 балл и 0 баллов за, хотя бы один, неверный ответ.

В.Переверзев[11] описывает метод «частичного балла» (partial credit), в котором за каждый правильно выбранный ответ дается 1 балл, за неправильно выбранный ответ – 0 баллов. Штрафные баллы в этом методе не предусмотрены.

На наш взгляд, использование заданий с выбором одного верного ответа предпочтительней. Единственным преимуществом заданий с выбором нескольких верных ответов является хорошая защищенность от угадывания. Однако весьма непросто создать задание содержащее и несколько верных ответов и большое количество очень хороших дистракторов. При слабых дистракторах защищенность от угадывания будет сильно снижаться. Лучше все же заменить одно задание с выбором нескольких верных ответов на несколько заданий с выбором одного верного ответа. Можно также использовать серийные задания (Пример № 13.), но остается проблема оценивания.

2.9. ЗАДАНИЯ С ГРАДУИРОВАННЫМИ ОТВЕТАМИ (ЗАДАНИЯ С ВЫБОРОМ НАИЛУЧШЕГО ОТВЕТА)

Одним из доводов против заданий с выбором одного верного ответа является утверждение, что испытуемые могут запомнить неверные ответы (дистракторы) и это приведет к снижению их уровня подготовленности. В заданиях с выбором наилучшего ответа можно дать все ответы верные, но в различной степени. Метод оценивания всех ответов (grading) сильно усложняет задачу испытуемого. Назовем такие задания - заданиями с градуированными ответами. В литературе эти задания имеют разные названия – «задания с выбором наилучшего ответа»[2], «задания с множественным выбором»[11]. В таких заданиях испытуемому надо не просто выбрать верный ответ, но еще дать свою оценку остальным ответам. Угадывание в такой ситуации практически исключено. Отметим, что в вариантах ответов иногда указывают и неверные ответы.

Пример № 30.

ЕСЛИ ЗА РАВНЫЕ ПРОМЕЖУТКИ ВРЕМЕНИ ТЕЛО ПРОХОДИТ РАВНЫЕ РАССТОЯНИЯ, ТО ЭТО ДВИЖЕНИЕ
1) равномерное
2) с постоянной скоростью
3) без ускорения
4) неравномерное

Разработчик задания предположил, что первому ответу, как наиболее правильному, соответствует оценка «5», второму, менее правильному» - «4», третьему – «3» и четвертому, неправильному – «2». Такие градуированные задания, (задания с множественным оцениванием), порождают ряд проблем.

Во-первых, крайне сложно создать градуированные ответы однозначно, оцениваемые разными людьми, что обусловлено их субъективизмом.

Во-вторых, достаточно сложная процедура оценивания задания в целом.

Рассмотрим эти проблемы. Нередки случаи, когда за один и тот же ответ учащегося два разных учителя ставят разные оценки. К.Ингенкамп[15] указывает, что расхождение может составлять 2 балла (от 3 до 5). Мало того, один и тот же учитель за один и тот же ответ (видеозапись, показанная с интервалом несколько месяцев) ставит

разные оценки. Рассмотрим вышеприведенное задание (Пример № 30.. Разработчик задания оценивает второй ответ на «4», а другой эксперт вполне может счесть, что этот ответ заслуживает оценки «5». Относительно легко градуированные задания создаются с использованием принципа кумуляции.

Пример № 31.

СИЛА ПЕРЕМЕННОГО ТОКА НА УЧАСТКЕ ЦЕПИ ОПРЕДЕЛЯЕТСЯ
1) емкостью
2) емкостью и индуктивностью
3) емкостью, индуктивностью и активным сопротивлением
4) емкостью, индуктивностью, активным сопротивлением и реактивным сопротивлением

Целью задания было выяснить – знает ли испытуемый, что сопротивление переменному току оказывают емкость (конденсатор), индуктивность (катушка индуктивности) и активное сопротивление (резистор). В этом задании использован принцип кумуляции, что позволяет произвести точное оценивание ответов. 3-й ответ самый полный (указаны три параметра) – оценка «5». Второй ответ менее полный (два параметра) – оценка «4», Первый ответ верный, но неполный (1 параметр) - оценка «3». Четвертый ответ – неверный – оценка «2». Логика оценивания была такая. Выбор четвертого ответа означает, что испытуемый неплохо знает, какие параметры влияют на силу переменного тока, но не понимает, что емкость и индуктивность как раз - таки задают реактивное сопротивление. Указание реактивного сопротивления сводит на нет, положительный эффект от указания емкости и индуктивности. Поэтому оценка самая низкая. Однако, следует отметить, что если первые три ответа оценивались на уровне «знание» таксономии Блума, то четвертый – на уровне «понимание». Такое смешение подходов к оцениванию недопустимо. Четвертый ответ необходимо переработать и перевести его на уровень «знание».

Четвертый ответ самый длинный и это важно, так как он является дистрактором. Если в задании есть очень короткие и очень длинные ответы, то следует избегать ситуации, когда эти, бросающиеся в глаза, ответы были верными. Чаще всего испытуемые, не зная верного ответа, пытаясь угадать, выбирают самый длинный.

Принцип кумуляции можно использовать и в скрытом виде.

Пример № 32.

УКАЖИТЕ ЧЕТЫРЕ ФИЗИЧЕСКИХ ЯВЛЕНИЯ[16]
1) таяние снега, дождь, радуга, метель
2) таяние снега, дождь, землетрясение, почернение серебряной монеты
3) таяние снега, скисание молока, гниение соломы, закат солнца
4) скисание молока, гниение соломы, почернение серебряной монеты, растворение соли в воде

Это задание также позволяет однозначно оценить ответы, так как имеется удобный числовой параметр – количество физических явлений, упомянутых в ответе. В первом ответе все четыре явления – физические, оценка «5». Во 2-м ответе три явления физические, одно – химическое, оценка «4». В 3-м ответе два – физические, два – химические, оценка «3». В 4-м ответе физических явлений нет, оценка «2».

Теперь обратимся к проблеме оценивания ответа испытуемого.

По нашему мнению, конструирование заданий с градуированными ответами с четким критерием оценивания ответов является сложной задачей, требующей немалого искусства от составителя. По этой причине лучше использовать градуированные задания с уменьшенным количеством градаций правильности ответов, например, только «5» и «2». Иными словами, среди предложенных есть несколько совершенно верных и несколько совершенно неверных ответов. В этом случае требования к критерию оценивания значительно ослабляются и задание составить гораздо легче. Легко видеть, что тогда задание с градуированными ответами превращается в задание с выбором нескольких правильных ответов.

Рассмотрим процедуру оценивания заданий с градуированными ответами. Как отмечалось в первой главе, технология мягкого (soft testing), непрямого тестирования развивается И.А.Моревым[17], под руководством которого разработано программное средство для ЭВМ - «STEACHER». Эта компьютерная программа позволяет осуществить тестирование в форме дидактической игры. Для оценивания заданий с градуированными ответами И.А.Морев вычисляет скалярный рейтинг

$$R = \left(\left(\prod_{k=1}^{4} \left(1 + \frac{P_k}{P_h^N} \right) \right)^{\frac{1}{4}} - 1 \right) \cdot 100\%$$

где h - нумерует уровни ответов, P_h^N - нормативный показатель уровня h.

Для полноты расчетов необходима еще оценочная таблица, кроме того, нормативный показатель в некоторых случаях оказывается отрицательным, что лишено смысла и, поэтому, требуются специальные правила пересчета.

О.Н.Фалалеева[18] предлагает другой метод оценивания, в котором вводятся три величины: X - "номер ответа", Y - "оценка ответа", Z - "номер тестового задания". Величины X, Y представляют собой распределение оценок по ответам на Z-вое тестовое задание. На рис.2.9.1 показано такое распределение для одного из заданий. Здесь по оси ординат отложены значения Y, а по оси абсцисс – X.. Кружочки соответствуют "истинным оценкам", выставленным экспертами в данной предметной области; треугольниками - оценки, выставленные испытуемым каждому ответу данного тестового задания.

Необходимо оценить эту совокупность (множество) оценок испытуемого. Оценивание зависит от величины отклонения ΔY в плоскости XY (рис.2.9.1).

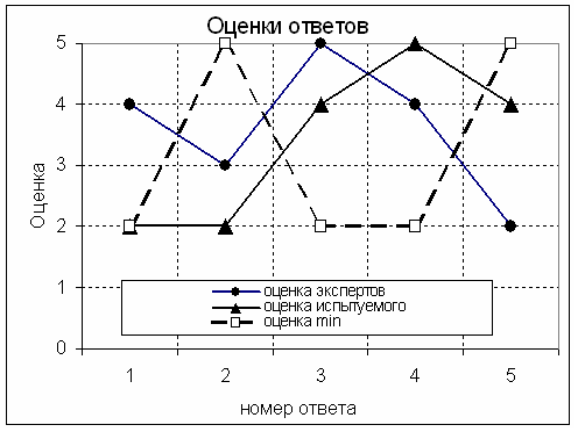


Рис.2.9.1. Оценивание заданий с градуированными ответами.

При нулевой разности всех ординат ΔY=0 испытуемый получит наивысшую оценку "5". На практике, конечно, суммарная разность ординат вероятнее всего будет ненулевой. В этом случае необходимо вычислить относительную долю суммарного отклонения. Для этого необходимо знать максимально возможное отклонение. В качестве максимального отклонения оценки k -го ответа возьмем наибольшую из двух величин (a и b) - расстояние между "истинным"

ответом и максимальной или минимальной оценками. В частности , Y_{max}="5" , Y_{min}="2" . Очевидно, что $Y_{min}+ a + b = Y_{max}$. Нас интересует только сдвиг $\Delta Y= a + b$, поэтому абсолютные значения Y_{min} и Y_{max}.не имеют значения. Суммарное отклонение для z-го задания будет равно

$$\Delta Y_z = \sum_{k=1}^{K} \Delta y_{zk}$$

где k - номер ответа, K - количество ответов в данном тестовом задании.

В нашем примере K=5. Теперь найдем $\Delta Y_{\max}$ - максимальное отклонение для данного тестового задания. Ясно, что эта величина будет различной для разных тестовых заданий. Величина $\Delta Y_{\max}$ будет складываться из максимальных отклонений для каждой оценки.

В нашем случае, например, для первого ответа ΔY_{z1}=4-2=2 и $\Delta Y_{z\max} = 2+2+3+2+3=12$.

Теперь найдем текущее отклонение:

$$\Delta Y_z = \sum_{k=1}^{K} \Delta y_{zk} = 2+1+1+1+2 = 7$$

или в относительных единицах (степень отклонения)

$$P_z = \frac{\Delta Y_z}{\Delta Y_{z\max}} = \frac{7}{12} = 0{,}58$$

Удобнее пользоваться не степенью отклонения P_z, а степенью совпадения S_z = 1- P_z.=0,42. Величина S_z характеризует оценку испытуемого за ответ на данное задание.

Таким образом, задания с градуированными ответами, очень привлекательны. Они не подвержены эффекту угадывания, но создать ответы с четкими градациями нелегко. Процедура оценивания довольно громоздкая и требует обязательного использования вычислительной техники.

2.10. ЗАДАНИЯ НА УСТАНОВЛЕНИЕ СООТВЕТСТВИЯ

В тех случаях, когда целью задания является выяснить – умеет ли испытуемый находить связи, ассоциации между явлениями, событиями, процессами, структурными единицами и т.д., используются задания на установление соответствия.

Структурно задание оформляется следующим образом. В верхней части задания приводится инструкция «Установить соответствие». Если в тесте такие задания выделены в отдельную серию, то инструкцию можно дать только в начале серии.

Под инструкцией указываются наименования двух колонок – левой и правой. В левой колонке размещены элементы первого множества (элементы левого столбца). Эти элементы помечены цифрами. В правой колонке размещены элементы второго множества, помеченные буквами (элементы правого столбца). Необходимо строго придерживаться следующего правила – количество элементов в правом столбце должно быть больше числа элементов в левом столбце. Если количество элементов в обоих столбцах одинаковое, то при установлении соответствия для последнего левого элемента останется единственный правый элемент, что приведет к исчезновению выбора и прекращению работы задания, что недопустимо. По этой причине количество правых элементов должно в полтора – два раза превышать количество левых элементов.

В самой нижней части задания приводится место для ответов – «Ответы: 1___, 2___ , 3___, 4___, ….». Рядом с номером элемента из левой колонки, испытуемый должен вписать букву соответствующего элемента из правой колонки.

При компьютерном тестировании возможны другие варианты оформления ответов. Например, последовательные щелчки мышью сначала по элементу левой колонки, затем по элементу правой колонки. При этом, выбранные элементы должны изменять свое текстовое (графическое) оформление с тем, чтобы испытуемый видел, какие элементы он уже использовал. В другом варианте, предлагается использовать переключатели «Check box», «Radio Button». Часто используется ввод букв, как указано в предыдущем абзаце, в специальное текстовое окно.

При оценивании задания поступают, так же как и в заданиях с выбором нескольких правильных ответов – либо давать 1 балл за полностью выполненное задание и 0 баллов за хотя бы одну ошибку, либо присваивать и снимать баллы за правильно и неправильно установленные соответствия. Можно также использовать метод

«частичного балла» (partial credit).

Пример № 33.

Установить соответствие

НАЗВАНИЕ		ФОРМУЛА
1. Закон Ома для участка цепи	А.	$I = \frac{R}{I}$
2. Закон Ома для замкнутой цепи	Б.	$I = \frac{E}{R + r}$
3. Закон Ома для участка цепи переменного тока	В.	$I = \frac{U}{R}$
	Г.	$I = \frac{E}{\sqrt{R^2 + \left(\omega L + \frac{1}{\omega C}\right)^2}}$
	Д.	$I = \frac{E}{\sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}}$

Ответы: 1 *В* , 2 *Б* , 3 *Д* .

Существует и другая (вторая) модификация заданий на установление соответствия – с многократным выбором[2,11]. В этих заданиях каждому элементу левого столбца могут соответствовать несколько элементов правого столбца. В.Аванесов считает задания с многократным выбором громоздкими и неудобными, порождающими множество ошибок, однако в медицинской практике они используются довольно широко.

Ясно, что в заданиях с многократным выбором число элементов правого столбца необязательно должно превышать число элементов левого столбца. Их количество может быть даже меньше.

Пример № 34[2].

Установить соответствие

СИНДРОМ	СИМПТОМ
1. Мозжечковый	А) атаксия
2. Вестибулярный	Б) головокружение
	В) дисметрия
	Г) адиадохокинез
	Д) нистагм
	Е) рвота
	Ж) интенция
	З) сканированная речь
	И) нарушения в калорической пробе
	К) гипотония

Ответы: 1-*А,Б,В,Г,Д,Ж,З,К.*
2 - *А,Б,Д,Е,И.*

Третья модификация заданий на установление соответствия предполагает наличие третьего множества. В таких заданиях элементы первого множества сопоставляются с элементами второго и третьего множеств.

Пример № 35[2].

ГОМОЛОГИЧЕСКИЙ РЯД	ФОРМУЛА	НАЗВАНИЕ
1. Алканы	I. CH_3OH	А) Этан
2. Алкены	II. C_2H_6	Б) Пропан
3. Алкины	III. C_6H_6	В) Пропен
4. Арены	IV. C_3H_6	Г) Этин
	V. C_2H_2	Д) Бензол
	VI. C_3H_8	Е) Пропин

Ответы: 1___, 2___, 3___, 4___.

2.11. ЗАДАНИЯ НА УСТАНОВЛЕНИЕ ПРАВИЛЬНОЙ ПОСЛЕДОВАТЕЛЬНОСТИ

Задания на правильную последовательность действий позволяют эффективно знания испытуемых в построении логических последовательностей, технологических цепочек, алгоритмов исполнения каких-либо процедур, построение хронологических последовательностей.

Структура заданий на правильную последовательность действий включает инструкцию «Установить правильную последовательность», вводную часть задания, список элементов, которые надо упорядочить. Испытуемый должен перед каждым элементом списка поставить порядковый номер, согласно логике содержания задания.

Пример № 36

Установите правильную последовательность
ЗАКОН ОМА ДЛЯ УЧАСТКА ЦЕПИ
☐ сопротивление
☐ напряжение
☐ сила тока
☐ прямо пропорционально
☐ обратно пропорционально

Правильная последовательность слов задается формулировкой закона Ома: «сила тока прямо пропорциональна напряжению на участке цепи и обратно пропорциональна сопротивлению участка цепи». Испытуемый вписывает номера в квадратики слева от элементов списка.

Другой вариант оформления таких заданий.

Пример № 37

Установите правильную последовательность
ЗАКОН ОМА ДЛЯ УЧАСТКА ЦЕПИ
1 сопротивление
2 напряжение
3 сила тока
4 прямо пропорционально
5 обратно пропорционально
Ответ: 3-4-2-5-1.

Отметим, что в данном примере есть еще один ответ, который можно формально можно считать правильным: 3-5-1-4-2. Однако стандартной формулировке закона Ома он не соответствует.

По мнению А.Н.Майорова задания на восстановление последовательности можно рассматривать как вариант задания на восстановление соответствия, когда одним из рядов является время, расстояние или иной континуальный конструкт, который подразумевается в виде ряда[12]. Переформулируем рассмотренное задание в задание на установление соответствия.

Пример № 38

Установить соответствие
ЗАКОН ОМА ДЛЯ УЧАСТКА ЦЕПИ

1. сопротивление	А) 1
2. напряжение	Б) 2
3. сила тока	В) 3
4. прямо пропорционально	Г) 4
5. обратно пропорционально	Д) 5

Ответ: 1 *Д* , 2 *В* , 3 *А* , 4 *Б* , 5 *Г*.

Нам удалось преобразовать задание, но достигли ли мы цели задания – выяснить – знает ли испытуемый формулировку закона Ома? Следует признать, что нет. Задание получилось запутанное, требуются значительные дополнительные усилия, чтобы, обладая полными знаниями по данному вопросу, правильно построить ответ. Испытуемый вполне может ошибиться именно на этапе построения ответа. Его неуспех будет обусловлен не слабой подготовленностью, а неудачным дизайном и несовершенной формой задания.

Таким образом, можно прийти к следующим выводам:

1) следует согласиться с А. Майоровым, что задания на установление правильной последовательности являются частным случаем заданий на установление соответствия с точки зрения логики их конструирования;

2) задания на установление последовательности имеют столь специфическую форму, что их не удается преобразовать в задания на установление соответствия без ухудшения их диагностических свойств. Иными словами, задания на установление последовательности

не сводимы к заданиям на установление соответствия с точки зрения оптимальности их формы;

3) задания на установление последовательности должны считаться классификационным элементом того же уровня, что и задания на установление соответствия, поскольку форма задания так же важна для построения тестового задания, как и его содержание.

2.12. ЗАДАНИЯ В ОТКРЫТОЙ ФОРМЕ

Для заданий в открытой форме мы будем рассматривать только задания вида «дополнение». Задания вида «свободное изложение» мы рассматривать не будем, поскольку они не технологичны и не могут использоваться в компьютерном тестировании. В заданиях свободного изложения проверку правильности осуществляет человек, от субъективизма которого мы хотим избавиться. В этом смысле задания со свободным изложением не являются заданиями в тестовой форме.

Задания в открытой форме принципиально отличаются от заданий в закрытой форме. Задания в закрытой форме содержат все необходимое для ответа, испытуемому только нужно отметить выбранные элементы в блоке вопросов. Задания же в открытой форме, требуют ввода дополнительной информации - дополнения.

Рекомендуется формулировать задания так, чтобы дополнение находилось в конце задания.

Пример № 39.

ЭЛЕКТРИЧЕСКИЙ ТОК СОЗДАЕТСЯ ДВИЖУЩИМИСЯ _____ .

Испытуемому надо дополнить задание словом «ЗАРЯДАМИ». В результате задание превращается в истинное высказывание.

Иногда не удается расположить дополнение в конце задания, тогда допустимо его расположение внутри задания.

Пример № 40.

СКОРОСТЬ СВЕТА В ВАКУУМЕ РАВНА _____ км/с

Крупным недостатком заданий в открытой форме является сложность подбора формулировки, четко и однозначно определяющей, какое дополнение необходимо вставить.

Пример № 41.

СИЛА ТОКА ИЗМЕРЯЕТСЯ В _____ .

В качестве правильного ответа здесь предполагалось дополнение «АМПЕРАХ». Но если испытуемый ответит «МИЛЛИАМПЕРАХ», это тоже будет верный ответ. В случае компьютерного тестирования испытуемый получит 0 баллов, просто потому, что эталон ответа, заложенный в память ЭВМ, не соответствует введенному слову. Произошло неверное оценивание ответа испытуемого, что приведет к увеличению ошибки измерения. Другой источник неверного оценивания - орфографические ошибки. Допустим, испытуемый ответил «АМПИРАХ». По существу он ответил правильно, ведь цель данного задания - выяснить, знает ли испытуемый наименование единицы измерения силы тока. Проверка же орфографии, грамматики - это цели совсем других заданий.

Пример № 42.

ВЕЛИКИМ РЕФОРМАТОРОМ РОССИИ БЫЛ ЦАРЬ ПЕТР _____ .

Эталон ответа – «Первый», однако ответы «Великий», «Алексеевич», по сути, тоже верные.

Из приведенных примеров ясно, что разработчик должен предусмотреть все возможные варианты правильных ответов. К сожалению, во многих случаях, это очень сложно сделать. Предпринимаются различные меры, с тем, чтобы упростить процесс создания заданий в открытой форме. В частности, предлагается сформировать множество верных ответов, затем в качестве эталона использовать корни слов, совокупности символов, характеризующих правильный ответ. В вышеприведенном примере в качестве эталона можно было бы использовать «АМП». Если во введенном слове содержатся эти символы, то ответ считать правильным.

Довольно легко конструируются задания в открытой форме для естественно-научных и технических дисциплин, где в качестве ответа надо вводить числа.

Пример № 43.

НАПРЯЖЕНИЕ 12 В ПРИЛОЖЕНО К УЧАСТКУ ЦЕПИ СОПРОТИВЛЕНИЕМ 13 ОМ. СИЛА ТОКА РАВНА _____ АМПЕР.

В отличие от словесных вариантов, в данном случае верный ответ единственный - 0,923. Отметим два момента. Необходимо обязательно указывать размерность ответа, в данном случае [АМПЕР]. Кроме того, необходимо указывать испытуемому с какой точностью ему следует проводить вычисления. В приведенном примере в качестве ответа получается бесконечная дробь

12/13 = 0,923076923076923076923076923076923 …

Рекомендуется указывать точность 10% - с одной стороны такие расчеты достаточно легко провести на калькуляторе, с другой - имеется возможность отличить верный ответ от неверного путем угадывания. Тогда верным ответом в нашем примере является 0,92. В необходимых случаях, когда результаты расчетов по правильным и неправильным формулам близки, следует указывать точность 1%. В некоторых случаях лучше указывать требуемое количество десятичных знаков после запятой.

При конструировании заданий в открытой форме следует придерживаться некоторых принципов, что позволит достичь высокого качества заданий. В.С.Аванесов предлагает шесть принципов, которыми необходимо руководствоваться разработчику.

1) Принцип ЛОГИЧЕСКОЙ ОПРЕДЕЛЕННОСТИ содержания задания.

Логически определенное задание дает возможность испытуемому правильно на него ответить, а содержание и форма задания способствуют этому.

2) Принцип ФАСЕТНОСТИ (ВАРИАТИВНОСТИ) содержания задания.

Пример № 44.

В СИСТЕМЕ СИ ЕДИНИЦЕЙ ИЗМЕРЕНИЯ { длины / времени / массы } ЯВЛЯЕТСЯ _____

Пример № 45.

ТОК СИЛОЙ { I }, ПРОХОДЯ ПО ПРОВОДНИКУ { R }, ЗА ВРЕМЯ { t } СЕКУНД ВЫДЕЛЯЕТ _____ ДЖОУЛЕЙ ТЕПЛА.

Это задание содержит три фасета, вариативность которых обеспечивает подстановка конкретных числовых значений вместо I, R и t. В компьютерном тестировании это позволяет создать сколько угодно вариантов для каждой переменной I, R и t.

3) Принцип ПАРАЛЛЕЛЬНОСТИ.

Этот принцип включает в себя три понятия

а) параллельности по содержанию;

б) параллельности по содержанию и по форме;

в) параллельности по содержанию, форме и по трудности заданий.

Параллельность по содержанию обеспечивается использованием принципа фасетности.

4) Принцип ОБРАТИМОСТИ.

Этот принцип предполагает использование обратимых утверждений.

Пример № 46.

В СИСТЕМЕ СИ ЕДИНИЦЕЙ ИЗМЕРЕНИЯ СИЛЫ ТОКА ЯВЛЯЕТСЯ _____.

Обратное утверждение:

В СИСТЕМЕ СИ АМПЕР ЯВЛЯЕТСЯ ЕДИНИЦЕЙ ИЗМЕРЕНИЯ _____.

Пример № 47 (П.М.Эрдниев[19]).

6 + 3 = ____; 6 + ___ = 9; ___ + 3 = 9.

Первое задание обычно не вызывает трудностей у младших школьников. Ответ на второе и третье ищется в форме поиска. Психологически эти задания выполняются различно. Второе и третье задания ориентированы на более высокие уровни таксономии Блума по сравнению с первым. Эти задания внешне обратимы, но не параллельны.

5) Принцип КРАТКОСТИ.

Многословие усложняет восприятие содержания задания.

Добиться краткости при сохранении ясности - сложная задача. Если это удалось, то задание от этого сильно выиграет.

5) Принцип НЕОТРИЦАТЕЛЬНОСТИ.

Не рекомендуется использовать отрицания в основе задания.

Пример № 48.

ПРИ РАВНОМЕРНОМ, ПРЯМОЛИНЕЙНОМ ДВИЖЕНИИ ТЕЛО НЕ ДВИЖЕТСЯ _____.

Эталон ответа «ускоренно». Однако при равномерном, прямолинейном движении тело также не движется «замедленно», криволинейно, по спирали и т.д.

7) Принцип ИМПЛИКАЦИИ «если … то …».

Пример № 49.

ЕСЛИ УСКОРЕНИЕ ОТСУТСТВУЕТ, ТО ДВИЖЕНИЕ ____.

Характеризуя тестовые задания в открытой форме, следует отметить, что их несомненное достоинство – полное исключение угадывания. Недостатком, и очень сильным, является сложность:

а) формулирования ясного, недвусмысленного задания. Может оказаться, что испытуемый понял задание не так, как задумал разработчик теста. В закрытых заданиях, даже неудачно сформулированных, испытуемый получает дополнительную информацию из списка ответов, что помогает ему однозначно понять задание;

б) всестороннего и полного анализа ответа испытуемого – нередки случаи, когда испытуемый предлагает, по сути, верный ответ, но его формулировка отличается от эталона. Все возможные варианты верных по существу ответов, должны быть предусмотрены в множестве эталонов к заданию;

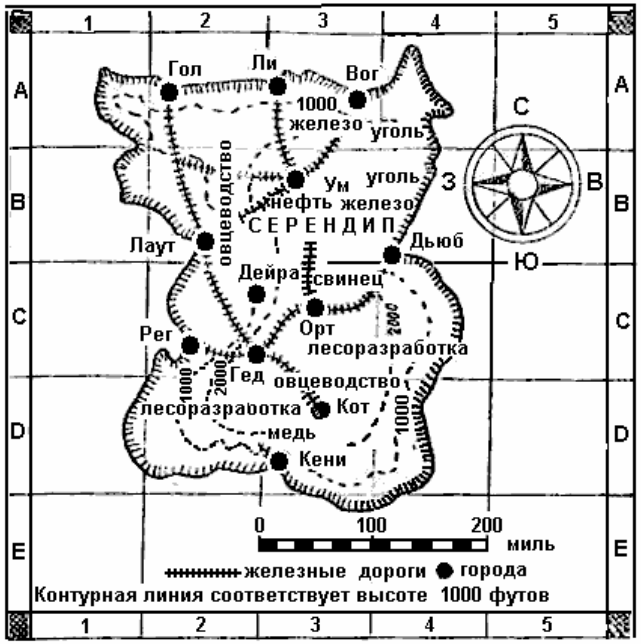
в) не технологичность – как правило, проверка ответов возлагается на экзаменатора, машинную проверку, особенно в гуманитарных областях, далеко не всегда удается реализовать.

В тех случаях, когда этот недостаток преодолен, можно получить хорошее тестовое задание, позволяющее повысить точность измерений уровня знаний. Отметим, что Георг Раш, при

экспериментальной проверке своей теории (глава 5) настаивал на применении тестовых заданий именно в открытой форме.

В целом, взвешивая все «за» и «против», мы считаем, что задания в закрытой форме предпочтительнее, поскольку практически исключают занижение индивидуального балла из-за неудачно сформулированного задания. Эффект угадывания можно заметно уменьшить, используя задания: серийные, на соответствие, на правильную последовательность, с градуированными ответами. Можно также вводить поправки на угадывание.

В заключение отметим большое значение невербальной информации в тестовых заданиях для проверки сложных умений и навыков, соответствующим высшим уровням таксономии Блума. В следующем примере в основную часть задания включена карта островного королевства «Серендип»[20].



Какой из следующих городов был бы лучшим местом для металлургического завода?

A - Ли (3A) **C - Кот (3D)**
B - Ум (3B) **D - Дьюб (4B)**

Испытуемый должен, используя карту острова, определить, где с наибольшей выгодой можно было бы построить металлургический завод. При этом во внимание надо принять большое количество факторов: расположение месторождений угля, железа и меди; наличие железных дорог; выходы к портам, для отправки готовой продукции; рельеф местности; экологические требования и т.д. При построении таких заданий следует очень тщательно прорабатывать основную часть задания с тем, чтобы выбор верного ответа был однозначным[21].

Литература к главе 2

[1] Аванесов В.С. Композиция тестовых заданий. Учебная книга для преподавателей вузов, учителей школ, аспирантов и студентов педвузов. 2 изд., испр. и доп. М.: Адепт 1998. -217с.

[2] Аванесов В.С. Форма тестовых заданий. -М.: Центр тестирования, 2005. -156 с.

[3] Перышкин А.В. Физика. 7 кл.: Учеб. для общеобразоват. учеб. заведений. -6-е изд. стереотип. – М.: Дрофа, 2002. -192 с.

[4] Распопов ВМ. Программирование и организация самостоятельной работы учащихся. М.: Высшая школа, 1965.

[5] Roid G.H., Haladyna T.V. A Technology for Test-item Writing. -N.Y.: Academic Press, 1982.

[6] Киевский С.В. орфографический тренажер “ГРАМОТЕЙ-КЛАСС” - http://www.ito.su/1997/B/B11.html.

[7] Ким В.С. Коррекция тестовых баллов на угадывание //Педагогические измерения, 2006, №4. –С.47-55.

[8] Кромер В. Еще раз о коррекции тестового балла // Педагогические измерения, 2007, №1. –С.89-94.

[9] Михайлычев Е.А. Дидактическая тестология. -М.: Народное образование, 2001. -432 с.

[10] Челышкова М.Б. Теория и практика конструирования педагогических тестов: Учебное пособие. –М.: Логос, 2002. -432 с.

[11] Переверзев В.Ю. Технология разработки тестовых заданий: справочное руководство. –М.: Е-Медиа, 2005. -265 с.

[12] Майоров А.Н. – Теория и практика создания тестов для системы образования. – М.: «Интеллект-центр», 2001. -296 с.

[13] Popham W.J. Criterion-referenced measurement. Englewood Cliffs. - N.J.: Prentice Hall, 1978.

[14] Роберт Ван Криген, Стивен Баккер. Подготовка и проведение экзаменов. Руководство для организации и разработки централизованных экзаменов. CITO, Национальный институт по оценке достижений в области образования. -Амхем, Нидерланды, 1995.

[15] Ингенкамп К. Педагогическая диагностика. -М.: Педагогика, 1991. - 240 с.

[16] Фалалеева О.Н. Банк тестовых заданий по физике для тестовой оболочки STEACHER (Didactor). Зарегистрировано ОФАП 24.05.2007. Свидетельство о регистрации отраслевой разработки №8404.

[17] Морев И. А. Образовательные информационные технологии. Часть 2.Педагогические измерения: Учебное пособие. – Владивосток: Изд-во Дальневост. ун-та, 2004. -174 с.

[18] Фалалеева О.Н. Оценивание учебных достижений методом мягкого тестирования. Вестн. МГОУ. Серия "Открытое образование". - 2(33). Том 2. - 2006. - М.: Изд-во МГОУ. - С. 126-130.

[19] Эрдниев П.М. Укрупнение дидактических единиц как технология обучения. В 2-х частях. Ч.1. -М.: Просвещение, 1992. -175 с.

[20] Стоунс Э. Психопедагогика. Психологическая теория и практика обучения / Пер. с англ — М.: Педагогика, 1984. — 472 с.

[21] Кувондиков О.К., Ким В.С. Методические указания по составлению тестовых заданий. -Самарканд, Изд. Самаркандского гос. ун-та, 1992, - 47 с.

ГЛАВА 3. СТАТИСТИЧЕСКАЯ ОБРАБОТКА РЕЗУЛЬТАТОВ ТЕСТИРОВАНИЯ

Статистическая обработка результатов тестирования позволяет с одной стороны, объективно определить результаты испытуемых, с другой – оценить качество самого теста, тестовых заданий, в частности оценить его надежность. Проблеме надежности уделено много внимания в классической теории тестов. Эта теория не потеряла своей актуальности и в настоящее время. Несмотря на появление, более современных теорий, классическая теория продолжает сохранять свои позиции.

3.1. ОСНОВНЫЕ ПОЛОЖЕНИЯ КЛАССИЧЕСКОЙ ТЕОРИИ ТЕСТОВ

Создателем классической теории тестов (Classical Theory of mental tests) является известный британский психолог, автор факторного анализа, Чальз Эдвард Спирмен (Charles Edward Spearman) (1863-1945 г.)[1]. Он родился 10 сентября 1863 года, и четверть своей жизни прослужил в британской армии. По этой причине, степень доктора философии он получил только в возрасте 41 года[2]. Диссертационное исследование Ч.Спирмен выполнял в Лейпцигской лаборатории экспериментальной психологии под руководством Вильгельма Вундта (Wilhelm Wundt). В тот период на Ч.Спирмена сильное влияние оказали работы Фрэнсиса Гальтона (Francis Galton) по тестированию интеллекта человека. Учениками Ч.Спирмена были

R.Cattell и D.Wechsler. В числе его последователей можно назвать A.Anastasi, J. P. Guilford, P.Vernon, C.Burt, A.Jensen.

Большой вклад в развитие классической теории тестов внес Льюис Гуттман (Louis Guttman, 1916-1987)[3].

Всесторонне и полно классическая теория тестов впервые изложена в фундаментальном труде Гарольда Гулликсена (Gulliksen H., 1950 г.)[4]. С тех пор теория несколько видоизменялась, в частности совершенствовался математический аппарат. Классическая теория тестов в современном изложении приведена в книге Crocker L., Aligna J. (1986 г.)[5]. Из отечественных исследователей впервые описание этой теории дал В.Аванесов (1989 г.)[6]. В работе Челышковой М.Б. (2002 г.)[7] приведены сведения о статистическом обосновании качества теста.

Классическая теория тестов основывается на следующих пяти основных положениях.

1. Эмпирически полученный результат измерения (X) представляет собой сумму истинного результата измерения (T) и ошибки измерения (E)[8]:

$$X = T + E \qquad (3.1.1)$$

Величины T и E обычно неизвестны.

2. Истинный результат измерения можно выразить как математическое ожидание E(X):

$$T = E(X)$$

3. Корреляция истинных и ошибочных компонентов по множеству испытуемых равна нулю, то есть $\rho_{TE} = 0$.

4. Ошибочные компоненты двух любых тестов не коррелируют:

$$\rho_{E1,E2} = 0$$

5. Ошибочные компоненты одного теста не коррелируют с истинными компонентами любого другого теста:

$$\rho_{E1,T2} = 0$$

Кроме этого, основу классической теории тестов составляют два определения – параллельных и эквивалентных тестов.

ПАРАЛЛЕЛЬНЫЕ тесты должны соответствовать требованиям (1-5), истинные компоненты одного теста (T_1) должны быть равны истинным компонентам другого теста (T_2) в каждой выборке испытуемых, отвечающих на оба теста. Предполагается, что $T_1=T_2$ и, кроме того, равны дисперсии $s_1^2 = s_2^2$.

Эквивалентные тесты должны соответствовать всем требованием параллельных тестов за исключением одного: истинные компоненты одного теста не обязательно должны равняться истинным компонентам другого параллельного теста, но отличаться они должны на одну и ту же константу *c*.

Условие эквивалентности двух тестов записывается в следующем виде:

$$T_1 = T_2 + c_{12}$$

где c_{12} - константа различий результатов первого и второго тестов.

На основе приведенных положений построена теория надежности тестов[9,10].

Далее, примем в качестве исходного положения следующее утверждение

$$s_X^2 = s_T^2 + s_E^2 \qquad (3.1.2)$$

то есть, дисперсия полученных тестовых баллов равна сумме дисперсий истинных и ошибочных компонентов.

Перепишем это выражение в следующем виде:

$$\frac{s_T^2}{s_X^2} = 1 - \frac{s_E^2}{s_X^2} \qquad (3.1.3)$$

Правая часть этого равенства представляет собой надежность теста (*r*). Таким образом надежность теста можно записать в виде:

$$r = 1 - \frac{s_E^2}{s_X^2} \qquad (3.1.4)$$

На основе этой формулы в последующем были предложены различные выражения для нахождения коэффициента надежности теста. Надежность теста представляет собой его важнейшую характеристику. Если неизвестна надежность, то результаты тестирования невозможно интерпретировать. Надежность теста характеризует его точность как измерительного инструмента. Высокая надежность означает высокую повторяемость результатов тестирования в одинаковых условиях.

В классической теории тестов важнейшей проблемой является определение истинного тестового балла испытуемого (Т). Эмпирический тестовый балл (Х) зависит от многих условий – уровня трудности заданий, уровня подготовленности испытуемых, количества заданий, условий проведения тестирования и т.д. В группе сильных, хорошо подготовленных испытуемых, результаты тестирования будут как правило, лучше,. чем в группе слабо подготовленных испытуемых. В этой связи остается открытым вопрос о величине меры трудности заданий на генеральной совокупности испытуемых. Проблема заключается в том, что реальные эмпирические данные получают на вовсе не случайных выборках испытуемых. Как правило, это учебные группы, представляющие собой множество учащихся достаточно сильно взаимодействующих между собой в процессе учения и обучающиеся в условиях, часто не повторяющихся для других групп.

Найдем s_E из уравнения (3.1.4)

$$s_E = s_X \sqrt{1 - r}$$

Здесь в явной форме показана зависимость точности измерения от величины стандартного отклонения s_X и от надежности теста r.

3.2. МАТРИЦА РЕЗУЛЬТАТОВ ТЕСТИРОВАНИЯ

Выполнение статистической обработки результатов тестирования начинается с формирования так называемой матрицы тестовых результатов[6].

МАТРИЦА ТЕСТОВЫХ РЕЗУЛЬТАТОВ a_{ij} – это матрица размерности NxM, содержащая числовые обозначения градации индикатора, связанного с изучаемой латентной переменной, где M-число индикаторов, N-число испытуемых.

Эта матрица представляет собой таблицу, строки, которой соответствуют испытуемым, а столбцы – индикаторным переменным. В случае тестирования учебных достижений индикаторными переменными являются тестовые задания. На пересечении строк и столбцов находится число, соответствующее ответу данного испытуемого на данное задание.

В политомическом случае ответ испытуемого характеризуется числами в некотором диапазоне, например от нуля до девяти. Допустим, 0 соответствует полному отсутствию знаний, а 9 – наличию полных знаний для данного тестового задания. Промежуточные варианты описываются числами* в диапазоне от 0 до 9. Пример политомической матрицы результатов тестирования приведен в таблице 3.2.1.

Таблица 3.2.1. Политомическая матрица результатов тестирования

№	Испытуемые	**Номера заданий**		
		1	**2**	**3**
1	Иванов	3	4	0
2	Сидоров	5	8	4
3	Петров	9	8	5
4	Алексеев	6	5	3
5	Михайлов	8	7	0

Отметим, что числа, расположенные в ячейках таблицы, отсчитываются по порядковой шкале (см. главу 1). Рассмотрим, например, задание №1. Числа в политомической матрице показывают,

* Точнее – цифрами. Цифры, числа, состоящие из цифр, в нашем случае это просто упорядоченное множество, каждый элемент которого, обозначен некоторыми символами, например числами (цифрами, если количество элементов не более 10).

что с первым заданием испытуемые справились с различным успехом. В порядке убывания их можно ранжировать: Петров (9), Михайлов (8), Алексеев (6), Сидоров (5), Иванов (3). С этими цифрами нельзя производить арифметические операции – складывать, вычитать, умножать, делить (см. главу 1). Это происходит оттого, что цифры от 0 до 9 являются не числами, а упорядоченными символами для описания градации знаний испытуемых, то есть, эти символы (0-9) размещены на порядковой шкале.

Матрица состоит векторов-строк, содержащих значения индикатора для испытуемого. Матрицу можно упорядочить как по строкам, так и по столбцам.

ПРОФИЛЬ ИСПЫТУЕМОГО – это последовательность значений индикатора в упорядоченной матрице тестовых результатов.

В дихотомическом случае ответы испытуемого характеризуется двумя символами (цифрами) - 0 и 1. Нулю соответствует неверный ответ, единице – верный ответ.

На практике чаще всего используется дихотомический случай, поэтому в дальнейшем мы будем рассматривать именно его.

БИНАРНАЯ МАТРИЦА - это матрица результатов тестирования для дихотомического случая. Пример дихотомической матрицы приведен в таблице 3.2.2.

Таблица 3.2.2. Бинарная матрица (11x9).

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1. Иванов	0	0	1	0	1	0	1	1	1
2. Сидоров	0	0	0	0	1	0	0	1	1
3. Петров	0	0	0	0	0	1	0	1	0
4. Алексеев	0	0	0	0	1	0	1	1	1
5. Михайлов	1	0	0	0	1	0	0	1	0
6. Федоров	1	1	1	0	0	0	1	1	1
7. Антонов	1	1	1	1	1	1	1	1	1
8. Кузнецов	1	0	1	1	0	1	1	1	1
9. Болдырев	0	0	0	0	0	0	1	1	0
10. Яковлев	1	0	1	0	1	1	1	1	1
11. Громов	1	1	1	0	1	1	1	1	1

Для дальнейшего анализа, нам потребуются значения X_i - индивидуального балла i-го испытуемого, количество верных ответов

R_j, на j-е задание, количество неверных ответов W_j на j-е задание, доля верных ответов p_j и доля неверных ответов q_j.

$$X_i = \sum_{j=1}^{M} a_{ij}$$

В нашем случае (M=9) индивидуальный тестовый балл, например, для второго испытуемого (i=2) равен:

$$X_2 = \sum_{j=1}^{9} a_{2j} = a_{21} + a_{22} + a_{23} + ... + a_{29} = 1+0+0+1+0+0+0+1+1 = 4$$

$$R_j = \sum_{i=1}^{N} a_{ij}$$

$$W_j = N - R_j$$

В нашем случае (N=11) для третьего задания (j=3) получаем:

$$R_3 = \sum_{i=1}^{11} a_{i3} = a_{13} + a_{23} + a_{33} + ... + a_{93} + a_{103} + a_{113} =$$
$$= 1+0+0+0+0+1+1+1+0+1+1 = 6$$

$$W_3 = 11 - R_3 = 11-6=5$$

Доля верных ответов p_j на j-е задание равна:

$$p_j = \frac{R_j}{N}$$

Параметр p_j принято называть мерой трудности задания, хотя логично было бы называть его мерой легкости тестового задания.

Доля неверных ответов равна $q_j = 1 - p_j$.

В нашем случае для третьего задания получим:

$$p_3 = \frac{R_3}{N} = \frac{6}{11} = 0.545$$

$$q_3 = 1 - p_3 = 1 - 0.545 = 0.455$$

В таблице 3.2.3 приведены вычисленные значения X_i и R_j. Для удобства визуального анализа ячейки с нулевыми значениями заштрихованы. В отличие от предыдущей таблицы, здесь фамилии испытуемых заменены их номерами. Крайний правый столбец содержит значения индивидуальных тестовых баллов испытуемых X_i. Самая нижняя строка содержит количество правильных ответов на задания R_j. Для удобства визуального контроля, ячейки с нулевыми значениями заштрихованы.

Таблица 3.2.3. Бинарная матрица (11x9) с индивидуальными тестовыми баллами испытуемых.

	1	2	3	4	5	6	7	8	9	x_i
1	0	0	1	0	1	0	1	1	1	5
2	0	0	0	0	1	0	0	1	1	3
3	0	0	0	0	0	1	0	1	0	2
4	0	0	0	0	1	0	1	1	1	4
5	1	0	0	0	1	0	0	1	0	3
6	1	1	1	0	0	0	1	1	1	6
7	1	1	1	1	1	1	1	1	1	9
8	1	0	1	1	0	1	1	1	1	7
9	0	0	0	0	0	0	1	1	0	2
10	1	0	1	0	1	1	1	1	1	7
11	1	1	1	0	1	1	1	1	1	8
R_j	6	3	6	2	7	5	8	11	8	56

Далее, необходимо упорядочить бинарную матрицу. Сначала выполним упорядочение по величине X_i. Для этого строки с большими значениями X_i переместим вверх, с меньшими значениями - вниз. В результате, строка №7 передвинулась на самый верх, а строка №9 - вниз. Аналогично выполняется упорядочение по значению R_j. В этом случае перемещаются столбцы. Столбец №8 становится крайним левым, а №4 - крайним правым (таблица 3.2.4).

Отметим, что после упорядочения, нумерация строк и столбцов нарушилась. Столбец №8 является первым, №7 - вторым, №9 - третьим и т.д. Номера столбцов сохранены именно такими, какими были номера заданий в тесте. Это важно для установления взаимно однозначного соответствия между заданиями в упорядоченной и

исходной матрицах. Это же самое относится и номерам строк (испытуемых).

Таблица 3.2.4. Бинарная матрица (11x9), упорядоченная по X_i и по R_j

	8	7	9	5	1	3	6	2	4	x_i
7	1	1	1	1	1	1	1	1	1	9
11	1	1	1	1	1	1	1	1	0	8
8	1	1	1	0	1	1	1	0	1	7
10	1	1	1	1	1	1	1	0	0	7
6	1	1	1	0	1	1	0	1	0	6
1	1	1	1	1	0	1	0	0	0	5
4	1	1	1	1	0	0	0	0	0	4
2	1	0	1	1	0	0	0	0	0	3
5	1	0	0	1	1	0	0	0	0	3
3	1	0	0	0	0	0	1	0	0	2
9	1	1	0	0	0	0	0	0	0	2
R_j	11	8	8	7	6	6	5	3	2	56

Из полученной таблицы видно, что задание №8 успешно выполнили все 11 испытуемых. Это задание не позволяет дифференцировать испытуемых, поэтому его следует удалить из теста. Отметим, что это требование относится только к нормативно-ориентированным тестам. Соответственно, если бы на какое-то задание не ответил ни один испытуемый (R_j=0), то это задание тоже должно быть удалено из теста. Кроме того, испытуемый №7 успешно выполнил все задания теста. Тест не дает информации об испытуемом, за исключением того, что для него все задания слишком легкие. Строку №7 также следует удалить из матрицы.

После удаления столбца №8 и строки №7 мы получим редуцированную бинарную матрицу (таблица 3.2.5).

В этой таблице для удобства добавлена новая нумерация строк и столбцов. В нижней части таблицы приведены значения R_j, W_j, p_j, q_j, и $p_j q_j$.

Важным параметром тестового задания является вариация (дисперсия) тестовых баллов $p_j q_j$. Чем больше вариация, тем лучше задание дифференцирует испытуемых.

На рис.3.2.1. показана зависимость вариации тестовых баллов от трудности задания. Видно, что максимальное значение, равное 0,25 достигается при $p_j = 0{,}5$. При $p_j = 0$ и $p_j = 1$ дисперсия задания равна нулю. Иными словами, если на задание не ответил ни один

испытуемый или успешно ответили все, то задание не может их дифференцировать по уровню подготовленности.

Таблица 3.2.5. Редуцированная бинарная матрица 10x8.

№	новые	1	2	3	4	5	6	7	8		
новые	старые	7	9	5	1	3	6	2	4	x_i	x_i^2
1	11	1	1	1	1	1	1	1	0	7	49
2	8	1	1	0	1	1	1	0	1	6	36
3	10	1	1	1	1	1	1	0	0	6	36
4	6	1	1	0	1	1	0	1	0	5	25
5	1	1	1	1	0	1	0	0	0	4	16
6	4	1	1	1	0	0	0	0	0	3	9
7	2	0	1	1	0	0	0	0	0	2	4
8	5	0	0	1	1	0	0	0	0	2	4
9	3	0	0	0	0	0	1	0	0	1	1
10	9	1	0	0	0	0	0	0	0	1	1
	R_j	7	7	6	5	5	4	2	1	37	181
	W_j	3	3	4	5	5	6	8	9		
	p_j	0,7	0,7	0,6	0,5	0,5	0,4	0,2	0,1		
	q_j	0,3	0,3	0,4	0,5	0,5	0,6	0,8	0,9		
	$p_j q_j$	0,21	0,21	0,24	0,25	0,25	0,24	0,16	0,09		

Бинарная матрица (таблица 3.2.5) имеет характерную особенность - почти все нули и единицы распределились относительно диагонали, идущей из левого нижнего угла в правый верхний.

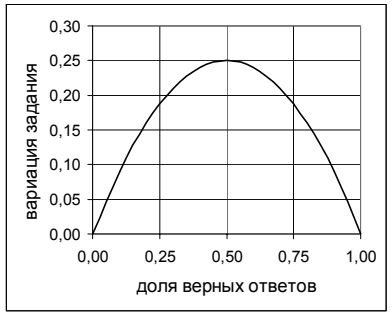


Рис.3.2.1. Вариация задания.

Согласно Гуттману это разграничение должно быть идеальным. Если испытуемый, верно ответил на трудное задание, то он тем более, должен справиться с более легкими заданиями. Это должно приводить к строгому разграничению единиц и нулей диагональю матрицы.

В действительности же это не совсем так. Например, в нашей бинарной матрице

профиль испытуемого №2 (№8 по старой нумерации) сильно отклоняется от правила Гуттмана. Этот испытуемый справился с самым трудным заданием №8 (№4 по старой нумерации), но не справился с более легкими заданиями №7 и №3. Профиль испытуемого искажен. Если бы единицы и нули поменялись местами, то есть испытуемый, верно ответил на трудные вопросы, но не справился с легкими, то говорят, у него *инвертированный* профиль.

Инвертированный профиль свидетельствует либо о неверной структуре знаний испытуемого, либо о нарушении процедуры тестирования (списывание, угадывание и т.д.), либо о недостатках тестовых заданий (по форме и (или) по содержанию.

3.3. ГРАФИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ ТЕСТОВЫХ БАЛЛОВ

Данные, представленные в таблице 3.2.5 удобно анализировать, используя их графическое представление. Для этого надо найти частоты тестовых баллов.

ЧАСТОТА ТЕСТОВОГО БАЛЛА - это количество испытуемых, имеющих данный тестовый балл.

Таблица 3.3.1. Частоты тестовых баллов.

X_i	1	2	3	4	5	6	7
Частота	2	2	1	1	1	2	1

Эта таблица построена следующим образом. В первой строке приведены все возможные значения индивидуальных баллов испытуемых. Во второй строке указаны частоты индивидуальных баллов. Например, тестовый балл (таблица 3.2.5) равный 1 имеют 2 испытуемых - №3 и №9 (по старой нумерации строк); 2 балла имеют также 2 испытуемых - №2 и №5; 3 балла - 1 испытуемый - №4; 4 балла - 1 испытуемый - №1; 5 баллов - 1 (№6); 6 баллов - 2 (№8 и №10); 7 баллов - 1 (№11).

На основании таблицы частот можно построить полигон частот (рис.3.3.1). По оси абсцисс отложены тестовые баллы, а по оси ординат - соответствующие частоты.

Рис.3.3.1. Полигон частот.

Вместо полигона частот иногда используют сглаженную кривую, проходящую максимально близко к экспериментальным точкам (рис.3.3.2).

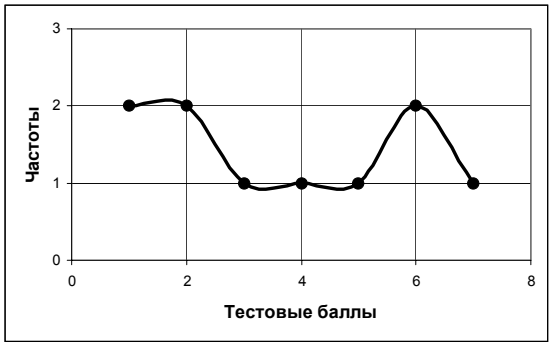


Рис.3.3.2. Сглаженная эмпирическая кривая.

Для графического представления данных можно также использовать гистограммы (рис.3.3.3).

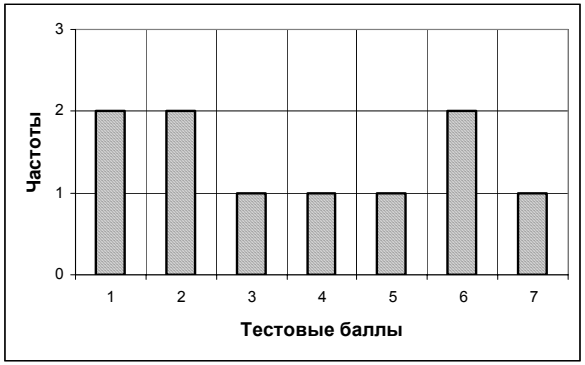


Рис.3.3.3. Гистограмма частот тестовых баллов.

Желательно, чтобы распределение частот тестовых баллов было близко к нормальному (Гауссовому).

3.4. МЕРЫ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ

Тестовые баллы испытуемых обычно группируются вблизи некоторых, наиболее вероятных значений, которые можно охарактеризовать тремя мерами центральной тенденции -модой, медианой и средним.

МОДА - это такое значение в множестве наблюдений, которое встречается наиболее часто[11].

Допустим 10 испытуемых получили следующие тестовые баллы:

Таблица 3.4.1

1	2	3	4	5	6	7	8	9	10
20	30	40	50	70	70	70	70	90	50

Мода будет равна Mo = 70, так как это значение повторяется чаще других (4 раза).

Соглашения об использовании моды[11].

1) Если все значения в группе встречаются одинаково часто, то мода отсутствует. Например, в группе (1, 1, 2, 2, 13, 13) моды нет.

2) Когда два соседних значения имеют одинаковые частоты и они больше частоты любого другого значения, мода есть среднее этих двух значений. Например, в группе (1, 2, 2, 5, 5, 5, 6, 6, 6, 9, 9, 10) мода равна 5,5.

3) Если два несмежных значения в группе имеют равные частоты и они больше частот любого другого значения, то существуют две моды. В этом случае говорят, что группа оценок является *бимодальной*. Например, в группе (1,4,4,4,7,7,9,9,9,10) модами являются 4 и 9. На рис.3.3.3 показано бимодальное распределение с модами 1,5 и 6.

Наибольшей модой в группе называется единственное значение, удовлетворяющее определению моды. Однако во всей группе может быть несколько меньших мод. Эти моды представляют собой локальные максимумы распределения частот.

МЕДИАНА - это значение, которое делит упорядоченное множество данных пополам, так что одна половина значений оказывается больше медианы, а другая - меньше.

Например, в группе (1,3,5,8,11,15,20) медианой будет 8. Если в группе четное число различных значений, то медиана находится посередине между двумя центральными значениями. В группе

(1,3,5,8,11,15) медианой будет 6,5. В сложных случаях, когда данные группируются вблизи медианы, придется использовать линейную интерполяцию.

СРЕДНЕЕ АРИФМЕТИЧЕСКОЕ определяется по формуле

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

где N - количество элементов в группе, X_i - величина i-го элемента группы. Например, в группе (1,3,5,8,11,15) среднее арифметическое будет равно (1+3+5+8+11+15) / 6 = 7,2.

Какую из мер центральной тенденции выбрать - решать исследователю*.

В педагогике очень часто в качестве меры центральной тенденции выбирается среднее арифметическое.

Найдем среднее арифметическое индивидуальных тестовых баллов из таблицы 3.2.5.

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{1}{10}\left(7+6+6+5+4+3+2+2+1+1\right) = 3.7$$

Это значение мы используем в дальнейшем.

* Рассмотрим пример вычисления «средней зарплаты». Пусть 7 человек имеют зарплату 500, 500, 500, 1000, 2000, 10 тыс. и 5 млн. руб. Если взять среднее арифметическое, то средняя зарплата равна 716357 руб. Если взять медиану, то средняя зарплата равна 1000 руб. Если взять моду, то 500 руб. Видимо медиана ближе к истине, чем среднее арифметическое.

3.5. НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Результаты нормативно-ориентированного тестирования при больших выборках обычно имеют распределение, близкое к нормальному.

Непрерывная случайная величина X имеет нормальный закон распределения (закон Гаусса) с параметрами μ и σ^2, если ее плотность вероятности имеет вид:

$$\varphi(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

где σ^2 - дисперсия, μ - константа, задающая сдвиг распределения по оси X (например, среднее арифметическое),

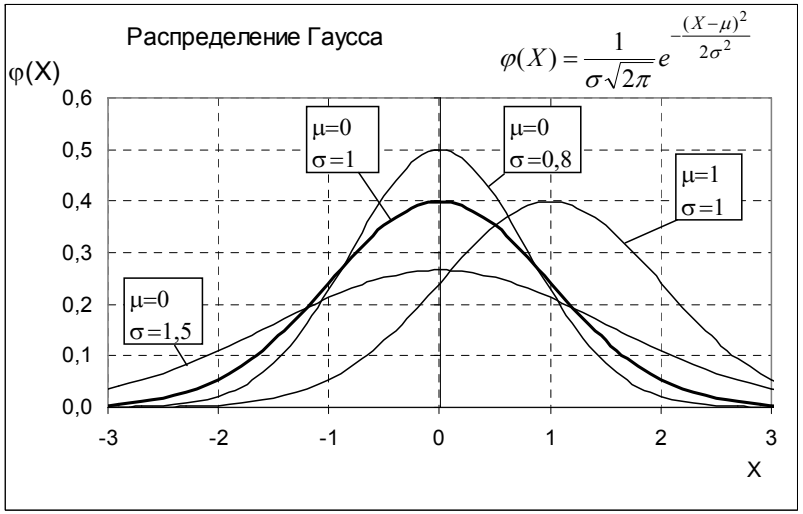


Рис.3.5.1. Распределение Гаусса (нормальное распределение).

Практически всем частотам соответствуют значения X от -3σ до $+3\sigma$ (рис.3.5.1).

Распределение Гаусса играет важную роль в статистической обработке результатов тестирования.

3.6. ДИСПЕРСИЯ ТЕСТОВЫХ БАЛЛОВ ИСПЫТУЕМЫХ

Нормативно-ориентированный тест должен хорошо дифференцировать испытуемых. Это означает, что индивидуальные тестовые баллы должны в достаточной степени отличаться друг от друга.

Вариацию тестовых результатов задают отклонения от среднего значения $\Delta = (X_i - \overline{X})$.

При полном совпадении всех индивидуальных баллов вариация равна нулю. Если индивидуальные баллы не совпадают, то отклонения могут быть положительными и отрицательными. Сумма всех отклонений будет равна нулю. Поэтому, чтобы охарактеризовать вариацию тестовых баллов используют квадрат отклонений. Сумма квадратов отклонений зависит от количества испытуемых N. Чтобы избавиться от этой зависимости, нам необходима обратно пропорциональная зависимость от N. В результате мы приходим к понятию дисперсии $s_x{}^2$.

$$s_x{}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2$$

Дисперсия пропорциональна не $1/N$, а $1/(N\text{-}1)$. Это сделано для того, чтобы для небольших N получить несмещенную оценку генеральной дисперсии[11].

Для удобства вычисления, преобразуем выражение для дисперсии.

$$\sum_{i=1}^{N}(X_i - \overline{X})^2 = \sum_{i=1}^{N}\left(X_i{}^2 - 2X_i\overline{X} + \overline{X}^2\right) = \sum_{i=1}^{N}X_i{}^2 - 2\overline{X}\sum_{i=1}^{N}X_i + \sum_{i=1}^{N}\overline{X}$$

Учтем, что

$$\sum_{i=1}^{n}X_i = n\overline{X}$$

$$\sum_{i=1}^{N}(X_i - \overline{X})^2 = \sum_{i=1}^{N}X_i{}^2 - 2\overline{X}\sum_{i=1}^{N}X_i + \sum_{i=1}^{N}\overline{X} = \sum_{i=1}^{N}X_i{}^2 - 2N\overline{X}^2 + N\overline{X}^2 = \sum_{i=1}^{N}X_i{}^2 - N\overline{X}^2 =$$

$$= \sum_{i=1}^{N}X_i{}^2 - N\left(\sum_{i=1}^{N}\frac{X_i}{N}\right)^2 = \sum_{i=1}^{N}X_i{}^2 - \frac{1}{N}\left(\sum_{i=1}^{N}X_i\right)^2 = \frac{1}{N}\left(N\sum_{i=1}^{N}X_i{}^2 - \left(\sum_{i=1}^{N}X_i\right)^2\right)$$

Используя полученное выражение, перепишем формулу для дисперсии

$$s_x{}^2 = \frac{1}{N(N-1)}\left(N\sum_{i=1}^{N} X_i{}^2 - \left(\sum_{i=1}^{N} X_i\right)^2 \right)$$

Подставим численные значения

$$s_x{}^2 = \frac{1}{10(10-1)}\left(10\sum_{i=1}^{10} X_i{}^2 - \left(\sum_{i=1}^{10} X_i\right)^2 \right) = \frac{1}{90}\left(10\cdot 181 - 37^2\right) = 4.9$$

Таким образом, дисперсия тестовых баллов в нашем примере равна 4.9. Подобные расчеты удобно проводить с использованием табличного процессора Microsoft Excel, входящего в стандартный офисный пакет. Для этого необходимо использовать статистическую функцию «ДИСП», для которой надо указать диапазон ячеек со значениями индивидуальных баллов испытуемых.

С дисперсией связан еще один важный параметр - стандартное отклонение

$$s_x = \sqrt{s_x{}^2} = \sqrt{4.9} = 2.214$$

Величина дисперсии тестовых баллов позволяет судить о качестве теста, о его дифференцирующей способности. Малая величина дисперсии говорит о том, что тест плохо различает испытуемых по уровню знаний, не позволяет с приемлемой точностью ранжировать их. Слишком большая дисперсия указывает на сильную неоднородность группы испытуемых, на возможные нарушения процедуры тестирования, на недостаточно ясные формулировки заданий и т.п. В случае оптимальной величины дисперсии, распределение тестовых баллов близко к нормальному.

М.Б.Челышкова[7] считает, что если среднее арифметическое примерно равно утроенному стандартному отклонению,

$$\overline{X} \approx 3s_x$$

то можно считать дисперсию оптимальной, а распределение тестовых баллов близким к нормальному.

Отметим, что это утверждение справедливо не для всех случаев. Возможны ситуации, когда среднее арифметическое гораздо больше утроенного стандартного отклонения, но распределение тестовых баллов, тем не менее, достаточно близко к нормальному.

Рассмотрим следующий модельный пример. Пусть в результате тестирования мы получили следующую таблицу частот.

Таблица 3.6.1

Баллы	77	78	79	80	81	82	83
Частота	1	2	7	10	6	2	1

Из таблицы видно, что средний тестовый балл равен 80.

Нормированная эмпирическая кривая распределения и нормальное распределение с дисперсией равной 1, показаны на рис.3.6.1.

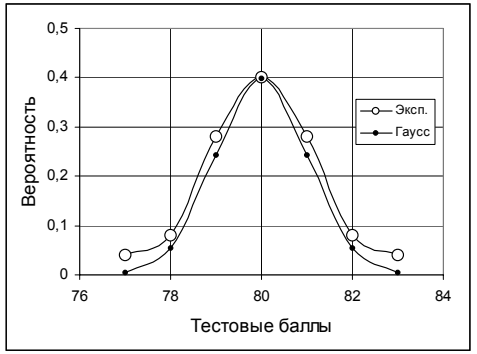


Рис.3.6.1. Эмпирическая кривая.

Легко видеть, что возможны такие эмпирические данные, когда кривая распределения будет почти гауссовой, но среднее арифметическое значение будет существенно превышать утроенное значение стандартного отклонения.

В качестве грубой оценки нормальности распределения можно рекомендовать проверку следующего соотношения:

$$\overline{X} - 3s_X \le X \le \overline{X} + 3s_X$$

- если почти все значения тестовых баллов X укладываются в этот интервал, то в первом приближении можно считать эмпирическое распределение нормальным.

Для корректного решения вопроса о степени близости эмпирических данных нормальному распределению необходимо использовать более строгие доказательства, например, проверить гипотезу о нормальном распределении генеральной совокупности по критерию Пирсона[12].

3.7. КОРРЕЛЯЦИОННАЯ МАТРИЦА

Тест, это не просто множество, а *система* тестовых заданий. Требование системности означает, что между заданиями существуют связи, которые можно обнаружить в результатах тестирования. Определение корреляции, как между заданиями, так и заданий с тестом в целом, позволит оценить системные качества теста. Благодаря такому анализу можно будет выполнить «чистку» - избавить тест от заданий, нарушающих его системные свойства.

Если две величины связаны между собой, то между ними есть корреляция. Виды корреляционной связи показаны в таблице 3.9.

Для выяснения вопроса о наличии связи между двумя величинами X и Y необходимо определить, существует ли соответствие между большими и малыми значениями X и соответствующими значениями Y или такой связи не обнаруживается. Значение каждого элемента X_i и Y_i определяется величиной и знаком отклонения от среднего арифметического[11]:

$$(X_i - \overline{X}) \cdot (Y_i - \overline{Y})$$

Если большие значения X_i соответствуют большим значениям Y_i, то это произведение будет большим и положительным, так как

$$X_i > \overline{X} \qquad \text{и} \qquad Y_i > \overline{Y}$$

То же самое будет наблюдаться и, когда малые значения X_i будут соответствовать малым Y_i, поскольку произведение отрицательных чисел будет положительным.

Если же большие значения X_i соответствуют малым значениям Y_i, то это произведение будет большим и отрицательным, что будет свидетельствовать об обратной зависимости между этими величинами.

В тех случаях, когда нет систематического соответствия больших значений X_i большим или малым Y_i, то знак произведения будет положительным или отрицательным для разных пар X_i и Y_i. Тогда сумма

$$\sum_{i=1}^{N} (X_i - \overline{X}) \cdot (Y_i - \overline{Y})$$

будет близка к нулю. Таким образом, эта сумма велика и положительна, когда X и Y сильно связаны прямой зависимостью,

близка к нулю в случае отсутствия связи и велика и отрицательна, когда X и Y сильно связаны обратной зависимостью[11].

Для того, чтобы эта сумма не зависела от количества значений X и Y, ее следует поделить ее на N-1. Полученная величина s_{XY} называется ковариацией X и Y и является мерой их связи:

$$s_{XY} = \frac{\sum_{i=1}^{N} (X_i - \overline{X}) \cdot (Y_i - \overline{Y})}{N - 1}$$

Для исключения влияния стандартных отклонений на величину связи, следует поделить ковариацию s_{XY} на стандартные отклонения s_X и s_Y:

$$r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y}$$

Полученная мера связи между X и Y называется коэффициентом корреляции Пирсона. Обозначение r происходит от слова *регрессия*. Подставив соответствующие выражения, получим формулу для коэффициента корреляции Пирсона r_{XY}[11]

$$r_{XY} = \frac{\sum_{i=1}^{N} (X_i - \overline{X}) \cdot (Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N} (X_i - \overline{X})^2} \sqrt{\sum_{i=1}^{N} (Y_i - \overline{Y})^2}}$$

Для вычислений более удобна следующая формула

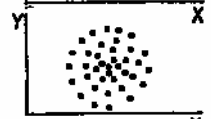
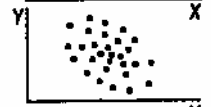
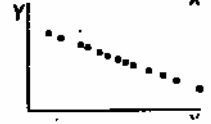
$$r_{XY} = \frac{N \sum_{i=1}^{N} X_i Y_i - \left(\sum_{i=1}^{N} X_i \right) \cdot \left(\sum_{i=1}^{N} Y_i \right)}{\sqrt{\left(N \sum_{i=1}^{N} X_i^2 - \left(\sum_{i=1}^{N} X_i \right)^2 \right) \cdot \left(N \sum_{i=1}^{N} Y_i^2 - \left(\sum_{i=1}^{N} Y_i \right)^2 \right)}}$$

Коэффициент корреляции Пирсона r_{XY} изменяется в пределах от -1 до +1. В таблице 3.7.1 приведены различные виды линейной зависимости и соответствующие значения r_{XY}.

Следует отметить, что в случае нелинейной связи между X и Y коэффициент корреляции может оказаться близким к нулю, даже если связь очень сильная.

Таблица 3.7.1. Типы корреляционной связи
(Гласс Дж., Стэнли Дж., 1976).

Интерпретация значений r_{xy}

Величина r_{xy}	Описание линейной связи	Диаграмма рассеивания
+1,00	Строгая прямая связь	
Около +0,50	Слабая прямая связь	
0,00	Нет связи (то есть ковариация X и $Y = 0$)	
Около −0,50	Слабая обратная связь	
−1,00	Строгая обратная связь	

Для решения вопроса о наличии связи между заданиями теста, надо, используя данные по столбцам из бинарной матрицы, рассчитать коэффициенты корреляции Пирсона для каждой пары заданий. Для расчетов используются различные статистические программы (SPSS, STATISTICA и др.). В простейшем случае можно использовать табличный процессор Excel с вызовом функции «ПИРСОН».

В случае дихотомического оценивания (1 - верно, 0 –неверно) выражение для коэффициента корреляции упрощается. Введем следующие обозначения:

p_m – доля верных ответов для задания с номером m;

q_m – доля неверных ответов для задания с номером m;

p_k – доля верных ответов для задания k;

q_k – доля неверных ответов для задания с номером k;

p_{mk} – доля верных ответов для задания с номером m и k.

Коэффициент корреляции Пирсона, для дихотомических данных называется *коэффициентом* «фи». Коэффициент ϕ_{mk}, описывающий связь между заданиями с номерами m и k записывается следующим образом[11]

$$\varphi_{mk} = \frac{p_{mk} - p_m p_k}{\sqrt{p_m q_m p_k q_k}}$$

Отметим, что коэффициент «фи» и коэффициент корреляции Пирсона дают в результате одно и то же значение, поскольку обе формулы эквивалентны. Рассмотрим пример вычисления коэффициента корреляции между 2-м и 5-м заданиями. Из таблицы 3.2.5 имеем: p_2=0.7, q_2=0.3, p_5=0.5, q_5=0.5. Для определения p_{25} надо подсчитать количество верных ответов на оба задания одновременно. Видно, что испытуемые с номерами 1-5 успешно справились с обоими заданиями (5 верных ответов). Испытуемые 6 и 7 правильно ответили на 2-е задание, но неправильно на 5-е (нет одновременно верных ответов). Испытуемые 8 и 9 не справились и со 2-м и с 5-м заданиями. Таким образом, p_{25} =5/10 = 0,5.

$$\varphi_{25} = \frac{p_{25} - p_2 p_5}{\sqrt{p_2 q_2 p_5 q_5}} = \frac{0.5 - 0.7 \cdot 0.5}{\sqrt{0.7 \cdot 0.3 \cdot 0.5 \cdot 0.5}} = 0.655$$

Результаты расчетов для всех заданий приведены в корреляционной матрице (таблица 3.7.2). Корреляционная матрица представляет собой квадратную матрицу размерности MxM, где M – количество заданий, симметричную относительно главной диагонали. В нашем примере матрица имеет 8 строк и столько же столбцов. Коэффициент корреляции Пирсона, скажем, между 2-м и 5-м заданиями находится на пересечении 2-й строки и 5-го столбца (0,655).

В самом последнем столбце располагается коэффициент корреляции каждого задания с тестовым баллом испытуемого (индивидуальным баллом) – r_{pb} – точечный бисериальный

коэффициент корреляции.

ТАБЛИЦА 3.7.2. Корреляционная матрица тестовых заданий.

	1	2	3	4	5	6	7	8	r_{pb}
1	1,000	0,524	-0,089	0,218	0,655	0,089	0,327	0,218	0,634
2	0,524	1,000	0,356	0,218	0,655	0,089	0,327	0,218	0,738
3	-0,089	0,356	1,000	0,000	0,000	-0,167	-0,102	-0,408	0,175
4	0,218	0,218	0,000	1,000	0,600	0,408	0,500	0,333	0,714
5	0,655	0,655	0,000	0,600	1,000	0,408	0,500	0,333	0,905
6	0,089	0,089	-0,167	0,408	0,408	1,000	0,102	0,408	0,505
7	0,327	0,327	-0,102	0,500	0,500	0,102	1,000	-0,167	0,548
8	0,218	0,218	-0,408	0,333	0,333	0,408	-0,167	1,000	0,365
$\sum r_{x_m x_k}$	2,942	3,388	0,590	3,278	4,151	2,338	2,488	1,936	4,584
$\overline{r}_{x_m x_k}$	0.368	0.423	0.074	0.410	0.519	0.292	0.311	0.242	0.573

Поскольку результаты выполнения тестовых заданий размещаются на дихотомической шкале, а индивидуальный балл испытуемого на интервальной, то формула для коэффициента корреляции Пирсона упрощается и преобразуется в r_{pb}. Выражение для точечного бисериального коэффициента корреляции имеет вид[11]

$$r_{pb} = \frac{\overline{X}_1 - \overline{X}_0}{s_x}\sqrt{\frac{n_1 \cdot n_0}{n(n-1)}}$$

где $\bar{X}_1$ - средний индивидуальный балл испытуемых, справившихся с данным заданием.

$\bar{X}_0$ - средний индивидуальный балл испытуемых, не справившихся с данным заданием.

n_1 – число испытуемых, выполнивших данное задание, n_0 – число испытуемых, не выполнивших его. $n = n_1 + n_0$ – общее количество испытуемых; s_x - стандартное отклонение для индивидуальных баллов всех испытуемых.

В нижних строках таблицы 3.7.2 приведены суммарные и среднее значения коэффициента корреляции для каждого задания.

Коэффициент корреляции r_{pb} очень важен, так как характеризует валидность отдельных заданий. Необходимо стремиться к тому, чтобы корреляция результатов по заданию и индивидуальными баллами была достаточно высокой. В.С.Аванесов[6] дает следующую рекомендацию: $r_{pb} \geq 0{,}5$.

Корреляция заданий друг с другом не должна быть слишком высокой ($r_{xy} \leq 0{,}3$), иначе задания начинают дублировать друг друга[6]. Если корреляция между двумя заданиями близка к единице, то одно из них лишнее.

Отрицательная корреляция задания с другими заданиями нежелательна. Если задание отрицательно коррелирует с большим количеством других заданий, то это означает, что исход ответов на него противоположен результатам по другим заданиям. По всей вероятности у такого задания либо имеются грубые ошибки в содержании и (или) оформлении (например, нет правильного ответа), либо проверяются знания из другой предметной области.

В нашем примере отрицательной корреляцией отличаются задания 1, 3, 6, 7, 8. Обращает на себя внимание то, что отрицательная корреляция у заданий 1, 6, 7 и 8 наблюдается именно с заданием 3. Это наводит на мысль, что проблематичным является задание 3. В пользу этого свидетельствует и самый низкий средний коэффициент корреляции (0,074) и, самое главное, низкая корреляция с индивидуальными баллами испытуемых (r_{pb} =0.175). Задание 3 следует удалить из теста. В результате отрицательная корреляция останется между 7 и 8 заданиями. Задание 8 находится под подозрением, так как у него r_{pb} =0.365. Это задание также следует удалить из теста. Если какое-либо задание отрицательно коррелирует с индивидуальными баллами ($r_{pb} < 0$), то такое задание, безусловно, подлежит удалению.

3.8. НАДЕЖНОСТЬ ТЕСТА

Важнейшей характеристикой теста является его надежность, определяющая воспроизводимость результатов тестирования, их точность. Допустим, у нас есть гипотетическая группа испытуемых, которые немедленно забывают содержание теста по его завершении. Тогда, в случае надежного теста, повторяя тестирование многократно, мы должны получать одни и те же индивидуальные баллы. Для малонадежного теста результаты будут меняться каждый раз.

Тест представляет собой систему заданий. Качество заданий определяет надежность теста в целом. Рассмотрим пример. Допустим, тест состоит из заданий в закрытой форме, в которых по ошибке не указанные правильные ответы. Слабые испытуемые, не зная ответа, будут пытаться его угадать. Сильные испытуемые, зная верный ответ, но не находя его среди предложенных, так же вынуждены будут случайным образом выбирать любой из ответов. В итоге, индивидуальные баллы будут представлять собой случайные последовательности, не повторяющиеся в разных сеансах тестирования. Воспроизводимость тестовых баллов будет полностью отсутствовать и надежность теста будет близка к нулю. Низкая надежность теста обусловлена низким качеством тестовых заданий.

Для определения надежности реальных тестов можно использовать коэффициент корреляции Пирсона для индивидуальных баллов разных сеансов тестирования. Для организации разных сеансов тестирования можно использовать либо параллельные тесты, либо повторное тестирование через определенный промежуток времени. Можно также использовать результаты одного сеанса тестирования. При этом выполняют расщепление теста, например, на четные и нечетные задания и, затем, находят корреляцию между этими двумя половинами.

Надежность теста определяется разными методами. Рассмотрим их.

Из классической теории теста следует, что надежность теста есть

$$r_t = 1 - \frac{{s_E}^2}{{s_X}^2} \qquad (3.8.1)$$

где S_E^2 - дисперсия ошибочного вклада тестовый балл, S_E^2 - дисперсия наблюдаемого тестового балла.

Когда ошибка отсутствует, коэффициент надежности равен единице. Если измеренный тестовый балл полностью обусловлен

ошибкой измерения, то надежность теста равна нулю.

Ошибка измерения зависит от надежности теста r_t.

$$s_E = s_X \sqrt{1 - r_t} \qquad (3.8.2$$

В работе[13] показано, что корреляция r_{jT} j-го задания с истинными тестовыми баллами T связана со средним значением его корреляции с другими заданиями теста[7]

$$r_{jT} = \sqrt{\bar{r}_j} \qquad (3.8.3)$$

Если тест содержит задания с высокой внутренней корреляцией, то он будет высоко надежным и ошибка измерений будет низкой.

Определение надежности теста необходимо выполнять на специально подобранной выборке испытуемых, репрезентативно представляющей всю генеральную совокупность. Выборка должна быть достаточно большой - 200-300 человек. Чем больше выборка, тем точнее определяется надежность теста.

Для вычисления надежности теста нужны результаты двух испытаний, которые организуются следующими способами:

1-й способ – тестирование с помощью двух параллельных тестов (parallel-form reliability);

2-й способ – повторное тестирование с помощью одного и того же теста (test-retest reliability);

3-й способ – расщепление теста (split-half method).

Первый способ, пожалуй, самый лучший, с точки зрения расчета надежности. Основной проблемой здесь является разработка параллельных тестов. Крайне сложно создать тесты параллельные и по содержанию и по результатам. Ранее нами приводился пример «параллельных» заданий, дающих разные результаты:

6 + 3 = ____; 6 + ___ = 9; ___ + 3 = 9.

Второй способ технически гораздо проще, однако здесь появляются новые факторы.

Во-первых, первое тестирование изменяет уровень подготовленности испытуемых. Это может произойти по разным причинам, в частности, запоминание заданий теста. Поэтому повторное тестирование необходимо проводить спустя некоторый интервал времени. Этот интервал должен быть как можно больше.

Во-вторых, к моменту повторного тестирования изменяются внешние условия – другие социальная среда, другие взаимодействия с членами микросоциальной группы, другое время года, и т.д. Кроме того, изменились и сами испытуемые, изменился их уровень знаний как специальных, так общекультурных. В результате повторное тестирование проводится в иных условиях и иной группе испытуемых. В этой связи желательно временной интервал между тестированиями выбирать как можно короче. Мы получили взаимоисключающие требования к интервалу повтора тестирования, следовательно, здесь придется идти на компромисс. Можно рекомендовать интервал в один месяц, хотя подобные рекомендации должны подтверждаться экспериментально.

Надо осознавать, что повторное тестирование в силу указанных причин, в принципе не позволяет получить параллельные результаты даже для идеального теста с надежностью равной единице.

Третий способ очень прост. На основании всего лишь одного тестирования мы можем оценить надежность теста. Полученные результаты тем или иным способом делятся на две группы. Например, в первую входят результаты по четным заданиям, во вторую – результаты по нечетным заданиям. Затем вычисляется коэффициент корреляции между этими группами. Недостаток этого способа обусловлен неидентичностью этих групп.

В качестве примера проанализируем надежность четырех гипотетических тестов, выполненных на одной и той же выборке испытуемых (таблица 3.8.1).

Таблица 3.8.1. Индивидуальные баллы по четырем тестам.

	Тест 1		Тест 2		Тест 3		Тест 4	
ФИО	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
1	80	80	80	70	80	70	80	20
2	70	70	70	80	70	20	70	30
3	60	60	60	60	60	40	60	40
4	50	50	50	50	50	80	50	20
5	40	40	40	20	40	20	40	35
6	30	30	30	30	30	45	30	45
7	20	20	20	30	20	50	20	80
	r_t	1.00	r_t	0,884	r_t	0,101	r_t	-0,769

Испытуемых – 7 человек. Каждый тест проводился два раза, индивидуальные баллы испытуемых приведены в столбцах X_1 и X_2.

В последней строке приведены значения надежности теста (коэффициента корреляции Пирсона для совокупностей X_1 и X_2).

Тест 1. Индивидуальные баллы полностью совпадают. Надежность теста r_t=1. Это идеальный случай, на практике не достижим.

Тест 2. Индивидуальные баллы различные, но наблюдается некоторое согласие. Большим и малым значениям X_1 приблизительно соответствуют большие и малые значения X_2. Тест обладает довольно высокой надежностью r_t=0,884.

Тест 3. Между результатами обоих тестирований отсутствует какая-либо связь. Надежность теста низкая (r_t=0,101), тест непригоден к использованию.

Тест 4. Между результатами обоих тестирований есть довольно сильная, но отрицательная корреляция (r_t= -0,769). Такой тест также нельзя использовать.

Тест можно использовать, если его коэффициент надежности не менее +0,7.

Приведем формулу для расчета коэффициента надежности при двукратном тестировании (параллельном или повторном)

$$r_t = \frac{N\sum_{i=1}^{N} X_i Y_i - \left(\sum_{i=1}^{N} X_i\right)\cdot\left(\sum_{i=1}^{N} Y_i\right)}{\sqrt{\left(N\sum_{i=1}^{N} X_i^2 - \left(\sum_{i=1}^{N} X_i\right)^2\right)\cdot\left(N\sum_{i=1}^{N} Y_i^2 - \left(\sum_{i=1}^{N} Y_i\right)^2\right)}} \qquad (3.8.4)$$

X_i и Y_i– индивидуальные баллы i-го испытуемого в первом и во втором тестированиях; N –количество испытуемых;

На рис.3.8.1 приведена графическая интерпретация полученных коэффициентов надежности всех четырех тестов.

Рассмотрим теперь пример вычисления надежности теста методом расщепления. Используем бинарную матрицу из таблицы 3.2.5. Уберем из нее старые номера заданий и испытуемых (таблица 3.7.2).

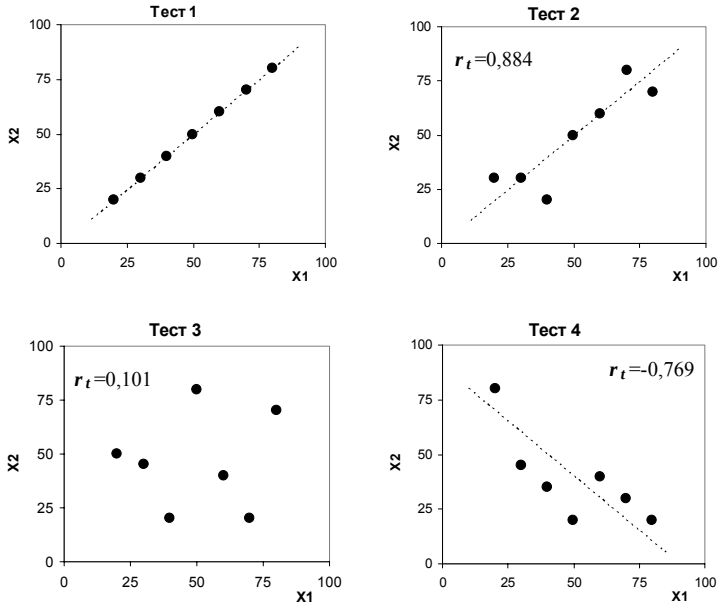


Рис.3.8.1. Графическая интерпретация надежности теста.

Таблица 3.8.2. бинарная матрица 10x8.

	1	2	3	4	5	6	7	8	X_i	Чет.	Нечет.
1	1	1	1	1	1	1	1	0	7	3	4
2	1	1	0	1	1	1	0	1	6	4	2
3	1	1	1	1	1	1	0	0	6	3	3
4	1	1	0	1	1	0	1	0	5	2	3
5	1	1	1	0	1	0	0	0	4	1	3
6	1	1	1	0	0	0	0	0	3	1	2
7	0	1	1	0	0	0	0	0	2	1	1
8	0	0	1	1	0	0	0	0	2	1	1
9	0	0	0	0	0	1	0	0	1	1	0
10	1	0	0	0	0	0	0	0	1	0	1

В последних столбцах таблицы приведены индивидуальные баллы по четным и нечетным заданиям. Например, для испытуемого №1 количество верных ответов в четных заданиях равно 3, а в

нечетных -4. Всего 7, что соответствует его индивидуальному баллу по всем заданиям.

Коэффициент надежности находим по формуле (3.7.3). В качестве X_i и Y_i используются соответственно данные из столбцов «Чет» и «Нечет» соответственно. Вычисления дают для коэффициента надежности следующее значение: r_t=0,569.

Поскольку для определения надежности использовалась лишь половина теста, то полученное значение r_t является заниженным. Для коррекции значения r_t используется формула Спирмена-Брауна

$$r_t' = \frac{2r_t}{1+r_t} \qquad (3.8.5)$$

где r_t' – исправленный коэффициент надежности; r_t – коэффициент надежности по половинкам расщепленного теста.

В нашем случае r_t' = 2·0,569/(1+0,569) = 0,725. Исправленное значение показывает удовлетворительную надежность теста (больше +0,7).

Другой способ определения надежности теста основан на использовании среднего коэффициента корреляции всех заданий между собой:

$$r_t = \frac{M\overline{R}}{1+(M-1)\overline{R}} \qquad (3.8.6)$$

Здесь M – количество заданий в тесте.

Из таблицы 3.7.2 следует, что

$$\overline{R} = \frac{0{,}368+0{,}423+0{,}074+0{,}410+0{,}519+0{,}292+0{,}311+0{,}242}{8} = 0{,}33$$

Тогда

$$r_t = \frac{8\cdot 0{,}33}{1+(8-1)\cdot 0{,}33} = 0{,}798$$

Приведем еще одну формулу, позволяющую рассчитать надежность теста по вариации тестового задания p_jq_j.

Эта формула носит название KR-20 (F.Kuder & M.Richardson)[14] – по имени ее создателей, число 20 – это номер формулы.

$$r_t = \frac{M}{M-1}\left(1 - \frac{\sum_{j=1}^{M} p_j q_j}{{s_X}^2}\right) \qquad (3.8.7)$$

где M – количество заданий, ${s_X}^2$ – дисперсия индивидуальных баллов испытуемых. Ранее, для дисперсии было получено значение ${s_X}^2$ =4,9.

Расчеты по таблице 3.2.5 дают

$$r_t = \frac{8}{8-1}\left(1 - \frac{0{,}21+0{,}21+0{,}24+0{,}25+0{,}25+0{,}24+0{,}16+0{,}09}{4{,}9} = 0{,}758\right)$$

Как видим, вычисления надежности по формулам (3.7.4), (3.7.5), (3.7.6) дают примерно одинаковые результаты.

Выше указывалось, что чем длиннее тест (чем больше в нем заданий) тем он надежнее (при прочих равных условиях). Формула Спирмена-Брауна позволяет оценить требуемую длину теста для заданного значения надежности.

Коэффициент надежности r_{tk} после изменения длины теста равен[15]

$$r_{tk} = \frac{k r_t}{1 + (k-1) r_t} \qquad (3.8.7)$$

где k-кратность измерения длины теста; r_t – коэффициент надежности до изменения длины теста.

Пусть начальная надежность теста равна 0,758 и количество заданий в тесте увеличивается в два раза. Тогда надежность нового теста равна:

$$r_{tk} = \frac{2 \cdot 0.758}{1 + (2-1) \cdot 0.758} = 0.862$$

Рассмотрим теперь, обратную задачу. Пусть начальная надежность теста равна 0,758 и мы хотим достигнуть надежности 0.862. Во сколько раз надо увеличить длину теста? Для расчетов воспользуемся формулой

$$k = \frac{r_{tk}(1 - r_t)}{r_t(1 - r_{tk})}$$

В нашем примере

$$k = \frac{0.862 \cdot (1 - 0.758)}{0.758 \cdot (1 - 0.862)} = 1.994 \approx 2$$

То есть длину теста надо увеличить в два раза.

Рассмотрим теперь вопрос об определении истинного балла испытуемого. Используя регрессионное уравнение[11], получим выражение[7]

$$T_i = \overline{X} + r_t (X_i - \overline{X})$$

Влияние r_t на T_i показано на графиках (рис.3.8.2).

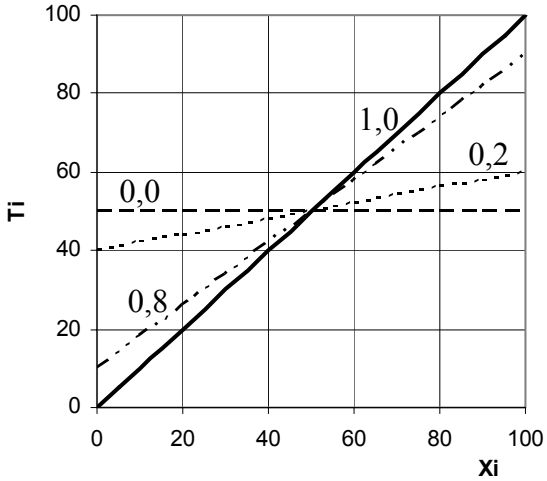


Рис.3.8.2. Влияние надежности теста на истинный балл.

Графики построены в предположении, что средний индивидуальный балл равен 50. Значения коэффициента надежности показаны возле соответствующих зависимостей.

При r_t =1 наблюдаемый X_i и истинный T_i баллы совпадают. Этому случаю соответствует прямая линия, проходящая через начало координат под углом 45^o к осям. Если надежность теста равна нулю, то определить истинный балл нельзя, для всех испытуемых получается одно и то же значение, равное среднему баллу. Соответственно, график представляет собой горизонтальную прямую, проходящую на уровне 50 баллов. При других значениях r_t получаются промежуточные случаи. На рисунке показаны графики для r_t =0,2 и r_t =0,8. Все зависимости образованы поворотом прямой линии относительно точки закрепления с координатами (50; 50).

Из приведенных графиков видно, что если наблюдаемый балл испытуемого меньше среднего, то $X_i < T_i$. Если же наблюдаемый балл больше среднего, то $X_i > T_i$. Иными словами, наблюдаемый балл у слабых испытуемых меньше, а у сильных - больше истинного индивидуального балла.

ОЦЕНКА ДОВЕРИТЕЛЬНОГО ИНТЕРВАЛА

Надежность теста определяет ошибку измерения индивидуального балла испытуемого, что позволяет найти стандартную ошибку измерения

$$s_E = s_X\sqrt{1-r_t}$$

Рассмотрим пример. Ранее, для модельной бинарной матрицы (таблица 3.2.5) нами было вычислено стандартное отклонение S_X = 2,214. Коэффициент надежности для этой же матрицы, рассчитанный по формуле Спирмена-Брауна, равен r_t= 0, 725. Тогда для стандартной ошибки измерения получим

$$S_E = 2{,}214 \cdot \sqrt{1-0{,}725} = 1{,}161$$

Найдем оценку доверительного интервала для доверительной вероятности α = 0,05. Предположим, что середина доверительного интервала совпадает с X_i, а не с T_i. Это, конечно, не так, но мы предположим, что наблюдаемый и истинный тестовый баллы не сильно отличаются. Это вполне справедливо для надежных тестов. Наше допущение приведет к сдвигу границ доверительного интервала, что вызовет ошибку в определении области локализации истинного тестового балла.

Ошибка, допускаемая при этом, получается приемлемой. Тогда половина доверительного интервала равна

$$\delta X_i = 1{,}96 S_E = 1{,}96 \cdot 1{,}161 = 2{,}27$$

Теперь найдем границы тестового балла, например, для второго испытуемого X_2 =6 (таблица 3.8.2). Минимальное значение равно 6-2,27=3,73 ≈ 4. Максимальное равно 6+2,27=8,27 ≈ 8. Следовательно, истинный балл испытуемого №2 находится в промежутке от 4 до 8 баллов.

Как видим, вопросу определения надежности теста, необходимо уделять самое пристальное внимание. Созданный на скорую руку «тест» - таковым не является. Это всего лишь совокупность заданий. В лучшем случае, это совокупность *заданий в тестовой форме*. Только статистическая проверка теста позволяет превратить его в *систему тестовых заданий*. Только указание его *надежности*, позволяет адекватно трактовать результаты тестирования.

Таким образом, вопросы определения надежности теста, его стандартной ошибки, области локализации истинного тестового балла очень важны для создания качественного педагогического теста и его дальнейшей сертификации.

3.9. ВАЛИДНОСТЬ ТЕСТА

Высокая надежность теста это необходимое, но недостаточное условие получения высококачественного теста. Тест еще должен быть валидным. Валидность – это важнейшая характеристика теста, без указания которой, его нельзя считать измерительным инструментом.

Анализируя сложную ситуацию с валидностью педагогических тестов, Е.Михайлычев[16] отмечает, что педагогу, заинтересовавшемуся валидностью, трудно будет разобраться в том, что же это такое.

Ниже мы приведем несколько определений валидности теста.

ВАЛИДНОСТЬ означает пригодность тестовых результатов для той цели, ради чего проводилось тестирование (В.Аванесов)[17].

ВАЛИДНОСТЬ - это характеристика способности теста служить поставленной цели измерения (М.Челышкова)[7].

ВАЛИДНОСТЬ - определяет, насколько тест отражает то, что он должен оценивать (А.Майоров)[18].

Приведенные определения в целом перекликаются и являются практически равноценными. Мы несколько уточним определение, сделав акцент на цель тестирования. Тестирование как измерительная процедура, дает информацию, на основе которой в дальнейшем должно быть принято то или иное управленческое решение. Обоснованность этих решений, зачастую сильно влияющих на судьбу испытуемых, определяется надежностью и валидностью теста.

ВАЛИДНОСТЬ – это характеристика теста, отражающая его способность получать результаты, соответствующие поставленной цели и обосновывающая адекватность принимаемых решений.

После создания теста начинается процесс его валидизации. Приведем определение:

ВАЛИДИЗАЦИЯ – процесс накопления подтверждений для доказательства валидности теста[19].

По нашему мнению ВАЛИДИЗАЦИЯ – это не столько сбор доказательств валидности теста, сколько процесс выполнения действий, повышающих его валидность. Вследствие этого будет расти и доказательная база валидности теста.

Выделяют три вида валидности – *содержательную*, *критериальную* и *конструктную*[20]. А.Майоров приводит следующую диаграмму видов валидности[18]:

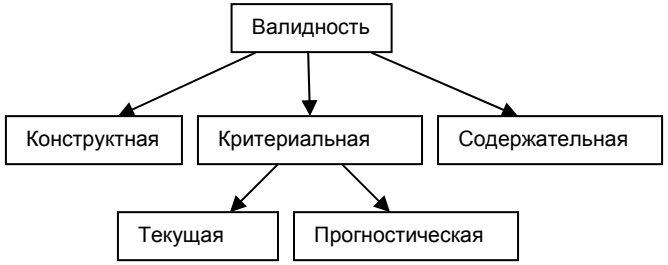


Рис.3.9.1. Виды валидности.

КОНСТРУКТНАЯ ВАЛИДНОСТЬ (концептуальная валидность) определяется в случаях, когда представление об измеряемом свойстве существует в форме абстрактного образа, модели. Для объяснения определенных качеств личности создается концептуальная модель, которая с помощью тестов подтверждается или опровергается.

КРИТЕРИАЛЬНАЯ ВАЛИДНОСТЬ (эмпирическая валидность) предполагает наличие внешнего критерия, корреляция с которым определяет валидность теста.

Имеется два вида критериальной валидности – текущая и прогностическая.

Текущая критериальная валидность (concurrent validity) характеризует способность теста измерять некоторые качества личности. Валидность теста подтверждается корреляцией с некоторым внешним критерием, существующим в данное время. Допустим, тест показал для некоторого испытуемого отличные знания по предмету, а школьные отметки, выставленные учителем – неудовлетворительные. Если мы в качестве внешнего, независимого и достоверного критерия выберем школьные отметки, то критериальная валидность теста – низкая, даже если он имеет высокую надежность.

Прогностическая критериальная валидность (predictive validity) характеризует способность теста предсказывать будущие качества, формирующихся в результате воздействия внешних обстоятельств или целенаправленной собственной деятельности. Этот тип валидности характеризует корреляцию результатов тестирования с внешним критерием, который появится в будущем.

СОДЕРЖАТЕЛЬНАЯ ВАЛИДНОСТЬ (content validity)

характеризует тест по степени его соответствия предметной области.

Согласно А.Анастази, содержательная валидность означает систематическую проверку содержания теста, с тем чтобы установить, соответствует ли оно репрезентативной выборке измеряемой области поведения. Такая процедура валидизации обычно применяется для тестов достижений[20].

Содержательная валидность необязятельно означает полноту отображения изучаемой дисциплины. Например, для нормативно-ориентированного теста, полнота охвата всех тем может быть меньше, чем для критериально-ориентированного. Здесь важнее глубина проработки отдельных подтем, вопросов. Это позволит с большей эффективность дифференцировать обучаемых. Под содержанием понимается не только совокупность фактов, понятий, терминов, но и умение применять имеющиеся знания, оценивать информацию, выполнять действия, соответствующие верхним уровням таксономии Блума.

Для обеспечения содержательной валидности необходим детальный анализ учебных программ, на основании чего составляется *спецификация* теста. Спецификация содержит перечень учебных тем, их важность, количество и тип тестовых заданий. Оценка содержательной валидности выполняется *экспертом* в данной предметной области.

Согласно П.Клайну содержательная валидность определяется следующим образом:

1) указать категорию лиц, для которой предназначен тест;

2) составить список знаний, умений, навыков, подлежащих тестированию;

3) выполнить внешнюю экспертизу полученного списка на предмет его полноты и обоснованности;

4) на основе списка составить перечень заданий;

5) выполнить внешнюю экспертизу полученных заданий;

6) после проверки преобразовать их в задания в *тестовой форме*. В дальнейшем, на этой основе создать *тестовые задания*, образующие тест, который *будет содержательно* валидным.

Проблема валидизации педагогического теста является, видимо, самой сложной в процедуре создания высококачественного измерительного инструмента.

Литература к главе 3

[1] Spearman C. Correlation calculated from faulty data //British Journal of Psychology, 1910, Vol.3, N2. -P.271-295.

[2] Richard H.Williams, Donald W.Zimmerman, Bruno D.Zumbo, Donald Ross. Charles Spearman: British Behavioral Scientist. //Human Nature Review, 2003, N3. – p.114-118.

[3] Guttman L. A basis for analyzing test-retest reliability. // Psychometrika, 1945, 10. P.255-282.

[4] Gulliksen H. Theory of mental tests. – N-Y. Willey, 1950. -486 pp.

[5] Crocker Linda, Algina James. Introduction to Classical and Modern Test Theory. –New-York: Harcourt Brace Jovanovich, 1986

[6] Аванесов В.С. Основы научной организации педагогического контроля в высшей школе. -М., 1989. -167 с.

[7] Челышкова М.Б. Теория и практика конструирования педагогических тестов: Учебное пособие. –М.: Логос, 2002. -432 с.

[8] Аванесов В.С. Основы научной организации педагогического контроля в высшей школе. -М., 1989. -167 с.

[9] Gulliksen H. Theory of mental tests. – N-Y. Willey, 1950. -486 pp.

[10] Lord F.M., Novick M. Statistical Theories of Mental Test Scoires. Addison-Westley Publ. Co. Reading, Mass. 1968. -560 pp.

[11] Глас Дж., Стэнли Дж. Статистические методы в педагогике и психологии. -М.: Прогресс, 1976. -495 с.

[12] Гмурман В.Е. Теория вероятностей и математическая статистика. Учеб. пособие для вузов. Изд. 7-е, стер. –М.: ВШ, 1999. -479 с.

[13] Клайн П. Введение в психометрическое проектирование. Справочное руководство по конструированию тестов. –Киев: ПАН Лтд, 1994. -184 с.

[14] Kuder, G.F., Richardson, M.W. The theory of the estimation of test reliability // Psychometrika, 1937, v.2, N3. -p.151-160.

[15] Кларин М.В. Инновационные модели обучения в зарубежных педагогических поисках. -М.: Арена, 1994. - 223 с.

[16] Михайлычев Е.А. Дидактическая тестология. -М.: Народное образование, 2001. -432 с.

[17] Аванесов В.С. Композиция тестовых заданий. Учебная книга для преподавателей вузов, учителей школ, аспирантов и студентов педвузов. 2 изд., испр. и доп. М.: Адепт 1998. -217с.

[18] Майоров А.Н. – Теория и практика создания тестов для системы образования. – М.: «Интеллект-центр», 2001. -296 с.

[19] Переверзев В.Ю. Технология разработки тестовых заданий: справочное руководство. –М.: Е-Медиа, 2005. -265 с.

[20] Анастази А., Урбина С. Психологическое тестирование. –Спб.: Питер, 2006. -688 с.

ГЛАВА 4. ТЕСТИРОВАНИЕ УЧЕБНЫХ ДОСТИЖЕНИЙ

В этой главе мы рассмотрим некоторые вопросы, связанные с применением тестовых технологий в учебном процессе.

4.1. УЧЕТ МОТИВАЦИИ ИСПЫТУЕМЫХ ПРИ ОРГАНИЗАЦИИ ТЕСТОВОГО КОНТРОЛЯ ЗНАНИЙ

При организации учебного процесса весьма важным фактором является учет мотивации к учению. Буквально одно слово преподавателя в определенных условиях может сильно повлиять на учебную мотивацию обучаемого. Согласно Л.И.Божович[1], мотив - это то, ради чего осуществляется деятельность индивида. Мотивация представляет собой совокупность побуждающих факторов, определяющих активность личности.

Если рассматривать учебный процесс как управляемую систему, то становится очевидной важность такого элемента системы как обратная связь. Именно обратная связь позволяет преподавателю получать информацию о текущем состоянии учебных достижений учащихся, что позволяет выполнять коррекцию хода учебного процесса и эффективно организовывать его. С другой стороны, обратная связь позволяет учащемуся осуществлять самоконтроль и самодиагностику своего процесса учения.

Обратная связь организуется в разнообразных формах, среди которых все большее применение находят педагогические тесты. Тесты являются объективным инструментом педагогической диагностики, позволяя организовать эффективную систему обратной связи в современных педагогических технологиях. Следует отметить, что вопросы мотивации в тестировании пока слабо изучены.

Мы предпримем попытку формального учета мотивации испытуемых при анализе результатов тестирования, в частности, при определении поправок на угадывание к тестовому баллу испытуемого.

На практике чаще всего используются тестовые задания с выбором одного или нескольких правильных ответов. Применение таких заданий осложняется попытками испытуемых угадать правильный ответ. Индивидуальный балл испытуемого в этом случае будет отличаться от истинного, что снижает диагностическую ценность тестирования. Вышесказанное обусловливает актуальность проблемы коррекции результатов тестирования.

В целях упрощения процесса тестирования можно, например, не учитывать вообще вероятность угадывания при обработке результатов тестирования. За правильный ответ испытуемый получает, например, 1 балл, за неправильный - 0 баллов. Поскольку часть баллов получена из-за угадывания, то испытуемые ставятся в неравные условия. Преимущество получают те, кто отличается сообразительностью, умеют анализировать задания по формальным и другим признакам, что помогает угадыванию правильного ответа.

Например, в тестовых заданиях по физике очень часто можно просто проверить размерности итоговых физических величин в ответах на задание. Если размерности не соответствуют заданию, то эти ответы отбрасываются. При этом уменьшается число ответов, среди которых легче угадать правильный. Иногда это число сокращается до единицы, то есть можно совершенно точно выбрать правильный ответ, даже не читая полностью задания. Существуют и другие способы анализа ответов, помогающих угадыванию.

Таким образом, игнорирование проблемы угадывания правильных ответов может серьезно снизить доверие к применению заданий с выбором одного правильного ответа, снизить учебную мотивацию.

Рассмотрим различные подходы к решению данной проблемы. В дальнейшем будем предполагать, что каждое задание теста содержит фиксированное количество ответов k, из которых только один правильный.

Хотя задания с выбором одного правильного ответа сильно критикуются за сравнительно высокую вероятность угадывания правильного ответа, эти задания, тем не менее, очень широко распространены. Поэтому анализ именно таких заданий является востребованным на практике. Хотя есть работы, где вместо заданий с выбором одного правильного ответа рекомендуется переходить, где

это оправдано, к заданиям с выбором нескольких правильных ответов. В таких заданиях вероятность угадывания резко снижается[2].

ФИКСИРОВАННАЯ ПОПРАВКА НА КОРРЕКЦИЮ ТЕСТОВОГО БАЛЛА

Допустим, что соблюдается условие равной привлекательности дистракторов в задании. Кроме того, дистракторы должны быть достаточно привлекательными по сравнению с правильным ответом. Обычно предполагается, что каждый из дистракторов должен выбираться не менее чем пятью процентами испытуемых. Тогда с увеличением количества ответов k в каждом задании вероятность угадывания падает. То есть, поправка должна быть обратно пропорциональна количеству ответов в задании.

В простейшем случае можно использовать фиксированную поправку, которая определяется следующим образом:

$$\Delta p_{const} = \frac{1}{k} \qquad (4.1.1)$$

Тогда

$$Y = \left(p - \Delta p_{const}\right)M \qquad (4.1.2)$$

где k – количество ответов в задании, L – количество заданий в тесте, Y – исправленный индивидуальный балл.

При $k = 4$ поправка равна 0,25 независимо от значения X.

Соответствующие зависимости показаны на рис.4.1.1.

Из рисунка видно, что исправленная зависимость получается параллельным сдвигом исходного графика на 25 единиц вниз. Индивидуальные баллы испытуемых уменьшаются на величину сдвига исправленного графика. Параллельность графиков означает, что все испытуемые теряют одно и то же количество баллов.

В однопараметрической модели G.Rasch (Item Response Theory) поправки на угадывание не вводятся. Это сделано в трехпараметрической модели:

$$P_{ij} = \left\{1; \theta_i, \beta_j, a_j, c_j\right\} = c_j + \left(1 - c_j\right)\frac{\exp a_j\left(\theta_i - \beta_j\right)}{1 + \exp a_j\left(\theta_i - \beta_j\right)}.$$

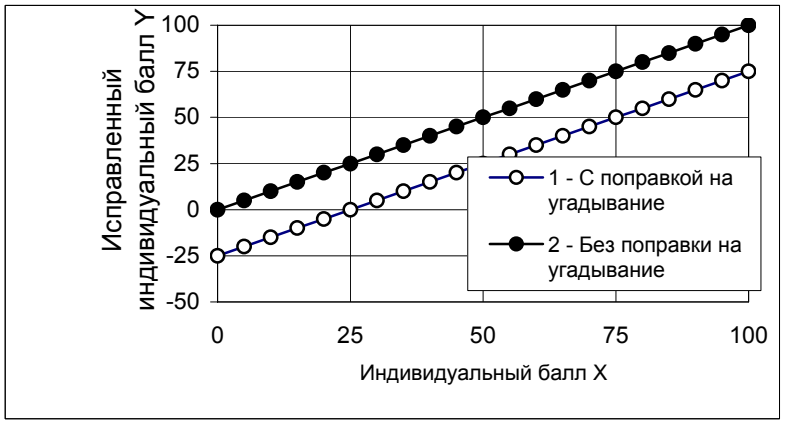


Рис.4.1.1. Исправленный индивидуальный балл с фиксированной поправкой.

Параметр c_j должен характеризовать вероятность угадывания. За разъяснением смысла остальных величин можно обратиться к работе В.Аванесова[3]. На рис.4.1.2 приведены соответствующие логистические кривые из работы B.Wright[4].

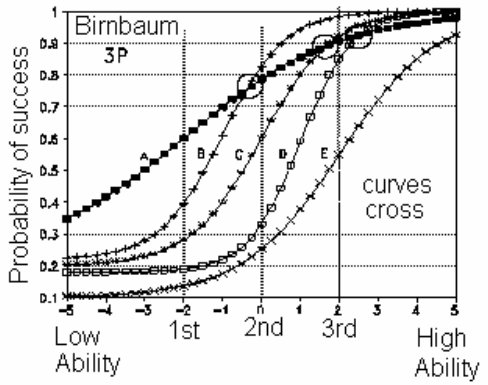


Рис. 4.1.2. Поправка на угадывание в 3-х параметрической модели IRT (B.Wright, 1992).

Из рисунка видно, что учет поправки на угадывание приводит к смещению кривых вверх по оси ординат.

Введение постоянных поправок лишь частично решает проблему и педагогически мало оправданно, так как в этом случае не различаются сильные и слабые испытуемые. В этом методе априори считается, что и сильный и слабый испытуемые в одинаковой степени пытаются угадать правильный ответ. Разумеется, это неверно. Сильный испытуемый не нуждается в угадывании, его знаний достаточно, чтобы с высокой вероятностью успешно справиться с заданием.

Таким образом, фиксированные поправки вводятся достаточно просто, но с педагогической точки зрения они могут быть не эффективны, поскольку не учитывают мотивацию сильных и слабых испытуемых. Для того чтобы учесть различие в мотивации к угадыванию у сильных и слабых испытуемых необходимо использовать поправку Δp_1, зависящую от доли правильных ответов

$$p = \frac{X}{M}$$

или от доли неправильных ответов $q = 1 - p$.

Итак, задача состоит в том, чтобы попытаться предугадать поведение испытуемого и внести коррекцию в его индивидуальный балл.

Представляется разумным считать, что чем ниже уровень знаний испытуемого, тем сильнее его стремление восполнить недостаток знаний простым угадыванием правильного ответа, без применения тех или иных методик выбора правильного ответа по формальным признакам. Если же испытуемый хорошо подготовлен, то есть обладает малым значением $q = 1 - p$, то его стремление к угадыванию будет очень слабым и поправка должна быть малой.

Иными словами, чем больше q, тем больше должна быть поправка. С другой стороны, чем больше k – количество ответов в задании, тем меньше должна быть поправка.

Обозначим через Y – исправленный индивидуальный балл испытуемого. Эту величину можно вычислить по формуле:

$$Y = X + \Delta X$$

где ΔX – поправка к индивидуальному баллу испытуемого.

Используем соотношения $X = p \cdot M$ и $\Delta X = \Delta p \cdot M$,

где Δp - поправка к доле правильных ответов испытуемого.

С учетом этих соотношений, исправленный индивидуальный балл перепишем в виде:

$$Y = (p + \Delta p) \cdot M$$

Примем, что поправка Δp_1 прямо пропорциональна q:

$$\Delta p_1 = \Delta p_{const}\, q$$

В этом случае исправленный индивидуальный балл равен

$$Y = (p - \Delta p_1) \cdot M = \left(p - \frac{q}{k}\right) \cdot M \qquad (4.1.3)$$

Соответствующая зависимость показана на рис. 4.1.3.

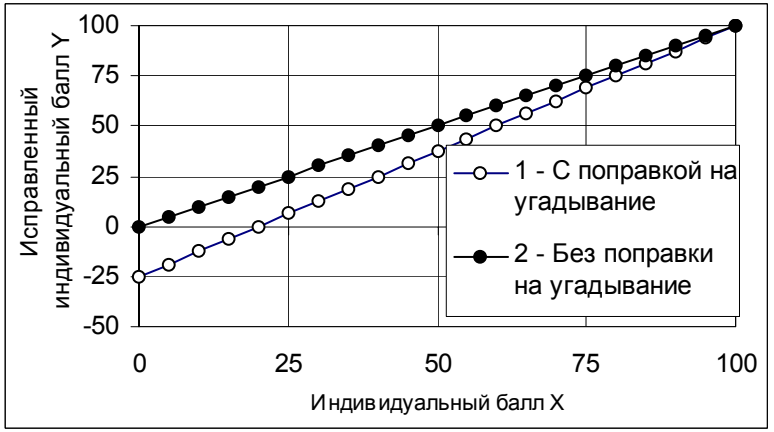


Рис. 4.1.3. Индивидуальный балл, исправленный с использованием Δp_1.

Из рис. 4.1.3 наглядно видно, что для сильных испытуемых с высокими значениями индивидуального балла поправки имеют малое значение и стремятся к нулю при X стремящемуся к M. Ситуация заметно улучшилась, по сравнению с тем, что показано на рис. 4.1.1. Отметим, что формула (4.1.3) почти не получила практического применения из-за своей слабой педагогической и психологической обоснованности. Отметим, что при $X = 25$ (четвертая часть всех заданий), в рассматриваемом приближении логично предположить,

что поправка равна индивидуальному баллу (ΔX = 25). Тогда исправленный балл должен быть равен нулю (Y = 0). Однако, из рис.4.1.3 следует, что Y = 6,25 балла.

Таким образом, требуется дать большее обоснование действию механизма поправки и желательно не ограничиваться линейными приближениями.

Предположим, что поправка Δp нелинейно зависит от доли неправильных ответов следующим образом:

$$\Delta p = \mu \cdot q^n, \qquad (4.1.4)$$

где n - натуральное число, μ - некоторый коэффициент, подлежащий определению.

Для определения μ примем во внимание два следующих обстоятельства.

Во-первых, в большинстве тестовых заданий, используемых на практике значение k находится в пределах 3 … 5.

Во-вторых, при малых значениях индивидуального балла испытуемого разумно предположить, что низкий уровень знаний способствует стремлению угадывать правильный ответ.

Что же считать низким уровнем знаний? Практически это следующие значения доли верных ответов

p_0 = 0,2 … 0,3.

Из этих двух обстоятельств следует предположение, что

$$p_0 = \frac{1}{k} = \Delta p_{const}$$

Иными словами, можно считать, что индивидуальный балл испытуемого полностью угадан и исправленный индивидуальный балл должен быть равен нулю, то есть

$$p_0 = \Delta p_{\text{const}} = \mu \cdot q^n.$$

Таким образом, получаем

$$p_0 = \frac{1}{k} = \mu q_0{}^n = \mu(1 - p_0)^n$$

или

$$\frac{1}{k} = \mu(1 - p_0)^n = \mu\left(1 - \frac{1}{k}\right)^n$$

отсюда имеем

$$\frac{1}{k} = \mu\left(1 - \frac{1}{k}\right)^n$$

Таким образом, коэффициент μ равен

$$\mu = \frac{1}{k}\left(\frac{k}{k-1}\right)^n \qquad (4.1.5)$$

Зная коэффициент μ, мы можем записать выражение для исправленного индивидуального балла испытуемого.

$$Y = \left(p - \frac{1}{k}\left(\frac{kq}{k-1}\right)^n\right) \cdot M \qquad (4.1.6)$$

Мы получили формулу, позволяющую вводить поправки на угадывание, используя различные нелинейные модели.

Ввиду важности обоснования этой формулы еще раз приведем оба положения, лежащие в основе рассуждений:

1) в практическом тестировании используются тестовые задания с k =3…5;

2) значения p_0 = 0,3 … 0,2 считаются низкими и полностью обусловленными угадыванием. Исходя из этого, мы можем приближенно считать, что $p_0 = 1/k$.

Используя выражение (4.1.6) для исправленного индивидуального балла испытуемого, рассмотрим различные модели коррекции и проведем их сравнительных анализ.

ФИКСИРОВАННАЯ ПОПРАВКА, n=0

Подставив n=0 в формулу (4.1.6), получим тестовый балл с фиксированной поправкой (см.формулу (4.1.2))

$$Y = X - \frac{N}{k}$$

Как указывалось выше, фиксированные поправки не учитывают мотивацию испытуемых и малопригодны в педагогическом тестировании.

ЛИНЕЙНАЯ МОДЕЛЬ, n = 1

В этом случае из формулы (4.1.6) имеем:

$$Y = \left(p - \frac{q}{k-1} \right) \cdot M \qquad (4.1.7)$$

Поскольку $X = pM$ и $W = qM$, то можно переписать это выражение в другом виде

$$Y = X - \frac{W}{k-1} \qquad (4.1.8)$$

где W - количество неверных ответов испытуемого.

Эта формула хорошо известна и давно используется в практике тестирования. Линейная модель широко применяется и формула (4.1.8) приведена, например, в работе В.С.Аванесова[5].

В таблице 4.1.1 приведены значения исправленного индивидуального балла в линейной модели, а на рис.4.1.4 - соответствующие зависимости.

Из таблицы 4.1.1 видно, что с ростом индивидуального балла испытуемого, его поправка на угадывание стремится к нулю. В этой модели предполагается, что при 25% правильных ответов, исправленный индивидуальный балл равен нулю (Y=0), то есть объем знаний испытуемого равен нулю и все правильные ответы получены путем угадывания.

Таблица 4.1.1. Поправка на угадывание в линейном приближении, при k = 4.

X, индивидуальный балл	ΔX, поправка	Y, исправленный индивидуальный балл
0	-33	-33
25	0	0
50	17	33
75	8	67
100	0	100

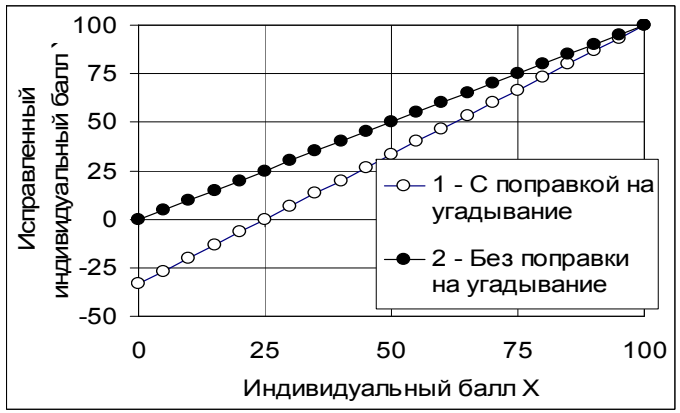


Рис.4.1.4. Коррекция индивидуального балла в линейной модели.

Если индивидуальный балл равен 100, то поправка отсутствует и исправленный индивидуальный балл совпадает с исходным, то есть тоже равен 100. В этом случае предполагается, что испытуемый не имел стимула отвечать наугад, так как располагал достаточно полным объемом знаний.

При низких индивидуальных баллах (X < 25) получаются отрицательные поправки, что не имеет смысла. Поэтому в этих случаях следует просто считать, что исправленный индивидуальный балл просто равен нулю (Y=0).

Отметим, что для X = 25 следует Y = 0, что соответствует логике наших рассуждений.

НЕЛИНЕЙНАЯ, ПАРАБОЛИЧЕСКАЯ МОДЕЛЬ, $n = 2$

Подставив $n = 2$ в формулу (4.1.6), получим для параболической модели выражение для исправленного индивидуального балла:

$$Y = \left(p - k\left(\frac{q}{k-1}\right)^2 \right) \cdot M \qquad (4.1.9)$$

Таблица 4.1.2. Поправка на угадывание в параболическом приближении, при $k = 4$.

X, индивидуальный балл	ΔX, поправка	Y, исправленный индивидуальный балл
0	-44	-44
25	0	0
50	11	39
75	3	72
100	0	100

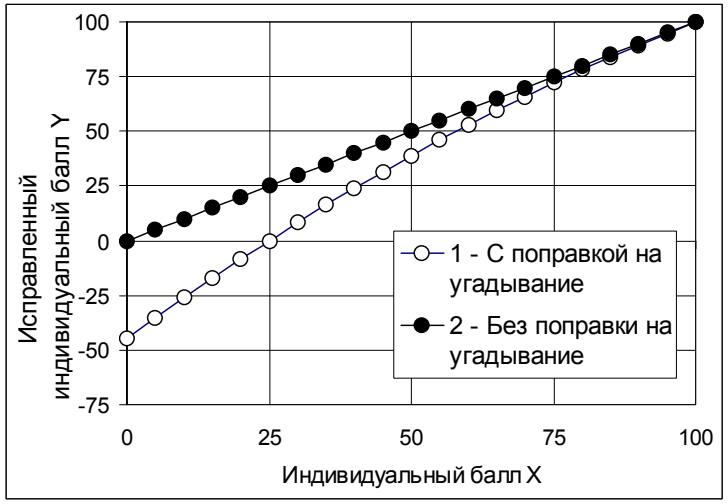


Рис. 4.1.5. Коррекция индивидуального балла в параболической модели.

В этой модели, как и предыдущем случае для $X = 25$ следует $Y = 0$. Из таблицы 4.1.2 следует, что наблюдается более жесткая реакция на угадывание для слабых испытуемых. Для сильных же испытуемых коррекция меньше, чем в линейной модели (4.1.7). Соответствующие графические зависимости приведены на рис. 4.1.5.

НЕЛИНЕЙНАЯ, КУБИЧЕСКАЯ МОДЕЛЬ, $n = 3$

Подставив $n = 3$ в формулу (4.1.6), получим для параболической модели выражение для исправленного индивидуального балла:

$$Y = \left(p - k^2\left(\frac{q}{k-1}\right)^3 \right) \cdot M \qquad (4.1.10)$$

Результаты расчетов приведены в таблице 4.1.3 и на рис. 4.1.6.

Таблица 4.1.3. Поправка на угадывание в кубическом приближении, при $k = 4$.

X, индивидуальный балл	ΔX, поправка	Y, исправленный индивидуальный балл
0	-59	-59
25	0	0
50	7	43
75	1	74
100	0	100

В кубической модели наблюдается усиление тенденции, проявившейся в параболической модели. К слабым испытуемым предъявляются еще более жесткие требования, а к сильным - значительно более мягкие.

Рассмотрение поправок для $n > 3$ нами не проводится, так как тенденция ясна и практического значения сильно нелинейные модели почти не имеют.

В психологическом плане, возрастание показателя n можно трактовать усиление недоверия к слабым испытуемым и наоборот, усиление доверия к сильным испытуемым.

Сфера применения тех или иных моделей определяется педагогическими условиями, в которых проводится тестирование. Возможны ситуации, когда можно обойтись вообще без коррекции тестового балла. Чаще всего коррекция все-таки нужна, что

обусловлено педагогической целесообразностью, стремлением повысить валидность тестовых результатов.

Нелинейные модели можно рекомендовать к применению в группах испытуемых с четко выраженным разделением на сильных и слабых.

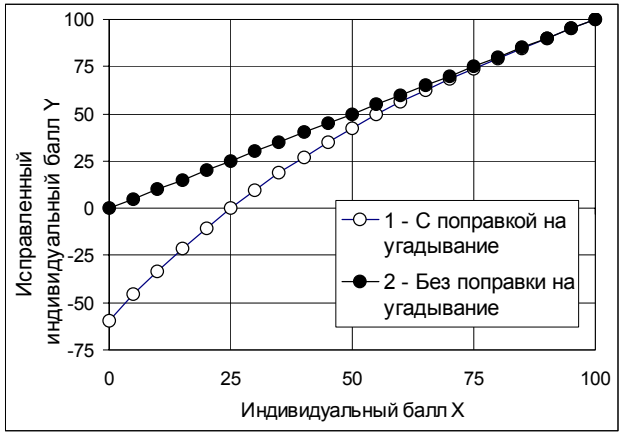


Рис.4.1.6. Коррекция индивидуального балла в кубической модели.

Таким образом, для правильной организации процесса тестирования необходимо вводить поправки на угадывание. При этом необходимо учитывать различие в мотивации к угадыванию у различных групп испытуемых.

4.2. ПРЕОБРАЗОВАНИЕ ТЕСТОВЫХ БАЛЛОВ В ОЦЕНКИ

Широкое внедрение тестирования в образовательный процесс высших и средних учебных заведений и трудности восприятия тестовых результатов в образовательной практике часто вынуждает исследователей трансформировать тестовые баллы в привычные оценки. И хотя такого рода перевод данных снижает дисперсию тестовых результатов и ухудшает дифференцирующую способность теста, реальная практика часто заставляет представлять тестовые баллы в обычной системе школьной и вузовской пятибальной шкалы. Подобная шкала отметок подвергается заслуженной критике, но, тем не менее, она обладает и рядом достоинств, что и позволяет ей прочно сохраняет свои позиции.

В.Аванесов отмечает, что оценки нередко путают с отметками. Отметки он считает численными аналогами оценочных суждений. Основная цель измерения в педагогике — это получение численных эквивалентов степени выраженности интересующего признака на интервальной шкале[2].

Несмотря на отмеченное принципиальное отличие оценок и отметок, в практике их почти всегда отождествляют. Видимо это обусловлено устоявшейся на практике терминологией. В частности, в рекомендациях Федерального центра тестирования[6] под термином «оценка» понимается именно «отметка».

Главным достоинством пяти- и четырехбалльной шкалы является ее простота, обусловленная ограниченной разрешающей способностью человека как измерительного инструмента. Педагогу достаточно легко отследить градации объема знаний в пределах 4-7 уровней. Если же ввести, например 20-ти балльную шкалу отметок, то отличить 19 баллов от 20 педагог просто не сможет.

При математической обработке результатов тестирования, преобразовании их в отметки по той или иной процедуре, вычислении средней отметки, следует иметь в виду, что отметки определены на порядковой шкале[7]. В частности, нельзя в качестве средней отметки использовать среднее арифметическое. Орловым А.И[8] показано, что, согласно законам нечисловой статистики, для определения среднего значения величины по порядковой шкале необходимо использовать не просто среднее арифметическое, а среднее арифметическое центральных членов вариационного ряда, то есть медиану. Среднее же арифметическое используется для интервальных шкал.

Важность корректного определения оценки обусловлена тем, что оценка является мощным педагогическим инструментом,

посредством которого педагог весьма эффективно может влиять на учебный процесс.

Процедура перевода тестовых баллов в отметки включает в себя либо таблицу соответствия некоторого диапазона тестовых баллов отметкам, либо некоторое математическое выражение, позволяющее определить отметку.

Остановимся сначала на таблицах. Разные авторы предлагают различные таблицы. Например, в работе В.Дубас[9] предлагается следующая таблица 4.2.1.

Таблица 4.2.1

«2»	«3»	«4»	«5»
0,4> V >=0,1	0,7> V >=0,4	0,9> V >=0,7	1= V >=0,9

В этой таблице V- относительный объем знаний.

Согласно этой таблице В.Дубас предлагает номограмму (рис. 4.2.1), позволяющая быстро осуществить процедуру перевода тестовых баллов в отметки.

Введем обозначения:

N – максимальное количество баллов;

X – индивидуальный балл испытуемого.

Рассмотрим пример использования номограммы. Допустим, что индивидуальный балл испытуемого составляет X=16. На номограмме отмечаем горизонтальную линию на высоте 16 единиц. Назовем эту линию линией индивидуального уровня. Эта линия индивидуального уровня пересекает графики почти всех отметок, а именно – «2», «3», «4» и «5».

Опуская перпендикуляры из точек пересечения линии индивидуального уровня с графиками отметок, мы найдем количество заданий *M* в тесте, для точного соответствия данной отметке. В таблице 4.2.2 приведены значения *M* в случае *X*=16 для различных отметок.

Таблица 4.2.2.

Оценка	«1»	«2»	«3»	«4»	«5»
N	-	40	23	18	16

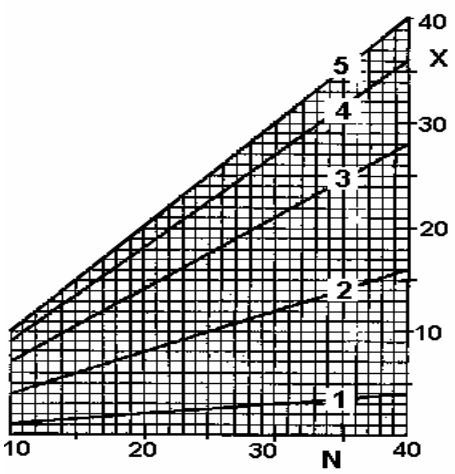


Рис. 4.2.1. Номограмма для определения отметки.

На практике нас обычно интересует обратная задача – найти отметку при заданном значении *N*. В этом случае ответ получается неоднозначный. Вернемся к нашему примеру *X*=16. Допустим, что испытуемый прошел тест из 30 заданий. На номограмме проводим вертикальную линию, соответствующую значению *N*=30 и отмечаем точку пересечения с линией индивидуального уровня. Эта точка удалена от графика «2» на 5 единиц, а от графика «3» на 4 единицы. Таким образом, получаем, что отметка находится между «2» и «3», но ближе к «3».

Несмотря на широкое применение вычислительной техники в учебном процессе, подобные «подручные» методы расчета могут оказаться полезными.

В некоторых случаях предпочтительнее использование тех или иных формул для определения отметок. В этих случаях, в уравнениях используются величины на интервальной шкале, а полученный результат (отметка) рассматривается на порядковой шкале). В частности, А.Молибог[10] предлагает приближенное соотношение следующего вида:

$$Y = 3{,}3 \lg(1/(1-V))$$

где Y — оценка (отметка) в баллах ($Y = 2, 3, 4, 5$);
V—объем знаний материала в долях от 1.

Кривая, соответствующая указанной выше зависимости, представлена на рис. 4.2.2.

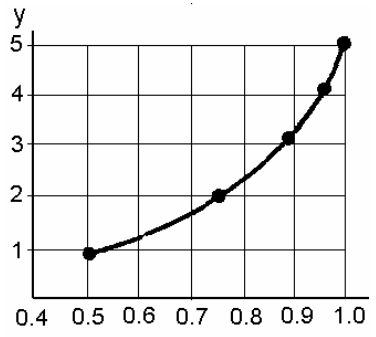


Рис. 4.2.2.

Здесь обращает на себя внимание сильно нелинейная зависимость величины отметки от относительного объема знаний, а также высокий уровень требований к знаниям испытуемых. В частности, оценке «3» соответствует значение V=0.75, что довольно много.

Например, Федеральный центр тестирования предлагает следующую таблицу (таблица 4.2.3) по переводу тестовых баллов в отметки. Из этой таблицы видно, что V=0.75 – это заведомо отличная оценка. Разумеется, подобные таблицы преобразования тестовых баллов в отметки изначально субъективны и отражают множество скрытых факторов.

Тем не менее, такие таблицы представляется интересным классифицировать по типу зависимости «отметка – тестовый балл».

Проследим общие закономерности разработки процедур преобразования тестовых баллов в оценки. Известно, что табулированные функции можно с той или иной степенью точности описать достаточно простыми уравнениями. В этой связи удобно анализировать не таблицы перевода, а поведение функций, описывающих эти таблицы.

В дальнейшем будем предполагать, что оценка **y** связана с индивидуальным баллом ***X*** испытуемого нелинейной зависимостью вида:

$$y = aX^n + b$$

где a, b, n – коэффициенты, подлежащие определению.

Таблица 4.2.3. Рекомендации по переводу тестового балла централизованного тестирования (вузовского) в пятибалльную шкалу оценок в 2005 году[11].

Предмет	«2»	«3»	«4»	«5»
1. Русский язык	0-36	37-50	51-65	66-100
2. Математика	0-34	35-48	49-67	68-100
3. Физика	0-37	38-47	48-65	66-100
4. Химия	0-33	34-48	49-69	70-100
5. Информатика	0-36	37-48	49-67	68-100
6. Биология	0-36	37-49	50-65	66-100
7. История России	0-38	39-49	50-62	63-100
8. География	0-38	39-48	49-61	62-100
9. Английский язык	0-36	37-49	50-66	67-100
10. Немецкий язык	0-36	37-48	49-65	66-100
11. Французский язык	0-35	36-50	51-65	66-100
12. Обществознание	0-36	37-50	51-63	64-100

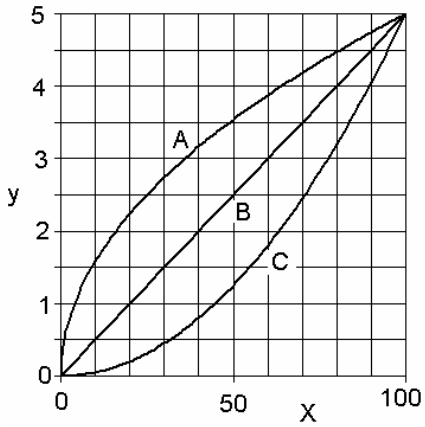


Рис. 4.2.3.

Из этих коэффициентов нас будет интересовать коэффициент n, определяющий тип зависимости (рис. 4.2.3).

При n=1 мы получаем линейную зависимость, обозначенную на рис. 4.2.3 символом В.

Кривая А соответствует случаю n<1 и характеризует сублинейную зависимость.

Кривая С соответствует случаю n>1 и характеризует надлинейную зависимость.

В случае линейной зависимости (n = 1) наблюдается прямая пропорциональная зависимость между оценками и индивидуальным баллом. Это самая простая зависимость, но на практике она, как правило, не используется.

В частности, из рис. 4.2.2 видно, что А.Молибог использует надлинейную зависимость.

В.Дубас считает, что диапазон "четверки" должен быть несколько уже диапазона "тройки". Из таблицы 4.2.1 также следует надлинейная зависимость с n = 1.4.

$$y = 0{,}006\,X^{1{,}4} + 1{,}8$$

Соответствующая надлинейная зависимость показана на рис. 4.2.4 кружочками.

Примеры сублинейных зависимостей можно взять из таблицы 4.2.3, содержащей рекомендации Федерального центра тестирования. На рис. 4.2.4 выборочно показаны зависимости для географии (треугольники) и химии (квадратики).

Шкала по географии представляет собой яркий пример сублинейной зависимости. Для химии это свойство выражено слабее. Для остальных предметов результаты занимают промежуточное положение между географией и химией.

Значение n < 1 (сублинейные зависимости) означает, что исследователь в первую очередь интересуется повышенной дифференцирующей способностью используемой шкалы в области низких отметок. Надлинейные же зависимости с n > 1 используются, когда стремятся повысить дифференцирующую способность шкалы преобразования в области высоких оценок.

Сублинейные зависимости видимо следует использовать для тестов, содержащих задания повышенной трудности. Это связано с тем, что в случае трудных заданий основная доля испытуемых будет получать относительно низкие индивидуальные баллы. Тогда, область повышенной дифференцирующей способности выгодно переместить к началу шкалы, т.е. в область низких отметок. Для тестов же с

относительно легкими заданиями желательно использование надлинейных зависимостей.

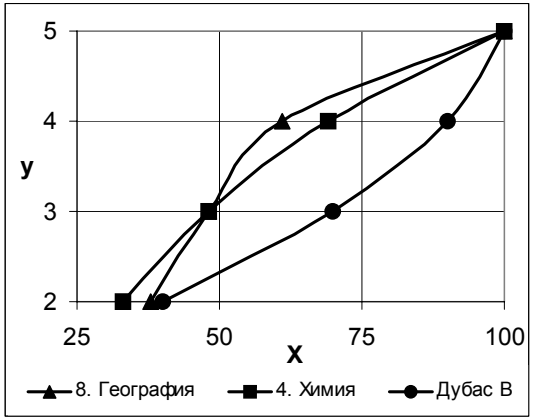


Рис. 4.2.4.

Таким образом, при разработке процедуры пребразования тестовых баллов в традиционные отметки (оценки) следует применять сублинейные или надлинейные зависимости, обращая внимание на характеристики как теста, так и испытуемых.

4.3. ЗАЩИЩЕННОСТЬ БАНКА ТЕСТОВЫХ ЗАДАНИЙ

Важной проблемой педагогического тестирования, является создание и использование банка тестовых заданий (БТЗ). Идея применения БТЗ заключается в генерации теста непосредственно в момент начала тестирования. При этом из БТЗ по заданной процедуре выбирается определенное количество тестовых заданий для конкретного теста. Из этого следует, что БТЗ должен содержать избыточное количество тестовых заданий.

Применение БТЗ повышает дидактическую эффективность тестера в связи с тем, что плохая «вскрываемость» банка стимулирует повторное изучение учебного материала и способствует повышению уровня обученности.

В этой связи, важной проблемой является обеспечение защищенности БТЗ.

В литературе приводятся данные о характеристиках защищенности банка тестовых заданий при различных уровнях «агрессивной» активности субъектов тестирования.

В работе[12] вводится коэффициент агрессивной активности субъектов тестирования k, изменяющийся в диапазоне от 0 до 1. В частности, при $k = 0$ считается, что среда пассивная и задания из БТЗ не копируется и такой банк считается нераспознанным. При $k = 1$ среда предельно агрессивна и процент узнавания тестовых заданий в БТЗ максимален. При $k = 0.5$ испытуемыми «копируется» и распространяется половина из предъявленных тестовых заданий. Для реальной практики тестирования принимается $k = 0.3$ и число тестовых заданий 45-55. Тогда БТЗ, содержащий 200 заданий, вскрывается за 60 сеансов тестирования, БТЗ из 300 заданий - за 90 сеансов, а БТЗ из 1000 тестовых заданий рассекречивается за 300 сеансов тестирования.

На наш взгляд это завышенная оценка вскрываемости БТЗ. Здесь не учтена мотивация испытуемых к рассекречиванию БТЗ. Эффективное значение активности k должно быть, на наш взгляд, ниже. Отвлекаясь от «агрессивной активности» испытуемых, рассмотрим, какие препятствия необходимо преодолеть, чтобы выполнить копирование тестового задания.

Во-первых, это непросто чисто технически. Простое переписывание достаточно легко обнаруживается персоналом и пресекается. Вполне возможно, что вслед за этим последует удаление испытуемого с тестирования. Требуется достаточно высокий уровень мотивации, чтобы идти на такой риск. Если испытуемый считает, что

сумеет пройти тест, то у него нет мотива к копированию заданий, так как результаты будут нужны не ему, а другим – тем, кто не пройдет тест.

Копирование «компьютерным» способом (клавиша Print Screen и тому подобное) может быть заблокировано программой-тестером, в качестве примера можно привести тестер Федерального центра тестирования. Для обхода блокировки потребуется немалый объем работы весьма квалифицированного программиста. Кроме того, для инсталляции программы «разблокиратора» возможно потребуется несанкционированный доступ к компьютеру с правами администратора, что тоже весьма проблематично и наказуемо. Во всяком случае, и этот путь рискован и также требует активности именно от слабых учащихся, которые уверены, что не пройдут тестирование.

Во-вторых, скопированные задания надо аккумулировать и распространять среди испытуемых. Поскольку сеансы тестирования разнесены во времени на промежуток от 1 часа и более (более короткие интервалы не стоит рассматривать, так как испытуемые не успевают изучить скопированные задания), то требуется многодневная организационная работа по аккумулированию и распространению тестовых заданий. Следует учесть, что, как правило, тестер разрешается запускать только в строго определенные моменты – согласно расписанию тестирования. Например, тестируется поток из 3 академических групп по 25 человек. Обычно, согласно расписанию, интервалы между тестированиями составляет в среднем 5 дней. После трех сеансов остаются только те, кто не выдержал тест. Для задолжников также назначаются пересдачи с не меньшими временными промежутками. В результате, для вскрытия БТЗ даже из 200 заданий может потребоваться 300 - 400 дней. Трудно представить себе испытуемого, способного в течение года собирать тестовые задания, чтобы подготовиться к тесту. Гораздо меньше усилий потребуется для того, чтобы просто выучить учебный материал и успешно пройти тест.

По вышеуказанным причинам, мотивация испытуемых к вскрытию БТЗ может оказаться довольно незначительной. Иными словами коэффициент k должен быть значительно меньше 1.

Наш практический опыт тестирования показывает, что БТЗ из 200 заданий практически не вскрываем. Не потому, что это очень трудно, а потому, что учащиеся не хотят этого делать. В качестве примера можно привести случай, когда испытуемым был передан тестер, которым можно было свободно пользоваться, в частности

многократно самостоятельно запускать его, то есть сеансы тестирования могли следовать непрерывно один за другим. Как только испытуемому удавалось набрать нужный процент правильных ответов - немедленно ставился зачет (допуск к лабораторной работе). Даже в таких сверх агрессивных условиях, испытуемые приходили к выводу, что проще и быстрее будет изучить учебный материал, чем тратить силы и время на расшифровку БТЗ.

Если же БТЗ содержит 1000 заданий и более, то их можно даже выложить в открытый доступ, то есть раскрыть БТЗ. Изучение такого количества заданий вполне эквивалентно изучению учебного материала на достаточно хорошем уровне. Таким образом, БТЗ, содержащий несколько тысяч заданий и более, можно вообще не защищать.

Таким образом, следует еще раз подтвердить, что компьютерное тестирование как элемент обратной связи, рассматриваемой в рамках педагогической кибернетики, может содействовать повышению эффективности процесса учения только при учете личностно-деятельностных педагогических закономерностей.

4.4. РАЗВИВАЮЩАЯ ФУНКЦИЯ ТЕСТА

Как известно, в образовательной системе России преобладает дидактическая доминанта, во главу угла которой ставится передача знаний, умений и навыков. В то же время, образовательные учреждения общего и профессионального образования должны способствовать общей социализации обучаемых, развитию их личностных качеств и ключевых компетенций, необходимых в современном обществе.

Большинство педагогических технологий, появившихся в последнее время, используют компетентностный подход, развивающий эвристические и творческие способности личности[13]. Важнейшим свойством педагогической технологии является ее диагностичность. Благодаря этому свойству педагогическая технология позволяет своевременно получать как актуальную информацию о состоянии учебного процесса в целом, так и результаты контроля по отдельным этапам обучения. Это позволяет организовать постоянный мониторинг образовательного процесса с целью прогнозирования индивидуальных траекторий обучаемых в ближайшем и отдаленном будущем.

Наиболее объективным инструментом педагогического контроля являются тесты. Педагогические тестирование служит не только целям контроля знаний. Как отмечает В.Аванесов[2], одной из функций педагогического тестирования является обучающая функция, которая наиболее ярко проявляет себя в программированном обучении. В более поздней работе[14] В.Аванесов указывает, что недостаточная информированность о реальном уровне знаний учеников и естественные различия в их способностях усвоить предлагаемые знания стали главной причиной появления адаптивных систем, основанных на принципах индивидуального обучения.

Рассмотрим последовательность тестовых заданий, позволяющих реализовать развитие творческих способностей личности за счет реализации обучающей функции педагогического контроля. В качестве примера, нами взята тема «Давление твердых тел, жидкостей и газов» дисциплины «Физика» для средних общеобразовательных учреждений[15].

Как правило, в тестах на эту тему используется следующее задание (рис.4.4.1).

1. НАИБОЛЬШЕЕ ДАВЛЕНИЕ НА ОПОРНУЮ ПОВЕРХНОСТЬ ОКАЗЫВАЕТ ТЕЛО

1) А
2) + Б
3) В

Графический способ оформления задания удачен с точки зрения перекодировки информации, что содействует более прочному усвоению материала. Вариативной величиной здесь выступает площадь основания. Благодаря тому, что во всех трех альтернативах используется одно и то же тело, испытуемый без труда приходит к

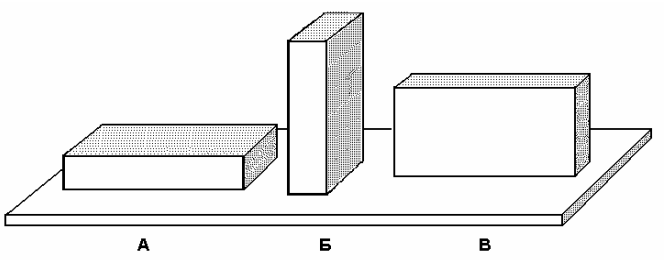


Рис. 4.4.1.

выводу, что сила веса тела – постоянна. В этом задании в основном проверяются знания испытуемого на репродуктивном уровне.

Представляется рациональным вместо одного такого задания предложить последовательность заданий возрастающей трудности. На рис.4.4.2 представлена иллюстрация к самому легкому заданию.

2. НАИБОЛЬШЕЕ ДАВЛЕНИЕ НА ОПОРНУЮ ПОВЕРХНОСТЬ ОКАЗЫВАЕТ ЦИЛИНДР

1) +А
2) Б
3) В

В этом задании изменяется только одна характеристика тела - высота. Плотность вещества и площадь основания во всех трех случаях одни и те же. От обучаемого требуется выполнить простое ментальное действие - сопоставить массу тела и его высоту, чтобы прийти к правильному ответу. Это задание репродуктивного типа и оно проверяет знание трех соотношений:

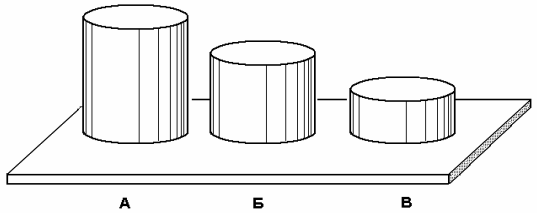


Рис. 4.4.2.

1) зависимость давления от силы давления и площади основания;
2) зависимость массы тела от его объема и плотности;
3) зависимость объема тела, в частном случае, от его высоты и площади основания.

Далее следует усложненный вариант этого задания (рис. 4.4.3).

3. ЦИЛИНДРЫ ОКАЗЫВАЮТ ДАВЛЕНИЕ НА ОПОРНУЮ ПОВЕРХНОСТЬ

1) А - наибольшее, В - наименьшее
2) А - наименьшее, В - наибольшее
3) +А, Б и В - одинаковое

Высота цилиндров неизменна, зато меняется площадь основания. Сложность задания заключается в том, что учащиеся функциональные зависимости давления от характеристик тела обычно группируют в два отдельных высказывания:

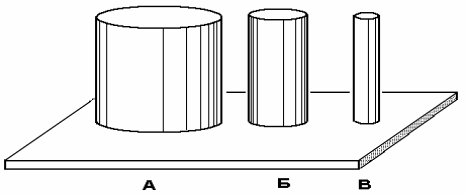


Рис. 4.4.3.

1) давление прямо пропорционально силе давления, которая в данном случае пропорциональна массе тела;

2) давление обратно пропорционально площади основания.

Если испытуемый нетвердо усвоил жесткую взаимосвязь этих высказываний и запомнил только второе, то он уверенно выбирает вариант «А», так как в этом случае площадь основания наименьшая. Только учет обоих высказываний приводит обучаемого к парадоксальному для него выводу - во всех трех случаях давление, оказываемое телами на опорную поверхность -одинаковое.

Следующее задание получено усложнением предыдущего (рис. 4.4.4). Здесь также высоты цилиндров равны. Отличительной чертой является то, что изменяются площади оснований и используются полые цилиндры.

4. ЦИЛИНДРЫ ОКАЗЫВАЮТ ДАВЛЕНИЕ НА ОПОРНУЮ ПОВЕРХНОСТЬ

1) А - наибольшее, Г - наименьшее, Б и В - промежуточное
2) А - наименьшее, Г - наибольшее, Б и В - промежуточное
3) +А, Б, В и Г - одинаковое

Обладая опытом правильного решения задания №3, испытуемый должен, проанализировать влияние толщины стенок полых цилиндров на величину давления и прийти к правильному выводу, что и в этом случае давление во всех четырех случаях одинаковое. Такой анализ непрост и требует достаточно длинных математических выкладок, которые позволяют испытуемому уяснить понятие давления. Прямого ответа на это задание в учебнике нет, то есть от испытуемого требуется провести небольшое самостоятельное исследование, активировать свой творческий потенциал.

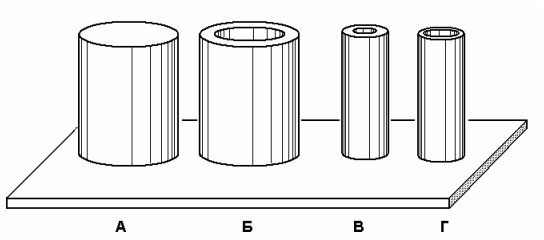


Рис. 4.4.4.

Задание №5 (рис. 4.4.5).

5. ЦИЛИНДРЫ ОКАЗЫВАЮТ ДАВЛЕНИЕ НА ОПОРНУЮ ПОВЕРХНОСТЬ

1) А - наибольшее, Б - наименьшее, В - промежуточное
2) А - наибольшее, Б - промежуточное, В - наименьшее
3) +А - промежуточное, Б - наименьшее, В - наибольшее

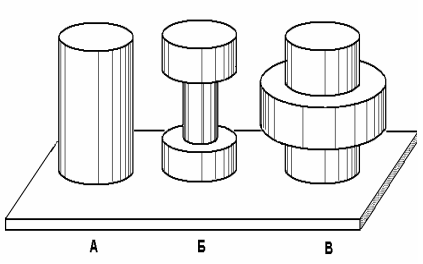


Рис. 4.4.5.

Это задание визуально кажется очень сложным, чему способствует замысловатая форма тел. В этом случае, используя эвристические возможности своего мышления, испытуемый приходит к выводу, что правилен вариант 3.

Дальнейшее усложнение формы тел приводит к заданию №6 (рис.4.4.6).

6. ТЕЛА ОКАЗЫВАЮТ ДАВЛЕНИЕ НА ОПОРНУЮ ПОВЕРХНОСТЬ

1) +А - наименьшее, Г - наибольшее, Б и В - промежуточное
2) А - наибольшее, Г - наименьшее, Б и В - промежуточное
3) А и Б - промежуточное, В - наименьшее, Г - наибольшее

Поскольку поперечное сечение и высоты среза тел одинаковые, то испытуемый приходит к выводу, что наибольшей массой обладает тело «Г», а наименьшей - тело «А».

Задание 7 (рис. 4.4.6).

7. ЖИДКОСТЬ В НЕВЕСОМОМ СОСУДЕ ОКАЗЫВАЕТ ДАВЛЕНИЕ НА ЕГО ДНО

1) А - наименьшее, Г - наибольшее, Б и В - промежуточное
2) А - наибольшее, Г - наименьшее, Б и В - промежуточное
3) А и Б - промежуточное, В - наименьшее, Г - наибольшее
4) +А , Б, В и Г - одинаковое

Испытуемому показывают тот же самый рисунок, что и в

предыдущем задании, но изменено агрегатное состояние вещества. Как это повлияет на величину давления? Испытуемый должен сопоставить свойства жидкостей и твердых тел, чтобы прийти к правильном уответу.

Ввиду того, что модуль сдвига у жидкостей равен нулю, правильный ответ разительно отличается от ответа на предыдущее задание. В данном случае испытуемому необходимо выполнить сопоставительный анализ свойств твердых тел и жидкостей, глубже осознать различие их механических свойств.

Задание 8 (рис. 4.4.6).

8. ГАЗ ОДНОЙ И ТОЙ ЖЕ ПЛОТНОСТИ ОКАЗЫВАЕТ ДАВЛЕНИЕ НА ДНО СОСУДА

1) А - наименьшее, Г - наибольшее, Б и В - промежуточное
2) А - наибольшее, Г - наименьшее, Б и В - промежуточное
3) А и Б - промежуточное, В - наименьшее, Г - наибольшее
4) +А , Б, В и Г - одинаковое

Испытуемому снова предъявляется тот же рисунок, что в задании №6. Это задание позволяет испытуемому глубже осознать тот факт, что газ удерживается в некотором объеме пространства только

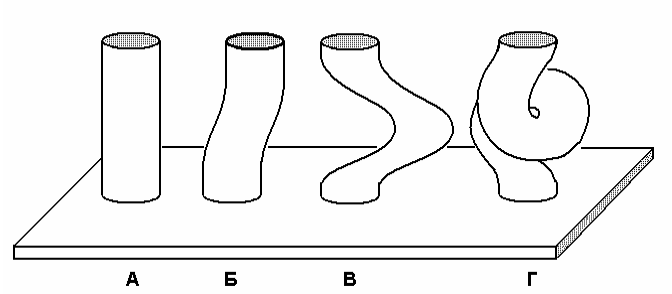


Рис. 4.4.6

благодаря закрытому сосуду, то есть свободный газ не может сохранять свой объем.

Задание 9 (рис. 4.4.7).

9. ТЕЛО ОКАЗЫВАЕТ ДАВЛЕНИЕ НА ОПОРНУЮ ПОВЕРХНОСТЬ

1) +А - наименьшее, Г - наибольшее, Б и В - промежуточное
2) А - наибольшее, Г - наименьшее, Б и В - промежуточное
3) А и Б - промежуточное, В - наименьшее, Г - наибольшее
4) А , Б, В и Г - одинаковое

Данное задание не опирается на школьные знания.

Поскольку образующая линия цилиндра и опорная точка шара в идеале имеют площадь поверхности равную нулю, то при определении давления возникают неопределенности. Подобные сингулярности в школьном курсе физики не рассматриваются, однако испытуемому с достаточно развитым творческим потенциалом под силу провести «научное исследование» с целью постижения субъективно новой для него истины. Для такого учащегося задание №9 согласно Л.Выготскому[16], ориентировано на зону ближайшего развития испытуемого. Испытуемому необходимо убедиться, что для реальных тел сингулярности исчезают. В точках соприкосновения тела с опорной поверхностью происходит упругая деформация, как тела, так и опорной поверхности (рис.4.4.7д). К подобному выводу можно прийти, осуществив домашний эксперимент с использованием резинового мяча и надувного матраца в качестве опорной поверхности.

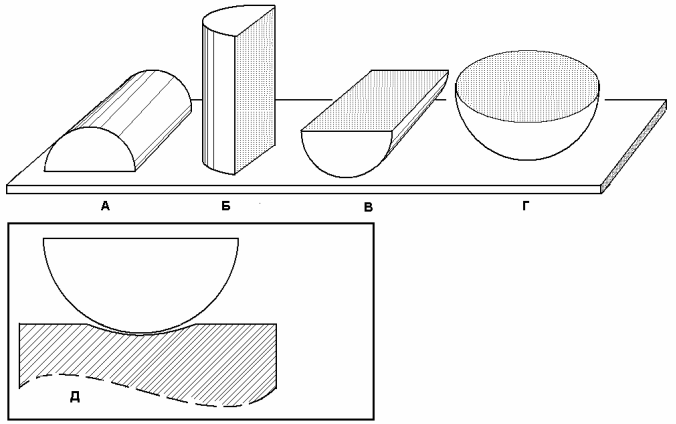


Рис. 4.4.7.

Размышления по поводу явлений, происходящих в точке контакта сферы и плоскости приводят к постановке вопроса о внутренней структуре вещества, о его атомарном строении. Подчеркнем, что подобные мыслительные действия порождаются вопросами, возникающими при рефлексии, вызванной заданием №9.

Проведем классификацию тестовых заданий, используя таксономию Блума[17]. Задания №1 и 2 требуют применения формулы

для давления, когда все исходные данные очевидны. Эти задания можно отнести к уровню «Знание».

Задания №3-6 требуют переформулировки выражения для давления, включения в нее дополнительных параметром (толщины стенок и т.п.), что позволяет отнести их к уровню «Понимание».

Задания № 7 и 8 предполагают у испытуемого наличие умений по сопоставлению различных фактов, нахождения сходства и различия, выявление причинно-следственных связей. Эти задания можно отнести к уровню «Анализ».

Задание №9 ориентировано на то, что испытуемый, не обладая нужными знаниями, самостоятельно добывает их, используя творческий процесс научного исследования. Это задание можно отнести к уровню «Синтез». По Блуму на этом уровне предполагается создание нового целого на основе изученных элементов.

Таким образом, тщательно подобранная система тестовых заданий возрастающей сложности позволяет реализовать развивающее обучение с помощью обучающей функции педагогического контроля знаний.

4.5. ДИАЛОГОВЫЙ ИНТЕРФЕЙС ДЛЯ ТЕСТОВЫХ ЗАДАНИЙ В ЗАКРЫТОЙ ФОРМЕ

В различных системах тестового контроля знаний из четырех основных форм тестовых заданий наибольшее применение получили две - задания в открытой форме и задания с выбором одного правильного ответа из предложенного списка. В обеих формах тестовых заданий необходимо иметь эталон верного ответа, с которым осуществляется сравнение ответа испытуемого.

На лабораторно-практических занятиях по некоторым дисциплинам для сдачи теоретического минимума можно использовать тестирование. Это удобно и преподавателю и студентам. Тестирование проводится на компьютере с помощью специальной тестирующей программы. При этом необходимо создать такие условия, чтобы студент мог готовиться к занятиям во внеаудиторное время, то есть тестирующая программа должна быть доступна. Далее необходимо уменьшить до минимума возможность угадывания верного ответа. Кроме того, необходимо хорошо защитить банк тестовых заданий, а также файлы с ключами верных ответов. Ниже будет приведено описание подобной тестирующей программы.

Задания с выбором одного правильного ответа из нескольких, справедливо критикуются за слабую защищенность и довольно высокую вероятность угадывания правильного ответа. С этим можно бороться различными методами, например, вводя поправки на угадывания[18].

Задания в открытой форме защищены гораздо лучше, вероятность угадывания практически сведена к нулю. Однако в этом случае возникает другая проблема - проблема корректного распознавания правильного ответа.

В литературе по тестированию известен такой пример[5]:

ПЕРВЫМ ПРЕЗИДЕНТОМ США БЫЛ _____.

Эталоном ответа являлось слово «Вашингтон». Но такие ответы как «Генерал», «Мужчина», «Уроженец штата Вирджиния» в принципе, также являются правильными.

Ниже мы становимся на рассмотрении этого вопроса более подробно.

Задания с выбором правильного ответа предпочтительнее, поскольку здесь отсутствует проблема распознавания правильного

ответа. Для ослабления эффекта угадывания можно использовать большое количество дистракторов. Обычно используются задания с двумя, тремя, четырьмя ответами. Реже - с пятью ответами. Задания с большим количеством ответов практически не используются по следующим обстоятельствам.

Во-первых, непросто создать тестовое задание с большим количеством качественных дистракторов. Это требует больших затрат времени и высокой квалификации разработчика теста. Иногда от разработчиков можно слышать утверждение, что в данном конкретном задании невозможно придумать больше одного или двух дистракторов. Однако после обсуждения различных подходов к построению тестового задания, в частности применения принципов сочетания, противоречия, противоположности, кумуляции, градуирования, удвоенного противопоставления и т.д., полно изложенных у В.Аванесова[19], разработчик убеждается, что все-таки можно было создать тестовое задание с требуемым количеством дистракторов. Естественно, это приходит с опытом. У опытных разработчиков тестов подобные вопросы обычно не возникают.

Во-вторых, испытуемому сложно ориентироваться в обилии дистракторов. Очевидно, что на тестовое задание содержащее, например, 50 ответов (1 правильный и 49 дистракторов) практически невозможно дать правильный ответ. Дистрактор может отличаться от верного ответа всего лишь одним словом, одним символом. Сравнить между собой все пятьдесят ответов, сопоставить их, проанализировать на достоверность - сложно чисто технически. Эта чрезмерная умственная и физическая нагрузка вызовет повышенную утомляемость испытуемого и, соответственно, низкие результаты тестирования, не связанные с его уровнем знаний. А потому такие задания никто не делает.

С другой стороны, если испытуемому технически обеспечить режим быстрого поиска нужного ответа, а сами дистракторы сделать легко различимыми, то такие тестовые задания, видимо, можно использовать.

В этом случае испытуемому придется оставить все попытки угадывания правильного ответа. В такой ситуации сильный, подготовленный испытуемый, проанализировав задания, сначала самостоятельно формулирует правильный ответ. Затем целенаправленно начинает поиск правильного ответа в списке ответов. Для повышения скорости поиска список ответов должен быть упорядочен, например, в алфавитном порядке по возрастанию (убыванию). В этом случае длина списка может быть очень большой -

от несколько десятков до нескольких сотен ответов. Но и такие задания тоже никто не делает.

Хотя фактически получается, что тестовое задание с выбором правильного ответа с очень большим количеством дистракторов эквивалентно тестовому заданию в открытой форме. При этом устранена проблема распознавания правильного ответа. И в том, и в другом случае угадывание становится практически невозможным.

Недостатком предлагаемого подхода является то, что далеко не всегда можно сформулировать тестовое задание, допускающее наличие большого количества однотипных, легко различимых, четко отличающихся друг от друга, дистракторов.

Рассмотрим пример задания, когда это удается сделать. Сначала сформулируем задание в открытой форме[14]:

ПЕРВЫМ ГРЕЧЕСКИМ ФИЛОСОФОМ СЧИТАЕТСЯ _____.

Эталоном ответа является слово «Фалес»

Далее, преобразуем это задание в задание с выбором правильного ответа из предложенных, например, двадцати ответов.

ПЕРВЫМ ГРЕЧЕСКИМ ФИЛОСОФОМ СЧИТАЕТСЯ

1) Фалес
2) Диоген
3) Гераклит
…
20) Аристотель

Список ответов упорядочиваем, например, по имени.

ПЕРВЫМ ГРЕЧЕСКИМ ФИЛОСОФОМ СЧИТАЕТСЯ

1) Аристотель
2) Гераклит
3) Диоген
…
20) Фалес

Очевидно, что подобные задания технически сложно реализовать на бумажных носителях, в бланковой форме - список из нескольких десятков элементов занимает очень много места. Совершенно иные возможности у разработчика тестов при использовании компьютеров. В частности, важнейшим элементом интерфейса диалоговых окон различных программных средств является так называемое окно списка, пример которого приведен на

рис. 4.5.1 (окно выбора размера шрифта в текстовом процессоре MS Word).

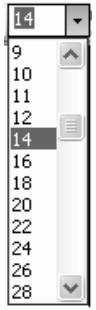


Рис.4.5.1. Окно списка.

Окно списка занимает очень небольшое место на экране, но после выполнения щелчка по «кнопке списка», появляется собственно список с полосой прокрутки. Такой интерфейсный механизм позволяет довольно быстро отыскать нужный элемент списка.

Таким образом, построение тестового задания с выбором правильного ответа с большим количеством дистракторов вполне достижимо при использовании вычислительной техники.

Для реализации описанного подхода, нами был разработана программа-тестер dbtest.exe, используемая на лабораторно-практических занятиях при изучении дисциплины «Базы данных» (ОПД.Ф.03 Федеральный компонент).

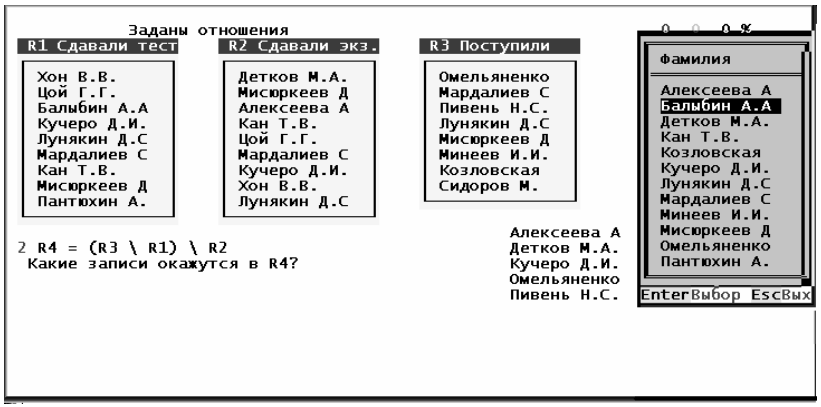


Рис. 4.5.2. Тестовое задание по реляционной алгебре.

На рис. 4.5.2. показан экран программы для одного из тестовых заданий открытой формы. Тема – «Реляционная алгебра».

На экран выводятся 4 окна. Левые три (R1, R2, R3) относятся к тексту задания. Правое крайнее окно №4 позволяет устранить вышеупомянутую проблему, связанную с применением тестовых заданий открытой формы. Речь идет о степени соответствия введенного ответа эталону, с которым сравнивается ответ.

В данном задании требуется ввести фамилии учащихся, попавших в множество R4. Что произойдет, если в процессе ввода будет допущена опечатка?

Как правило, компьютерные программы считают, что введен ошибочный ответ – нет совпадения с эталоном. С этим нельзя согласиться – ведь по существу испытуемый отвечает правильно. Для преодоления этого недостатка тестирующей программы предлагается сравнивать, например, только фрагменты слов, обычно для этого используется корень слова.

В данном тестере используется подход, описанный выше. Испытуемый выбирает нужную фамилию из списка, содержащего 67 элементов (фамилий), показанного в крайнем правом окне №4. Это напоминает задание с выбором, когда выбирается один из 67 возможных ответов. Отличие заключается в том, что, используя этот список, испытуемый формирует множество R4, путем многократного выбора элементов из списка №4. Список фамилий множества R4 появляется слева от окна №4. Количество выбираемых элементов (в данном примере - фамилий) не ограничивается – это определяется содержанием задания. По мере выбора фамилий, длина списка R4 возрастает.

Обычно тестирующие программы хранят в своей памяти эталоны ответов для всех заданий. В рассматриваемом тестере используется другой подход - выполняется моделирование действий испытуемого. Результат моделирования и есть эталон ответа для данного конкретного задания. Например, для тестового задания, показанного на рис. 4.5.2, выполняются все необходимые операции реляционной алгебры с учетом их приоритета, и определяется состав множества R4. Множества R1, R2 и R3 формируются с помощью генератора случайных чисел с использованием списка в окне №4. Как указывалось выше, список содержит 67 фамилий. Этот список находится на внешнем запоминающем устройстве в незашифрованном виде, длина списка практически не ограничена.

При каждом перезапуске программы для данного тестового задания будут формироваться разные комбинации R1, R2, R3 и вычисляться новое значение R4.

В итоге получается, что к тестирующей программе не нужен файл с ключами верных ответов, поскольку эталоны ответов вычисляются программой непосредственно в процессе компьютерного моделирования действий испытуемых. Банк тестовых заданий практически не вскрываем, так как исходные данные к заданиям

создаются непосредственно во время тестирования также благодаря компьютерному моделированию.

Таким образом, в некоторых случаях можно создавать тестовые задания с выбором правильного ответа по своим свойствам эквивалентные заданиям в открытой форме, но лишенные проблемы распознавания эталона верного ответа. Применение методов компьютерного моделирования мыслительных действий испытуемого позволяет создать тестовые задания с легко заменяемым фасетом с большим количеством элементов, что позволяет хорошо защитить программу-тестер.

Литература к главе 4

[1] Божович Л. И. Изучение мотивации поведения детей и подростков. – М.: Педагогика, 1972. − 351 с.

[2] Аванесов В.С. Основы научной организации педагогического контроля в высшей школе, 1989. –М., МИСИС. –168 с.

[3] Аванесов В.С. Применение тестовых форм в Rasch Measurement //Педагогические измерения, 2005, 4. -С.3-20.

[4] Wright B. IRT in the 1990s: Which Models Work Best? //Rasch Measurement Transactions, 1992, 6:1, 196-200.

[5] Аванесов В.С. Форма тестовых заданий. -М., 2005. -156 с.

[6] Рекомендации по переводу тестового балла централизованного тестирования (вузовского) в пятибалльную шкалу оценок в 2005 году http://www.rustest.ru/test/scale100in5.php

[7] Глас Дж., Стэнли Дж. Статистические методы в педагогике и психологии. –М.: Прогресс,1976. -495 с.

[8] Орлов А.И. Теория измерений и педагогическая диагностика. //Педагогическая информатика, 2004, №1. –С.22-31.

[9] Дубас В. Об оценивании знаний при программированном контроле //Физика в школе, 1990, №3. -С 83.

[10] Молибог А.Г. Программированное обучение (вопросы научной организации педагогического труда). –М., Высшая школа, 1967. - 243 с.

[11] Рекомендации по переводу тестового балла централизованного тестирования (вузовского) в пятибалльную шкалу оценок в 2005 году http://www.rustest.ru/test/scale100in5.php

[12] Попов Д.И., Попова Е.Д. Модель расчета рассекречивания банков тестовых заданий // Педагогические измерения, №4, 2005. –С.117-125.

[13] Селевко Г.К. Технологии развивающего образования. -М.,2005. -192 с.; Гусарова Е.Н. Современные педагогические технологии. -М., 2005. -176 с.

[14] Аванесов В.С. Современные методы обучения и контроля знаний: Уч.пос.Владивосток: Дальрыбвтуз, 1999. -123 с.

[15] Перышкин А.В. Физика. 7 кл.:Учебник для общеобразовательных учебных заведений. -М.: Дрофа, 2002. -192 с.

[16] Выготский Л.С. Педагогическая психология. -М., 1991.

[17] Bloom B.S. Human Characteristics and School Learning. New York, 1976

[18] Ким В.С. Коррекция тестовых баллов на угадывание //Педагогические измерения, 2006, №4. –С.47-55.

[19] Аванесов В.С. Композиция тестовых заданий. Учебная книга для преподавателей вузов, учителей школ, аспирантов и студентов педвузов. 2 изд., испр. и доп. М.: Адепт 1998. -217 с.

ГЛАВА 5. ПРИМЕНЕНИЕ ITEM RESPONSE THEORY В ТЕСТИРОВАНИИ УЧЕБНЫХ ДОСТИЖЕНИЙ

Классическая теория тестирования, рассмотренная в третьей главе, несмотря на хорошо разработанный математический аппарат, прозрачность и ясность получаемых выводов, имеет принципиальные недостатки. В частности, тестовые баллы испытуемых зависят от трудности заданий в тесте, а трудность задания зависит о выборки испытуемых. Большим недостатком классической теории является нелинейность тестовых баллов испытуемых.

За рубежом уже несколько десятилетий развивается современная теория тестирования - Item Response Theory (IRT), являющаяся частью более общей теории латентно-структурного анализа. Отдельно следует указать теорию Георга Раша (G. Rasch)[1] - Rasch measurement, которую иногда называют однопараметрической (теорией) IRT.

На русском язык название Item Response Theory переводится различным образом. Ю.Нейман и В.Хлебников[2] предлагают называть ее «Теория моделирования и параметризации педагогических тестов» (ТМППТ). В.Аванесов[3] - «Математико-статистическая теория оценки латентных параметров заданий теста и уровня подготовленности испытуемых». Поскольку в отечественной литературе общепринятого названия пока нет, мы будем называть ее без перевода - IRT.

5.1. ОСНОВНЫЕ ПОЛОЖЕНИЯ IRT

Перечислим преимущества IRT перед классической теорией тестов[4]:

1) IRT (особенно это относится к модели Раша) превращает измерения, выполненные в дихотомических и порядковых шкалах, в линейные измерения, в результате качественные данные анализируются с помощью количественных методов;

2) мера измерения параметров модели Раша является линейной, что позволяет использовать широкий спектр статистических процедур для анализа результатов измерений;

3) оценка трудности тестовых заданий не зависит от выборки испытуемых, на которых она была получена;

4) оценка уровня подготовленности испытуемых не зависит от используемого набора тестовых заданий;

5) неполнота данных (пропуск некоторых комбинаций испытуемый - тестовое задание) не является критичным.

Полный перечень преимуществ модели Раша приведен в работе[5].

Сформулируем несколько определений, необходимых для изложения дальнейшего материала.

ЛАТЕНТНЫЙ ПАРАМЕТР – это свойство личности, недоступное для прямого наблюдения.

Латентными параметрами являются, например, чувство патриотизма, толерантность, уровень знаний, и т. п. О величине латентного параметра можно судить по ее индикатору (индикаторной переменной). Главное достоинство индикатора – его доступность для прямого наблюдения. Измеряя значение индикатора, мы можем судить о значении латентного параметра, с которым он связан. Например, индикатором может являться тестовое задание. Значением индикатора является числовое (символьное) выражение реакции испытуемого, на это тестовое задание. По этому индикатору мы можем судить об уровне знаний, соответствующих данному тестовому заданию.

ИНДИКАТОР – это некоторое средство воздействия (вопрос, тестовое задание), связанный с определенным латентным параметром, реакция на который, доступна для непосредственного наблюдения.

Допустим, нас интересует латентный параметр «Уровень знаний по физике». Для этого мы создаем КОНСТРУКТ – систему индикаторов, позволяющих оценить латентный параметр. В нашем примере конструктом является тест по физике, а индикаторами –

тестовые задания.

ОСНОВНЫЕ ДОПУЩЕНИЯ IRT

1) существуют латентные (скрытые) параметры личности, недоступные для непосредственного наблюдения. В тестировании это уровень подготовленность испытуемого и уровень трудности задания;

2) существуют индикаторные переменные, связанные с латентными параметрами, доступные для непосредственного наблюдения. По значениям индикаторных переменных можно судить о значениях латентных параметров;

3) оцениваемый латентный параметр должен быть одномерным. Это означает, что, например тест, должен измерять знания только в одной, четко заданной, предметной области. Если условие одномерности не выполняется, то необходимо переработать тест, удалив задания, нарушающие его гомогенность.

Существуют и другие допущения, носящие специальный характер и связанные с математико-статистическим аппаратом IRT для обработки эмпирических данных[6].

ОСНОВНОЙ ЗАДАЧЕЙ IRT является переход от индикаторных переменных к латентным параметрам.

В IRT устанавливается связь между двумя множествами значений латентных параметров. Первое множество составляют значения латентного параметра, определяющего уровень подготовленности испытуемых θ_i, где i - номер испытуемого, изменяющийся в интервале от 1 до N (N - количество испытуемых). Второе множество составляют значения латентного параметра, характеризующего трудность j-го задания β_j. Индекс j меняется в пределах от 1 до M, где M - количество заданий в тесте.

Георг Раш предположил, что уровень подготовленности испытуемого θ_i и уровень трудности задания β_j размещены на одной шкале и измеряются в одних и тех же единицах - логитах. Аргументом функции успеха испытуемого является разность $\theta_i - \beta_j$.

Если эта разность положительна и велика, то соответственно высока вероятность достижения успеха i-го испытуемого в j-м задании. Если же эта разность отрицательна и велика по модулю, то вероятность достижения успеха i-го испытуемого в j-м задании будет низкой. В этом принципиальное различие подходов Гуттмана и Раша. По Гуттману в первом случае вероятность успеха в точности равна единице, а во втором - нулю. В отличие от Гуттмана Раш оперирует вероятностями, а не детерминированными константами.

5.2. МАТЕМАТИЧЕСКИЕ МОДЕЛИ IRT

В качестве математической модели, связывающей успех испытуемого с уровнем его подготовленности и трудностью задания выбирается логистическая функция. Для модели Раша она имеет вид

$$P_j(\theta) = \frac{e^{1,7(\theta-\beta_j)}}{1+e^{1,7(\theta-\beta_j)}} \qquad (5.2.1)$$

$$P_i(\beta) = \frac{e^{1,7(\theta_i-\beta)}}{1+e^{1,7(\theta_i-\beta)}} \qquad (5.2.2)$$

Масштабный множитель 1,7 используется для совместимости модели G.Rasch с моделью A.Fergusson, где вероятность правильного ответа на задание выражена интегралом нормального распределения (5.2.3), что позволяет использовать вместо логистических кривых хорошо изученную интегральную функцию нормированного нормального распределения[7]

$$P_j(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\theta-\beta_j} e^{\frac{1}{2}x^2}\,dx \qquad (5.2.3)$$

Модель Раша носит название «1 Parametric Logistic Latent Trait Model» (1PL), а модель A.Fergusson - «1 Parametric Normal Ogive Model» (1PN). Поскольку модель Раша описывает вероятность успеха испытуемого как функцию одного параметра ($\theta_i - \beta_j$), то иногда ее называют однопараметрической моделью IRT.

Взаимодействие двух множеств θ_i и β_j образует данные, обладающие свойством «совместной аддитивности» (conjoint additivity). Правильное использование модели Раша позволяет отделить оценки испытуемых от оценок трудности заданий и наоборот. Это свойство Rasch Measurement носит название separability parameter estimates[8] - «независимость оценок заданий от испытуемых и оценок испытуемых от параметров заданий».

На рис.5.2.1. показаны три характеристические кривые согласно уравнению (5.2.1) с трудностями заданий -2, 0 и +2 логита (первое самое легкое, второе - среднее, третье самое трудное). Из приведенных зависимостей видно, что чем выше уровень

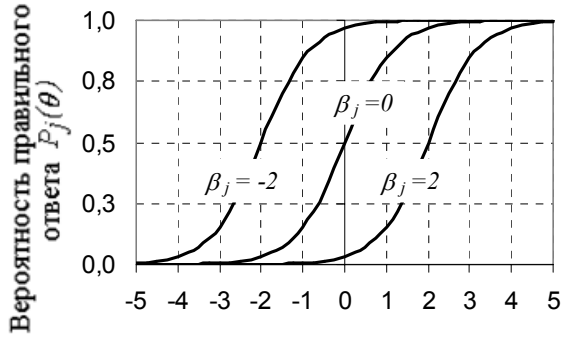


Рис.5.2.1. Характеристические кривые заданий (ICC) в модели (1PL).

подготовленности θ испытуемого, тем выше вероятность успеха в том или ином задании. Например, для испытуемого с θ =0 вероятность правильно ответить на первое задание близка к единице, на второе равна 1/2 и на третье почти равна нулю. Отметим, что в точках, где $\theta = \beta$ вероятность правильного ответа равна 0,5. То есть, если трудность задания равна уровню подготовленности (ability) испытуемого, то он с равной вероятностью может справиться или не справиться с этим заданием.

Характеристические (логистические) кривые для заданий теста в англо-язычной литература называются *Item Characteristic Curve* (ICC).

На рис.5.2.2. показаны три характеристические кривые испытуемых согласно уравнению (5.2.2) - «Person Characteristic Curve» (PCC). Показаны графики для трех испытуемых с уровнем подготовленности -2 логита (самый слабый), 0 логитов (средний) и +2 логита (сильный испытуемый).

Из приведенных зависимостей видно, что чем выше уровень подготовленности, тем выше вероятность правильного ответа на задание. Например, задание с трудностью $\beta = 0$ первый испытуемый (θ=-2) практически не сможет выполнить, второй θ = 0) имеет вероятность выполнения задания равную 0,5, третий θ=+2) легко справится с заданием, так как для него вероятность успеха почти равна единице.

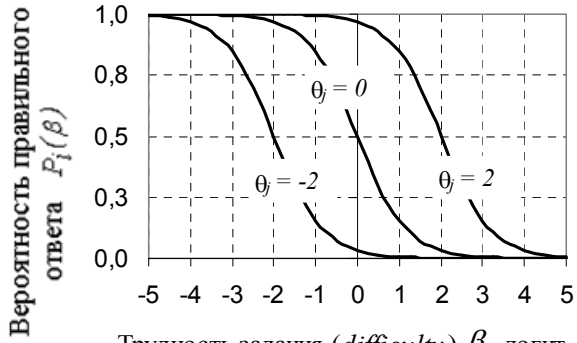


Рис.5.2.2. Характеристические кривые испытуемых (PCC) в модели 1PL.

ДВУХПАРАМЕТРИЧЕСКАЯ МОДЕЛЬ БИРНБАУМА

Как видно из приведенных зависимостей, крутизна характеристических кривых в области P_j=0,5 одинакова, то есть дифференцирующая способность является константой. Для дихотомической модели эта константа равна 0,25.

Если тест содержит задания с различной дифференцирующей способностью, то однопараметрическая модель 1PL не может описать такие эмпирические данные. Для преодоления этой трудности А.Бирнбаум (A.Birnbaum)[9] ввел еще один параметр - *a* (item discrimination parameter).

$$P_j(\theta) = \frac{e^{1,7a_j(\theta-\beta_j)}}{1+e^{1,7a_j(\theta-\beta_j)}} \qquad (5.2.4)$$

$$P_i(\beta) = \frac{e^{1,7a_j(\theta_i-\beta)}}{1+e^{1,7a_j(\theta_i-\beta)}} \qquad (5.2.5)$$

Параметр a_j определеяет наклон (крутизну) характеристической кривой *j*-го заданий. Примеры характеристических кривых показаны на рис.5.2.3. Видно, что чем больше aj тем круче идет кривая, тем выше дифференцирующая способность задания.

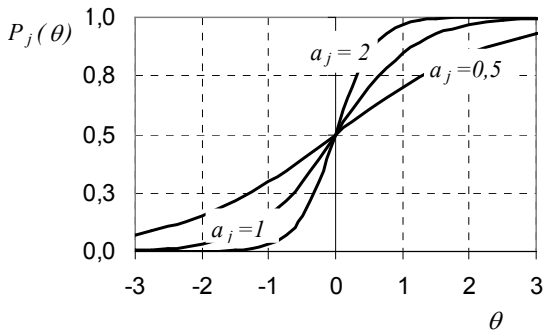


Рис.5.2.3. ICC в двухпараметрической модели 2PL

Для еще лучшего соответствия эмпирическим данным А.Бирнбаум ввел третий параметр c - параметр угадывания*.

$$P_j(\theta) = c_j + (1 - c_j)\frac{e^{1,7a_j(\theta-\beta_j)}}{1+e^{1,7a_j(\theta-\beta_j)}} \qquad (5.2.6)$$

$$P_i(\beta) = c_j + (1 - c_j)\frac{e^{1,7a_j(\theta_i-\beta)}}{1+e^{1,7a_j(\theta_i-\beta)}} \qquad (5.2.7)$$

Из уравнений (5.2.5) и (5.2.6) видно, что при c_j=0 и a_j=1 эти уравнения переходят в однопараметрическую модель. По этой причине иногда говорят, что модель Раша является частным случаем двух и трехпараметрической моделей Бирнбаума. Формально это так, но по существу это неверно. К обсуждению этой проблемы мы вернемся далее.

На рис.5.2.4. приведены примеры характеристических кривых для трех заданий с трудностью $\beta = 1$, дискриминационным параметром $a_j = 1$ и различными параметрами угадывания $c_j = 0$, $c_j = 0{,}25$, $c_j = 0{,}5$.

Из приведенных графиков видно, что наличие параметра угадывания приводит к пропорциональному смещению ICC вверх на величину c_j.

*В.Аванесов отмечает, что F.Lord называл c_j параметром псевдоугадывания. Это указывает на то, что в величину c_j дают вклад и другие факторы, помимо угадывания.

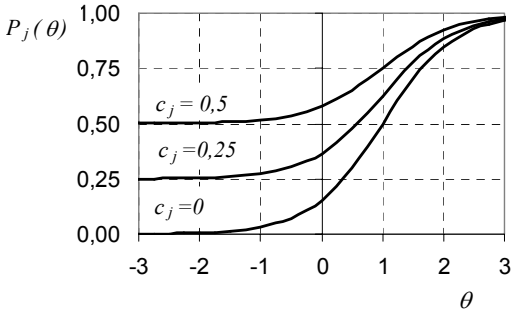


Рис.5.2.4. ICC в трехпараметрической модели 3PL, a_j=1, β_j=1.

В качестве теоретической оценки c_j можно использовать обратную величину от количества ответов в заданиях с выбором. Например, в тесте используются задания с четырьмя ответами, тогда c_j = 1/4 = 0,25. Это значение должно уточняться при анализе эмпирических данных.

МОДЕЛЬ RASCH MEASUREMENT

Обсудим вопрос о степени пригодности моделей IRT для целей измерения латентных параметров.

Характерной особенностью модели Раша является то, что характеристические кривые (ICC) не пересекаются (рис.5.2.1). Это означает, что если некоторое задание «А» легче задания «Б», то это соотношение сохраняется во всем интервале изменения θ.

Совершенно иная картина наблюдается для двух- и трехпараметрической моделей. На рис.5.2.3 это хорошо видно. Задание с a_j = 0,5 в области положительных значений θ является самым трудным из представленных трех заданий, то есть вероятность правильного ответа на это задание самая низкая. В области же отрицательных значений θ это же задание теперь уже самое легкое - вероятность правильного ответа на него наибольшая. Получается, что для слабых учащихся это самое легкое задание, а для сильных учащихся - самое трудное.

Аналогичная картина наблюдается и для трехпараметрической модели. На рис.5.2.4. показан редкий случай непересекающихся характеристических кривых, так как для них выбраны одинаковые параметры β_j=1 и a_j=1, то есть все три задания имеют одинаковую

трудность и одинаковый параметр дифференцирующей способности.

На рис.5.2.5 приведен другой пример.

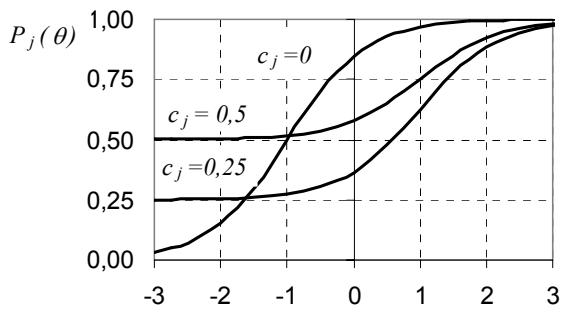


Рис.5.2.5. Пересекающиеся ICC в трехпараметрической модели.

Здесь у задания с параметром c_j=0 изменена трудность β_j= -1, что немедленно вызвало пересечение характеристических кривых. Задание с c_j=0 в области θ < -2 является самым трудным. В области -1,5 < θ < -1 это задание легче задания с c_j=0,25 и труднее задания с c_j=0,5. В области θ > -1 задание с c_j=0 является самым легким.

Подобное пересечение ICC практически всегда происходит для двух- и трехпараметрической моделей.

Таким образом, только однопараметрическая модель Раша соответствует требованиям, предъявляемым к качественному измерительному инструментарию. Именно модель RASCH MEASUREMENT больше всего пригодна для построения теста, как измерительного инструмента.

5.3. ВЫЧИСЛЕНИЕ θ_i и β_j ИЗ ЭМПИРИЧЕСКИХ ДАННЫХ

Рассмотрим процедуру вычисления θ_i и β_j из эмпирических данных. В качестве исходных данных возьмем бинарную матрицу из таблицы 3.2.5. Дальнейшие расчеты выполним, следуя М.Челышковой[6].

Сначала необходимо вычислить доли верных p_i и неверных $q_{i=1-}$ p_i ответов испытуемых.

$$p_i = \frac{X_i}{M}$$

где X_i - индивидуальный балл испытуемого, M - количество заданий в тесте.

Например, для 2-го испытуемого имеем

$$p_2 = \frac{6}{8} = 0{,}75$$

$q_2 = 1 - p_2 = 1 - 0{,}75 = 0{,}25$.

Далее вычисляем начальные значения уровня подготовленности испытуемых по формуле

$$\theta_i^{\,0} = \ln \frac{p_i}{q_i}$$

Для 2-го испытуемого имеем

$$\theta_2^{\,0} = \ln \frac{0{,}75}{0{,}25} = 1{,}099$$

Аналогичные расчеты выполняются для всех десяти испытуемых (таблица 3.2.5) и заносятся в таблицу 5.3.1.

Далее вычисляем начальное значение трудности заданий β_j.

Здесь j пробегает значения от 1 до M, где M -количество

$$\beta_j^{\,0} = \ln \frac{q_j}{p_j}$$

испытуемых. В качестве примера рассчитаем начальное значение трудности 2-го задания. Величины p_j и q_j рассчитаны нами ранее и приведены в таблице 3.2.5.

$$\beta_2^{\,0} = \ln \frac{q_2}{p_2} = \ln \frac{0{,}3}{0{,}7} = -0{,}847$$

Таблица 5.3.1. Начальные значения уровня подготовленности испытуемых

i	X_i	p_i	q_i	θ_i^0	$(\theta_i^0)^2$
1	7	0,875	0,125	1,946	3,786
2	6	0,750	0,250	1,099	1,207
3	6	0,750	0,250	1,099	1,207
4	6	0,750	0,250	1,099	1,207
5	4	0,500	0,500	0,000	0
6	3	0,375	0,625	-0,511	0,261
7	2	0,250	0,750	-1,099	1,207
8	2	0,250	0,750	-1,099	1,207
9	1	0,125	0,875	-1,946	3,786
10	1	0,125	0,875	-1,946	3,786
				$\sum(\theta_i^0)^2 =$	17,655

Расчеты для всех восьми заданий сведены в таблицу 5.3.2.

Таблица 5.3.2. Начальные значения трудности заданий

j	R_j	p_i	q_i	β_j^0	$(\beta_j^0)^2$
1	7	0,700	0,300	-0,847	0,718
2	7	0,700	0,300	-0,847	0,718
3	6	0,600	0,400	-0,405	0,164
4	5	0,500	0,500	0,000	0
5	5	0,500	0,500	0,000	0
6	4	0,400	0,600	0,405	0,164
7	2	0,200	0,800	1,386	1,922
8	1	0,100	0,900	2,197	4,828
				$\sum(\beta_j^0)^2 =$	8,514

Теперь мы можем вычислить средние значения уровня подготовленности испытуемых и трудности заданий.

$$\overline{\theta} = \frac{\sum_{i=1}^{N}\theta_i^0}{N} = \frac{1{,}946 + 1{,}099 + 1{,}099 + 1{,}099 + 0 - 0{,}511 - 1{,}099 - 1{,}099 - 1{,}946 - 1{,}946}{10} = -0{,}136$$

$$\overline{\beta} = \frac{\sum_{j=1}^{M}\beta_j^0}{M} = \frac{-0{,}847 - 0{,}847 - 0{,}405 + 0 + 0 + 0{,}405 + 1{,}386 + 2{,}197}{8} = +0{,}236$$

В таблицах 5.3.1. и 5.3.2 мы имеем значения параметров на разных интервальных шкалах. Нам надо свести их в единую шкалу стандартных оценок. Для этого необходимо вычислить дисперсии S_θ и S_β, используя данные из таблиц 5.3.1 и 5.3.2.

$$S_\theta = \frac{\sum_{i=1}^{N}\left(\theta_i^0\right)^2 - N\left(\overline{\theta}\right)^2}{N-1} = \frac{17{,}655 - 10\cdot(-0{,}136)^2}{10-1} = 1{,}941$$

$$S_\beta = \frac{\sum_{j=1}^{M}\left(\beta_j^0\right)^2 - M\left(\overline{\beta}\right)^2}{M-1} = \frac{8{,}514 - 8\cdot(0{,}236)^2}{8-1} = 1{,}153$$

Далее вычисляем угловые коэффициенты

$$a_\theta = \sqrt{\frac{1 + S_\beta/2{,}89}{1 - S_\theta S_\beta/8{,}35}} = \sqrt{\frac{1 + 1{,}153/2{,}89}{1 - 1{,}941\cdot 1{,}153/8{,}35}} = 1{,}911$$

$$a_\beta = \sqrt{\frac{1 + S_\theta/2{,}89}{1 - S_\theta S_\beta/8{,}35}} = \sqrt{\frac{1 + 1{,}941/2{,}89}{1 - 1{,}941\cdot 1{,}153/8{,}35}} = 2{,}284$$

Наконец, мы можем записать оценки параметров θ и β на единой интервальной шкале[10].

$$\theta_i = a_\theta\theta_i^0 + \overline{\beta},$$
$$\beta_j = a_\beta\beta_j^0 + \overline{\theta},$$

Для нашего примера получим

$\theta_i = 1{,}911\cdot\theta_i^0 + 0{,}236$

$\beta_j = 2{,}284\cdot\beta_j^0 - 0{,}136$

Все результаты сведены в таблицы 5.3.3 и 5.3.4 (второй столбец).

Из таблицы 5.3.4. следует, что

$$\Sigma_\beta = \sum_{j=1}^{N}\beta_j = +3{,}226$$

То есть, заданий с положительными β_j больше, чем с отрицательными. Данный тест не сбалансированный, он содержит больше трудных заданий, чем легких.

Рекомендуется стремиться к тому, чтобы $\sum_\beta$ было близко к нулю.

Нам осталось вычислить стандартные ошибки измерения $S_E(\theta_i)$ и $S_E(\beta_j)$ для θ_i. и β_j

Таблица 5.3.3. Расчетные параметры для уровня подготовленности испытуемых

i	θ_i	$S_E(\theta_i)$	θ_i'
1	3,955	2,043	2,436
2	2,335	1,560	1,365
3	2,335	1,560	1,365
4	2,335	1,560	0,523
5	0,236	1,351	-0,157
6	-0,740	1,396	-0,781
7	-1,863	1,560	-1,431
8	-1,863	1,560	-1,431
9	-3,483	2,043	-2,217
10	-3,483	2,043	-2,217

Таблица 5.3.4. Расчетные параметры для трудности заданий теста

j	β_j	$S_E(\beta_j)$	β_j'
1	-2,071	1,576	-1,545
2	-2,071	1,576	-1,669
3	-1,062	1,474	-0,603
4	-0,136	1,445	-0,502
5	-0,136	1,445	-0,256
6	0,790	1,474	0,102
7	3,030	1,806	1,854
8	4,882	2,408	2,620

$$S_E(\theta_i)=\frac{a_\theta}{\sqrt{p_i(M-X_i)}}=\frac{a_\theta}{\sqrt{Mp_i(1-p_i)}}=\frac{a_\theta}{\sqrt{Mp_iq_i}}$$

$$S_E(\beta_j)=\frac{a_\beta}{\sqrt{p_j(N-R_j)}}=\frac{a_\beta}{\sqrt{Np_j(1-p_j)}}=\frac{a_\beta}{\sqrt{Np_jq_j}}$$

Например, для первого испытуемого получим

$$S_E(\theta_1)=\frac{1{,}911}{\sqrt{8\cdot 0{,}875\cdot 0{,}125}}=2{,}043$$

Для первого задания стандартная ошибка равна

$$S_E(\beta_1)=\frac{2{,}284}{\sqrt{10\cdot 0{,}7\cdot 0{,}3}}=1{,}576$$

Вычисленные значения стандартных ошибок приведены в таблицах 5.3.3 и 5.3.4 (третий столбец).

5.4. МЕТОД НАИБОЛЬШЕГО ПРАВДОПОДОБИЯ

Значения вычисленных параметров θ_i и β_j могут измениться на других выборках испытуемых. При больших объемах выборки можно вычислить значения θ_i' и β_j', к которым в результате итерационной процедуры, будут стремиться θ_i и β_j.

Обычно итерационная процедура выполняется методом наибольшего правдоподобия Р.Фишера.

Можно показать, что функция правдоподобия имеет вид[2]

$$L(a_{ij},\theta_i,\beta_j)=\exp\left[\sum_{i=1}^{N}\sum_{j=1}^{M}a_{ij}(\theta_i-\beta_j)\right]\cdot\left[\prod_{i=1}^{N}\prod_{j=1}^{M}(1+\exp(\theta_i-\beta_j))\right]^{-1}$$

где a_{ij} - элементы бинарной матрицы результатов тестирования.

В качестве оценок наибольшего правдоподобия θ_i' и β_j'. принимают такие значения θ_i и β_j, при которых функция правдоподобия достигает глобального максимума. Поскольку функции L и $ln\ L$ достигают максимума при одних и тех же значениях своих аргументов, то удобно искать максимум функции $ln\ L$, называемой логарифмической функцией правдоподобия

$$\ln L=\sum_{i=1}^{N}\sum_{j=1}^{M}a_{ij}\theta_i-\sum_{j=1}^{N}\sum_{i=1}^{M}\beta_j-\sum_{i=1}^{N}\sum_{j=1}^{M}\ln\left[1+\exp(\theta_i-\beta_j)\right]$$

Для нахождения максимума логарифмической функции правдоподобия надо найти частные производные функции по каждому ее аргументу и приравнять нулю

Мы получили систему уравнений правдоподобия. Эта система уравнений решается в итерационном цикле путем последовательной

$$\frac{\partial\ln L}{\partial\theta_i}=\sum_{j=1}^{M}a_{ij}-\sum_{i=1}^{N}p_{ij}=0$$

$$\frac{\partial\ln L}{\partial\beta_j}=-\sum_{i=1}^{N}a_{ij}+\sum_{j=1}^{M}p_{ij}=0$$

подстановки найденных значений аргументов в качестве исходных.

Цикл прерывается, когда различие в аргументах не станет меньше наперед заданной величины. Система уравнений правдоподобия нелинейна и для организации итерационного цикла требует применения вычислительной техники.

Результаты вычислений для нашего примера приведены в таблицах 5.3.3 и 5.3.4 (четвертый столбец). Видно, что данные во втором и четвертом столбцах заметно различаются. Это связано с тем, что наша модельная выборка недопустимо мала. При больших выборках это различие невелико.

Кроме метода наибольшего правдоподобия существуют и другие методы нахождения устойчивых оценок латентных параметров. В частности, A.J.Stenner, B.D.Wright & J.M.Linacre[11] предложили другую итерационную процедура, время прохождения которой, в среднем, в два раза меньше, чем в методе наибольшего правдоподобия[12].

5.5. ПОСТРОЕНИЕ ХАРАКТЕРИСТИЧЕСКИХ КРИВЫХ ДЛЯ ЗАДАНИЙ ТЕСТА (ICC)

Используя таблицы 5.3.3 и 5.3.4, мы можем построить характеристические кривые испытуемых (PCC) и заданий (ICC). В качестве примера рассмотрим построение ICC. Для этого используем формулу (5.2.1), где β_j будет параметром, а θ - переменной величиной.

График ICC строится, например, по 15-20 точкам. Рассмотрим процесс определения первой точки для первого задания. Из таблицы 5.3.4 видим, что первое задание имеет трудность $\beta_j^{'}$= -1,545. Пусть переменная θ будет меняться в интервале от -5 до +5 с шагом 0,5 логита. Тогда у нас получится 21 точка. Точке №1 соответствует θ= -5. Вычислим значение вероятности успеха испытуемого с уровнем подготовленности -5 для задания с трудностью -1,545.

$$P_1(-5) = \frac{e^{1,7(-5-(-1.545))}}{1+e^{1,7(-5-(-1.545))}} = 0,0028$$

Аналогично вычисляются значения P_2(-4,5), P_3(-4), … P_{21}(+5). Полученные точки соединяются плавной кривой. Затем точно также рассчитывается характеристическая кривая для второго задания с трудностью $\beta_j^{'}$= -1,669 и т.д.

Результаты по всем заданиям приведены на рис.5.5.1.

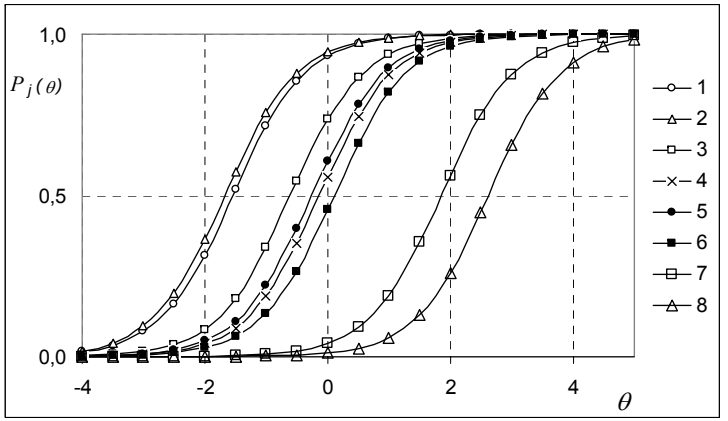


Рис.5.5.1. Характеристические кривые (ICC) для 8-ми заданий.

Из приведенных графиков видно, что задания неравномерно покрывают требуемый диапазон уровней подготовленности испытуемых (от -4 до +4), особенно в области +1 логита. С другой стороны, задания 1, 2 и 4, 5 дублируют друг друга, Часть из них можно удалить из теста без ущерба его общей дифференцирующей способности.

Представляет интерес сравнить эти теоретические кривые с нашими экспериментальными данными (рис.5.5.2 и 5.5.3).

Методика сопоставления экспериментальных данных с моделью Раша рассмотрена далее в параграфе 5.7.

На рисунках приведены в качестве примера, данные для двух заданий (№5 и №6).

Видно, что задание №6 не соответствует модели Раша. Это может быть обусловлено как несовершенством задания (по форме и/или по содержанию), так и нарушениями в процедуре тестирования. Более подробно анализ результатов тестирования в модели Раша будет рассмотрен далее.

Кратко рассмотрим еще один вопрос - о сопоставлении эмпирических данных.

Поскольку логиты размещены на интервальной шкале, то отсутствует понятие абсолютного нуля. За нуль можно взять любую точку на шкале логитов, то есть характеристические кривые

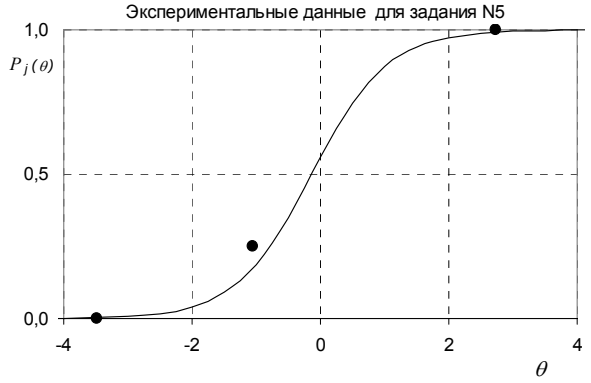


Рис.5.5.2

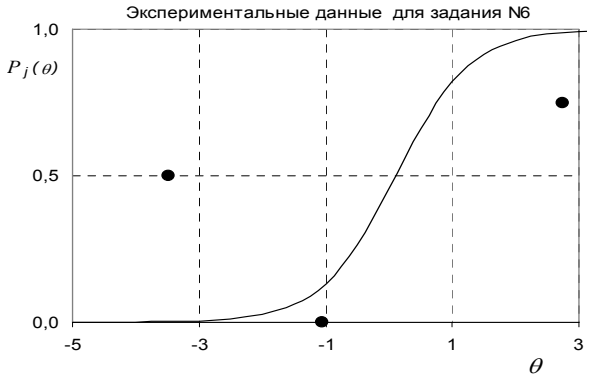


Рис.5.5.3

инвариантны относительно сдвига на заданную константу. Отсутствие нуля приводит к необходимости «сшивания» результатов разных тестов.

Для этой цели используются «узловые»[2] или «якорные»[13] задания. Якорные задания обеспечивают перекрытие тестов. Рекомендуется выбирать три задания в качестве якорных, причем два задания должны располагаться на краях метрической шкалы, а третье примерно посередине. Крайние якоря желательно размещать в области 1,5 - 2,5 логита[2]. Отметим, что наши исследования[14,15] показывают, что якорные задания лучше размещать вблизи 2,5 - 3 логитов.

5.6. ИНФОРМАЦИОННАЯ ФУНКЦИЯ

Согласно А.Бирнбауму[9] количество информации, обеспеченное *j*-м заданием теста в данной точке θ_i - это величина, обратно пропорциональная стандартной ошибке измерения данного значения θ_i с помощью *j*-го задания. Для описания информации, соответствующей заданию вводится информационная функция $I(\theta)$[9, 16].

$$I_j(\theta) = \frac{\left(P_j'(\theta)\right)^2}{P_j(\theta) \cdot Q_j(\theta)}$$

Для однопараметрической модели $P_j' = 1{,}7 P_j Q_j$, тогда

$I_j(\theta) = 2{,}89 P_j(\theta) Q_j(\theta)$, где $Q_j(\theta) = 1 - P_j(\theta)$-вероятность неверного ответа на *j*-е задание. Поскольку

$$Q_j(\theta) = \frac{1}{\left(1 + e^{1{,}7(\theta - \beta_j)}\right)}$$

то выражение для информационной функции перепишем в следующем виде

$$I_j(\theta) = 2{,}89 \frac{e^{1{,}7(\theta - \beta_j)}}{\left(1 + e^{1{,}7(\theta - \beta_j)}\right)^2} \qquad (5.6.1)$$

Для двухпараметрической модели
$I_j(\theta) = 2{,}89 a_j^2 P_j(\theta) Q_j(\theta)$,

В трехпараметрической модели информационная функция имеет вид[6]

$$I_j(\theta) = \frac{2{,}89 a_j^{\,2} (1 - c_j)}{\left(c_j + e^{1{,}7 a_j (\theta - \beta_j)}\right) \cdot \left(1 + e^{-1{,}7 a_j (\theta - \beta_j)}\right)^2}$$

Отметим, что численный коэффициент 2,89 появился из-за наличия масштабного множителя 1,7. Если считать его равным единице, то, например, для однопараметрической модели получим (F.Baker)[18] $I_j(\theta) = P_j(\theta) Q_j(\theta)$. В этом случае максимальное значение информационной функции равно 0,25.

Построим информационную функцию для однопараметрической модели, используя выражение (5.6.1)

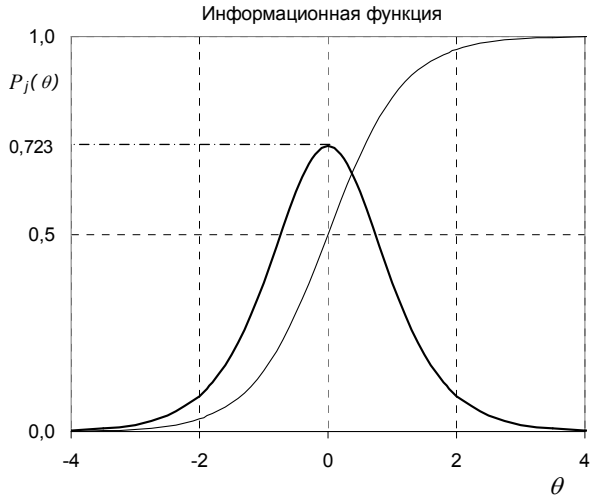


Рис.5.6.1. Информационная функция задания с нулевой трудностью.

На рис.5.6.1. показана характеристическая кривая задания с уровнем трудности $\beta=0$ и информационная функция для этого задания.

Видно, что максимум информационной функции достигается при таком значении θ, когда имеет место перегиб характеристической кривой задания, то есть вероятность выполнения задания равна 0,5.

Таким образом, задание наиболее информативно, когда его трудность примерно равна уровню подготовленности испытуемого.

Информационные функции обладают свойством аддитивности[18]

$$I(\theta) = \sum_{j=1}^{N} I_j(\theta)$$

Это означает, что можно построить информационную функцию всего теста. На рис.5.6.2. приведен пример теста из трех заданий с трудностями -1 (график №1), 0 (график №2) и +1 (график №3). Для этих трех информационных функций тестовых заданий построена информационная функция всего теста (график №4).

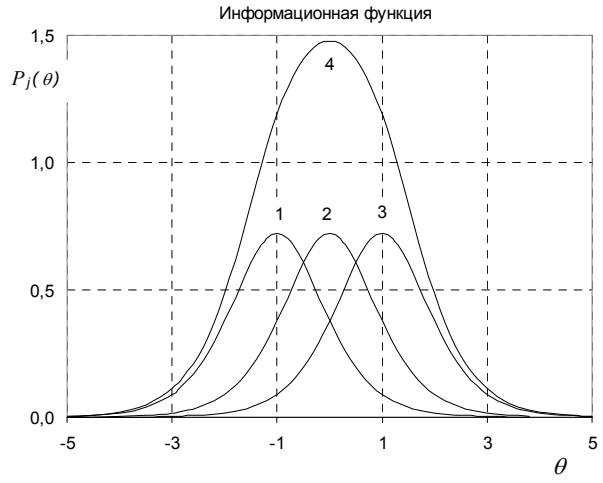


Рис.5.6.2. Информационная функция «хорошего» теста.

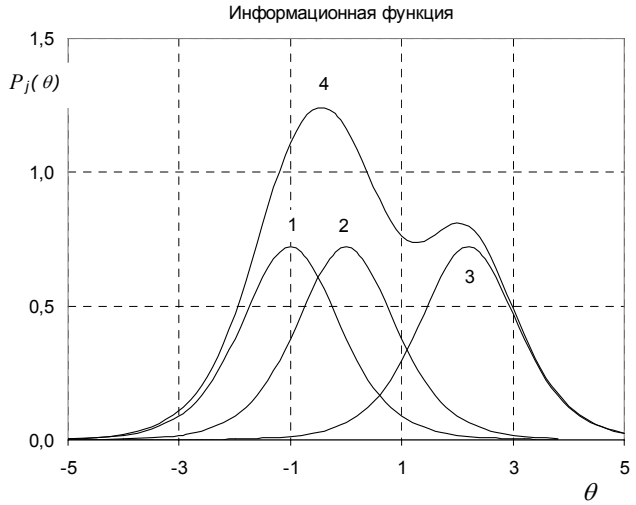


Рис.5.6.3. Информационная функция «плохого» теста.

Информационная функция теста должна иметь один четко выраженный максимум (рис.5.6.2). Если это не так, то тест нуждается в доработке, в него необходимо добавить задания с трудностями, соответствующими областями провала информационной функции теста.

На рис.5.6.3. приведены информационные функции заданий с трудностями -1 (график №1), 0 (график №2) и 2,2 (график №3). Для теста, состоящего из этих трех заданий, информационная функция (график №4) имеет два максимума. Этот тест явно нуждается еще в заданиях с трудностями в области +1 логит. Улучшения информационной функции теста можно добиться и не изменяя числа заданий в нем. Для этого необходимо сдвинуть задание №3 влево, то есть уменьшить его трудность.

Характер информационной функции для двух- и трехпараметрической моделей в целом сохраняется.

В двух параметрической модели наклон характеристической кривой в точке $\theta = \beta$ может заметно превышать величину 0,25 для однопараметрической модели. С одной стороны это хорошо, так как информационная функция имеет довольно острый пик, а с другой – уменьшается рабочая область задания. Если в однопараметрической модели одно задание удовлетворительно перекрывает диапазон $\beta \pm 1$ логит, то в двухпараметрической модели для этого же диапазона могут потребоваться два и более задания в зависимости от величины параметра дифференцирующей способности a_j.

В трехпараметрической модели параметр псевдоугадывания c_j заметно снижает точность оценок θ и β, а также замедляет сходимость итерационных процедур, используемых для поиска устойчивых значений θ и β. Информационная функция достигает максимума в точке[16]

$$\theta_{\max} = \beta_j + \frac{1}{3{,}4a_j}\left(1 + \sqrt{1 + 8c_j}\right)$$

Расчеты информационной функции показывают, что с уменьшением c_j происходит рост информативности задания. Это ясно и из общих соображений. Если c_j характеризует угадывание, то ее большая величина свидетельствует об очень большом вкладе угадывания в результаты тестирования. Естественно, что эти результаты имеют мало общего с реальными знаниями испытуемых, то есть информативность теста в этом случае очень низкая.

5.7. АНАЛИЗ РЕЗУЛЬТАТОВ ТЕСТИРОВАНИЯ НА ОСНОВЕ RASCH MEASUREMENT

Педагогический тест, как средство измерения учебных достижений, может дать достоверный результат только в случае его корректного применения. Корректность применения теста – это многоаспектное понятие, включающее в себя вопросы конструирования и дизайна теста, вопросы разработки и применения тестов и, разумеется, интерпретации результатов тестирования. В данной работе основное внимание уделено вопросам корректности интерпретации результатов педагогического тестирования, проводимого на основе модели Г.Раша. Анализ результатов обычно проводится на основе классической теории тестов или на основе Item Response Theory.

После выполнения работ по созданию теста и сбора данных на репрезентативной выборке испытуемых, производится интерпретация результатов. Этот этап принципиально отличается от технологии, принятой, скажем в экспериментальной физике. Там экспериментальные данные пытаются описывать с помощью той или иной теории. Если теоретическая зависимость между исследуемыми величинами не соответствует наблюдаемой в эксперименте, то делается вывод, что теория недостаточно развита и требует дальнейшей разработки. В теории педагогических измерений может применяться иной подход. Если в физике законы природы не зависят от исследователя, то тесты в немалой степени зависят от его воли. Это принципиально важный момент.

IRT в настоящий момент является общепризнанной теорией. В качестве латентных параметров модели выступают как характеристики тестируемых, так и самого теста. Ю.Нейман и В.Хлебников[2] делают вывод, что «…уникальность моделей семейства Г.Раша состоит в том, что они задают определенный механизм преобразования формальных наблюдений за исходом событий в объективные измерения на метрической шкале латентных стимулов этих событий». Это очень важно, так как недостаточно глубокое осознание этого факта, может приводить к тому, что положения педагогических измерений могут критически восприниматься специалистами в области точных наук

Таким образом, несоответствие эмпирических данных модели Раша означает, что, например, имеются неточности в формулировке заданий, были нарушения в процедуре тестирования и т.д. Как отмечает В.Аванесов[17], в литературе можно встретить немало критики по поводу неприменимости модели Раша к множеству «тестов», и

поэтому ведется поиск других моделей, более адекватных полученным результатам. Но здесь есть один очень важный вопрос. В теории Г.Раша никогда не ставилась задача адекватного описания данных. Напротив, это пример другой философии измерения - model based measurement, где утверждается противоположное – не модель должна соответствовать эмпирическим данным, а данные должны соответствовать модели. Об этом можно спорить, но в соответствии с философией Rasch шкалу (педагогический тест) образуют только те задания, которые отвечают данной модели измерения. Все остальные в тест не включаются.

Итак, при анализе результатов тестирования, нам необходимо проверить соответствие эмпирических данных модели Раша.

Согласно Ф.Бейкеру[18] для этого всех N тестируемых, выполняющих M заданий теста распределяют по шкале θ (ability) по своим диапазонам уровня подготовленности. Испытуемые делятся на J групп вдоль шкалы θ так, чтобы все тестируемые внутри данной группы имели одинаковый уровень подготовленности θ_j. Всего внутри группы с номером j окажутся m_j тестируемых, где j принимает значения из интервала $j = 1,2,3,\ldots,J$.

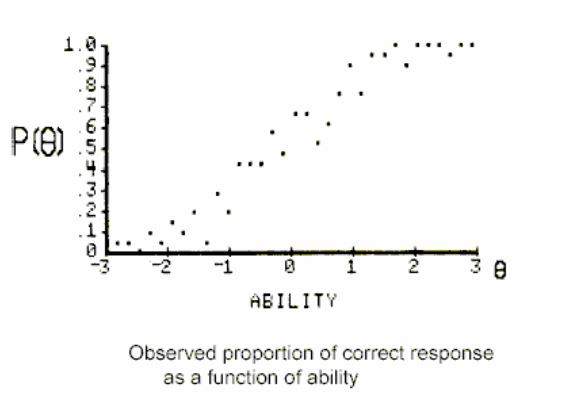
В пределах каждой группы r_j тестируемых отвечают правильно на данное задание теста. Таким образом, для уровня подготовленности (уровня знаний) равного θ_j вероятность правильного ответа на данное задание равна

$$p(\theta_j) = \frac{r_j}{m_j}$$

Величина $p(\theta_j)$ является экспериментальным значением вероятности правильного ответа на данное задание. На рисунке 5.7.1 показаны данные из работы Ф.Бейкера[18].

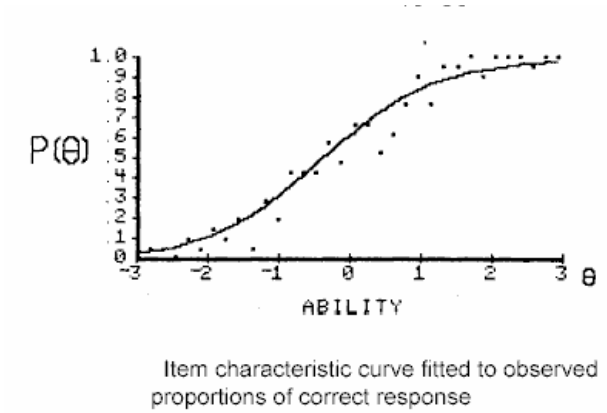
На следующем этапе проверяется, насколько хорошо эмпирические данные описываются IRT-моделью. Результат сравнения показан на рисунке 5.7.2.

Из рисунка 5.7.2 видно, что наблюдается хорошее согласие эмпирических данных с IRT. В целом задача разработчика тестов состоит в том, чтобы разработать такие тестовые задания и так осуществить процедуру тестирования, чтобы получить результаты, аналогичные тем, что показаны на рисунке 5.7.2.



Observed proportion of correct response as a function of ability

Рис. 5.7.1.



Item characteristic curve fitted to observed proportions of correct response

Рис. 5.7.2.

Проанализируем результаты тестирования учащихся средних общеобразовательных учреждений по теме «Механика» учебной дисциплины «Физика». Нормативно-ориентированный тест содержал 30 заданий закрытого типа заданной специфической формы. Всего было протестировано 60 испытуемых, т.е. использовалась бинарная матрица размером 30х60. После упорядочения матрицы по строкам и столбцам по стандартной процедуре были рассчитаны логистические кривые по модели Раша. Для этого использовалась методика, подробно описанная в предыдущих параграфах данной главы.

Для модели 1PL вероятность успеха в j-м задании равна

$$P_j = \frac{1}{1+e^{-d(\theta-\beta_j)}}$$

где d – фактор шкалирования, равный 1,702.

На рисунке 5.7.3 приведены результаты расчетов для всех 30 заданий теста. Экспериментальные значения P_j, полученные по методике[18], приведены на рисунках 5.7.4-5.7.7. Экспериментальные данные показаны выборочно для четырех заданий различного уровня

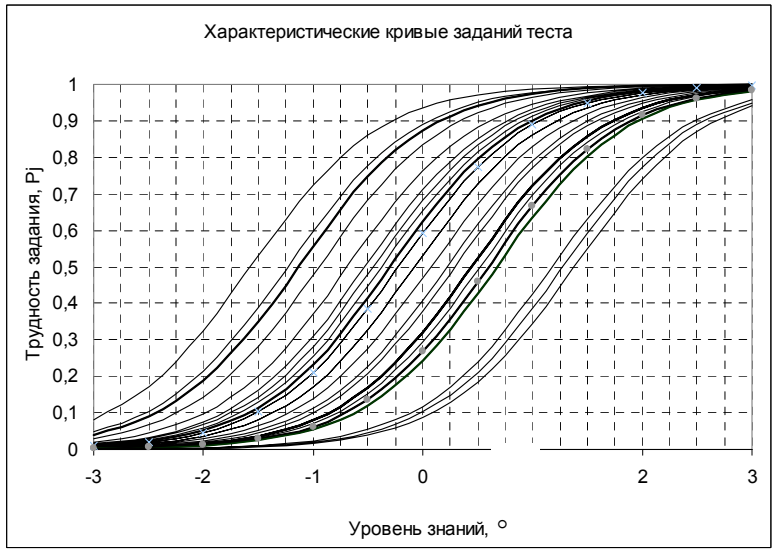


Рис.5.7.3.

трудности – 3, 8, 20, 30 задания.

По результатам тестирования сразу можно получить матрицу, анализируя которую можно избавить тест от некоторых неподходящих заданий[19]. Дальнейшие расчеты возможны в трех вариантах:

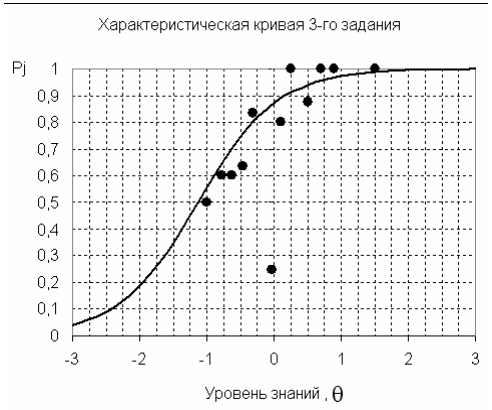


Рис.5.7.4.

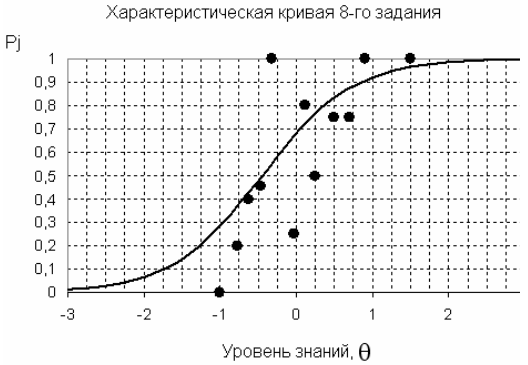


Рис.5.7.5.

Однопараметрическая логистическая модель (1PL), или модель Раша[20];

Двухпараметрическая логистическая модель (2PL) Бирнбаума;

Трехпараметрическая логистическая модель (3PL) Бирнбаума.

Как известно, модели 2PL и 3PL предлагались для лучшего

согласования теории с наблюдаемыми эмпирическими данными. Если считать, что согласования следует добиваться не видоизменением теории, а получением других эмпирических данных, то следует принять модель Раша.

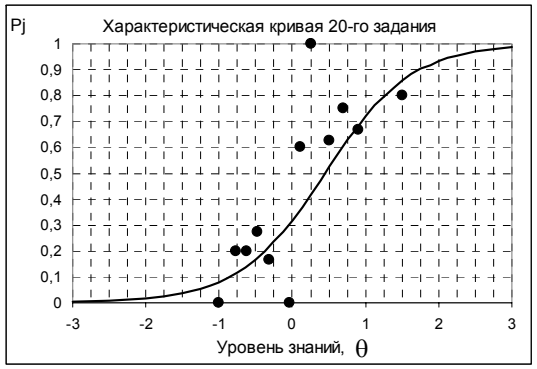


Рис.5.7.6.

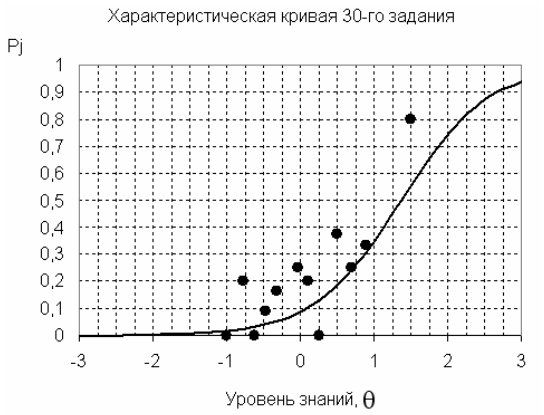


Рис.5.7.7.

Иными словами, если экспериментальные данные не соответствуют модели Раша, то необходимо переработать тестовые задания и повторно провести эксперимент, добиваясь лучшего

согласия с теорией, как указывалось выше.

Следуя такой парадигме, в данной работе все построения проводились по модели Г.Раша.

Из рисунка 5.7.3 видно, что задания теста по шкале уровня знаний θ перекрывают диапазон примерно от -3,5 до +3,5 логитов. Графики показаны последовательно слева- направо от 1-го (самого легкого) до 30-го самого трудного задания. Характеристические кривые некоторых заданий, а именно 3 и 4; 10 и 11; 13, 14 и 15; 19, 20 и 21; 23 и 24; 25, 26 и 27 перекрываются. В связи с этим 4, 11, 14, 15, 19, 21, 24, 26, 27 задания могут быть удалены из теста без ущерба его измерительным свойствам.

На семействе характеристических (логистических) кривых тестовых заданий отчетливо наблюдается явная недостаточность отдельных заданий. Наличие «провалов» в монотонной последовательности характеристических кривых указывает на необходимость дополнительной оптимизации теста путем добавления новых тестовых заданий или переработки имеющихся. Переработкой тестовых заданий необходимо добиться появления добавочных характеристических кривых в интервале от -1,5 до - 0,5 и от +0,7 до +1,2 логита (на уровне $P_j = 0{,}5$).

Экспериментальные данные для P_j имеют примерно одинаковое согласие с моделью Раша, которое можно считать удовлетворительным. Приведенные на рисунках 4-7 характеристические кривые некоторых заданий иллюстрируют это. При анализе вся совокупность тестируемых разбивалась на 12 групп (J=12).

Экспериментальные точки для характеристической кривой 3-го задания группируются в области от -1 до +1 логита для P_j от 0,5 до 1,0. Это относительно легкое задание и экспериментальные точки приблизительно соответствуют верхнему участку характеристической кривой. Задания 8 и 20 находятся примерно в средней области тестовых заданий (рисунок 5.7.3) и соответствуют заданиям средней сложности. Экспериментальные точки в этом случае группируются вблизи линейной области характеристических кривых P_8 и P_{20}

Задание №30 самое трудное и экспериментальные точки в основном сосредоточены вблизи нижнего загиба характеристической кривой P_{30}.

Для проверки гипотезы H_o на соответствие полученных эмпирических данных одномерной модели IRT для всех заданий теста проводилось вычисление критерия χ^2 согласно[18]

Расчетное значение критерия χ^2 оказалось в пределах от 7

$$\chi^2 = \sum_{j=1}^{J} m_j \frac{(p(\theta_j) - P(\theta_j))^2}{P(\theta_j)Q(\theta_j)}$$

до 15 для различных заданий теста.

Таким образом, несмотря на довольно заметный разброс данных, что вероятнее всего обусловлено недостаточной репрезентативностью выборки (60 испытуемых), все же можно констатировать более или менее удовлетворительное согласие экспериментальных результатов с одномерной моделью IRT.

5.8. АНАЛИЗ РЕЗУЛЬТАТОВ ТЕСТИРОВАНИЯ В RUMM

При разработке тестовых заданий важно оценить их качество, что делается в рамках той или иной модели. В данном параграфе продолжен[21],[22] анализ тестовых заданий, выполняемый на основе теории G.Rasch. Для анализа данных использовались задания по учебной дисциплине «Базы данных» (Федеральный компонент, ОПД.Ф.03) и программное средство RUMM (Rasch Unidimensional Measurement Model), разработанное под руководством профессора D.Andrich[23]. Программное средство RUMM успешно используется для оценки качества тестовых заданий во многих странах мира.

Для однопараметрической модели измерения (1PL) вероятность успеха *i*-го испытуемого в *j*-м задании равна

$$P_{ij} = \frac{1}{1 + e^{-d(\theta_i - \beta_j)}}$$

где d – коэффициент шкалирования, равный 1,702.

θ_i (ability) - степень подготовленности испытуемого

Исходный набор заданий содержал 72 задания с выбором одного правильного ответа из четырёх, предлагавшихся на выбор. Всего было протестировано 40 испытуемых.

Все испытуемые были распределены по шкале θ по своим диапазонам уровня подготовленности. Испытуемые были поделены на K групп (классов интервалов) вдоль шкалы θ так, чтобы все тестируемые внутри данной группы имели одинаковый уровень подготовленности θ_k. Всего внутри группы с номером k окажутся m_k тестируемых, где k принимает значения из интервала $k = 1,2,3,\ldots,K$.

В RUMM-2020 значение K по умолчанию устанавливается равным 3, но при необходимости его можно изменить, используя параметр «Class_Intervals» в диалоговом окне «Analysis Control». Чем большим берётся число классов, тем больше «эмпирических» точек представляется на графике заданий. Однако в этом случае требуется иметь и большее число испытуемых. Вот почему при небольшом числе испытуемых минимально допустимым принимается число классов, равное трём, так как по двум точкам на теоретической кривой ICC трудно судить о соответствии задания модели Раша.

В настоящем исследовании в первый классовый интервал данных были включены 13 испытуемых, во второй - 12 и в третий - 15 испытуемых. Точки, соответствующие этим интервалам, имеют

значения θ, равные соответственно 0.636, 2.751 и 3.638.

Далее в работе приведены рисунки с изображениями характеристических кривых некоторых заданий - Item Characteristics Curves (ICC). Для каждой группы приведены примеры ICC для двух заданий. На рис.5.8.1 приведены примеры ICC, иллюстрирующие смысл некоторых параметров, характеризующих ICC. На рис.5.8.2-5.8.9 приведены ICC для заданий анализируемого теста.

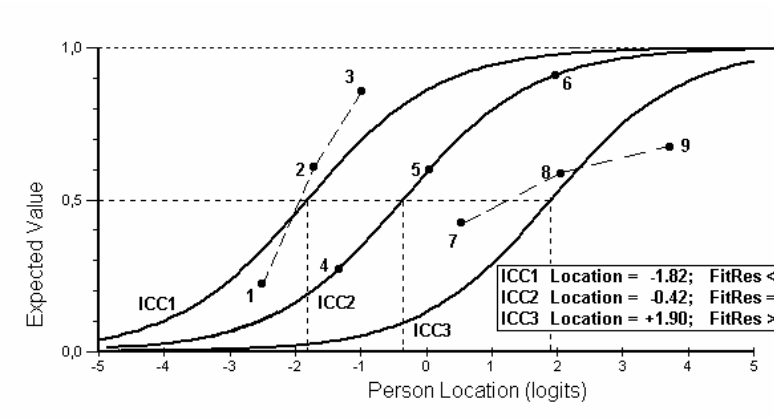


Рис. 5.8.1. Пример характеристических кривых трех заданий.

Для оценки степени соответствия данных модели Раша в RUMM2020 используется распределение хи-квадрат ($\chi^2_{probability}$). Чем ближе значения этого распределения к единице, тем лучше соответствие данных модели. соответственно, чем ближе к нулю, тем хуже соответствие задания модели измерения по теории Раша.

Для каждой кривой приведены следующие параметры, которые рассмотрим на примере ICC-53 (рис. 5.8.2). Описание основных свойств параметров приведено в работе А.Маслака[4]. Ниже следует это описание с нашими дополнениями.

I0053 - код (идентификатор) тестового задания;

Descriptor for item 53 - заголовок (название) тестового задания 53. Вообще-то при редактировании вводимых данных в RUMM можно использовать в качестве названия произвольный текст. В данном случае выбрано значение по умолчанию;

Locn = 0,902 - трудность тестового задания в логитах.

На рис. 5.8.1 в качестве примера показаны три характеристические кривые ICC1, ICC2, ICC3 для некоторых заданий со значениями Locn (Location) равными -1.82, -0.42, +1.90. Для каждой

кривой Locn – это значение Person location, при котором вероятность правильного ответа на данное задание равно 0.5.

FitRes = -0,358 - суммарное отклонение ответов испытуемых на данное задание от ожидаемых на основе модели Раша. Если параметр FitRes = 0, то мы имеем совпадение ответов испытуемых с моделью Раша. Большие по абсолютной величине значения FitRes свидетельствуют о расхождении экспериментальных данных с моделью Раша. Схематически это показано на рис. 5.8.1, где в качестве примера показаны характеристические кривые, имеющие различные значения параметра FitRes.

Для характеристической кривой ICC1 экспериментальным точкам 1,2 и 3 соответствует отрицательное значение параметра FitRes. Здесь мы имеем дело со сверхдифференцирующей способностью тестового задания. Эти экспериментальные данные плохо соответствуют модели Раша. Необходимо дополнительно проверить значение параметра $\chi^2_{probability}$. Если оно менее чем 0.05, то задание следует исключить из теста.

Для кривой ICC2 (точки 4, 5 и 6) параметр FitRes=0, что свидетельствует о соответствии экспериментальных данных модели Раша.

Для кривой ICC3 (точки 7, 8 и 9) параметр FitRes > 0, что свидетельствует о плохом соответствии модели Раша. Это тестовое задание со слабой дифференцирующей способностью. Для решения вопроса об исключении задания из теста необходимо, как и в случае ICC1, проверить значение $\chi^2_{probability}$.

ChiSq[Pr] = 0,872 - мера соответствия данных модели Раша на основе проверки эмпирического и теоретического значений распределения хи-квадрат ($\chi^2_{probability}$). Если ChiSq[Pr] меньше критического значения 0.05, то задание следует исключить из теста;

Slope = 0,25 – наклон ICC в точке перегиба ($\theta_j = \beta_j$). Этот параметр характеризует теоретическую дифференцирующую способность задания – способность тестового задания различать испытуемых по уровню их знаний. В дихотомическом случае наклон всех ICC одинаков, что хорошо видно на рис. 5.8.1.

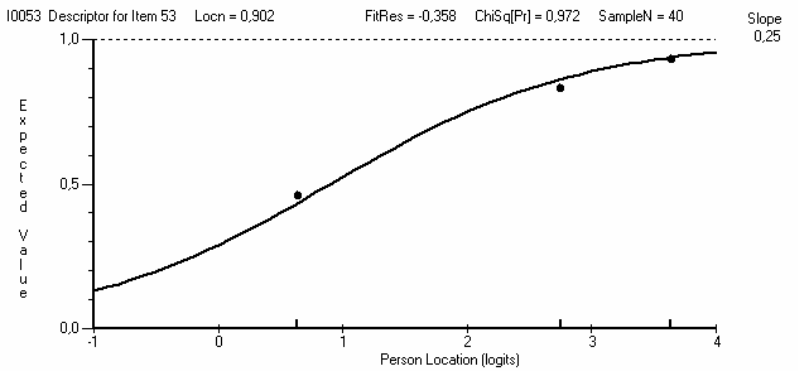


Рис. 5.8.2. ICC для задания №53 с $\chi^2_{prob} = 0.972$, входящего в состав первой группы.

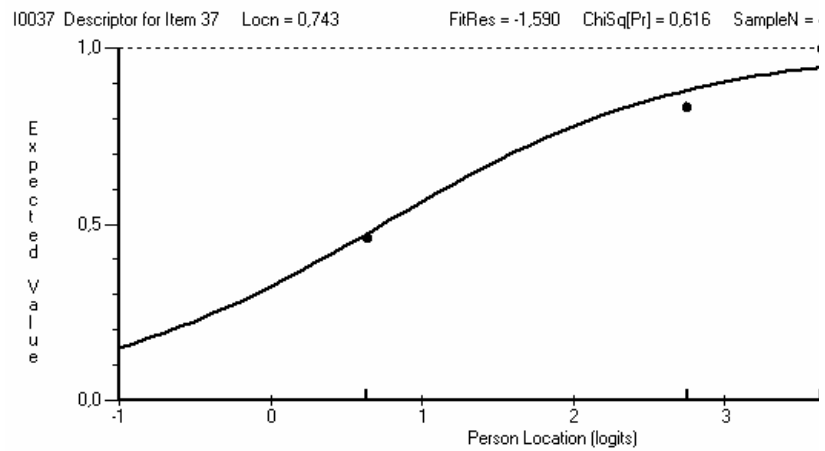


Рис. 5.8.4. ICC для задания №37, $\chi^2_{prob} = 0.616$, входящего в состав второй группы.

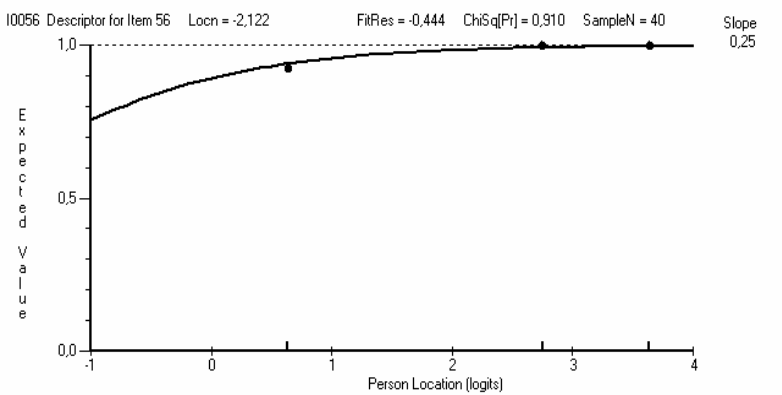


Рис. 5.8.3. ICC для задания №56 с $\chi^2_{prob} = 0.910$, входящего в состав первой группы.

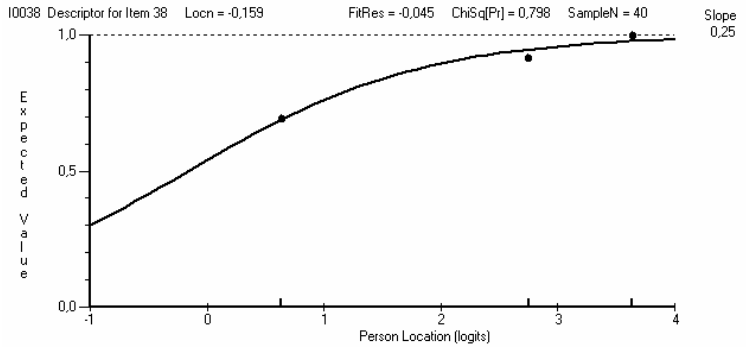


Рис. 5.8.5. ICC для задания №38 с $\chi^2_{prob} = 0.798$, входящего в состав второй группы.

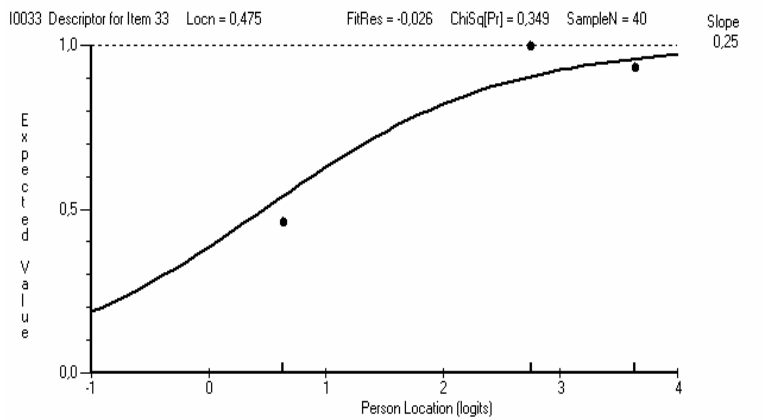


Рис. 5.8.6. ICC для задания №33 с χ^2_{prob} = 0.349, входящего в состав третьей группы.

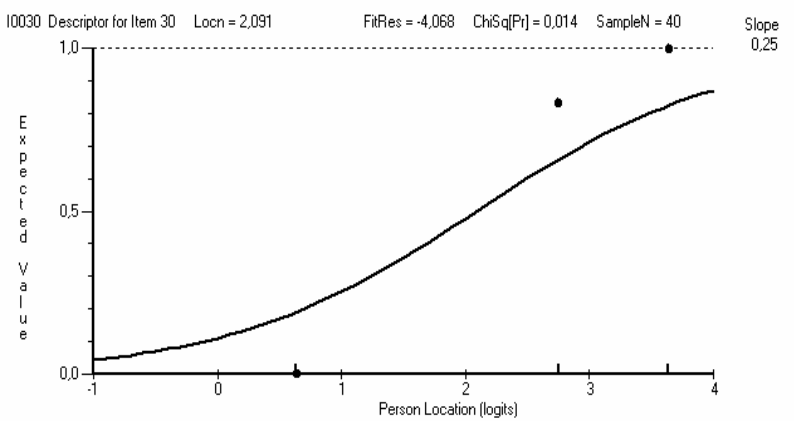


Рис. 5.8.8. ICC для задания №30 с χ^2_{prob} = 0.014, входящего в состав четвертой группы.

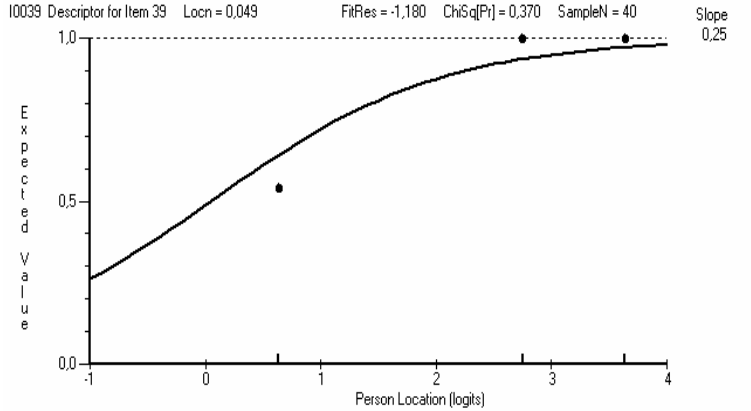


Рис. 5.8.7. ICC для задания №39 с χ^2_{prob} = 0.370, входящего в состав третьей группы.

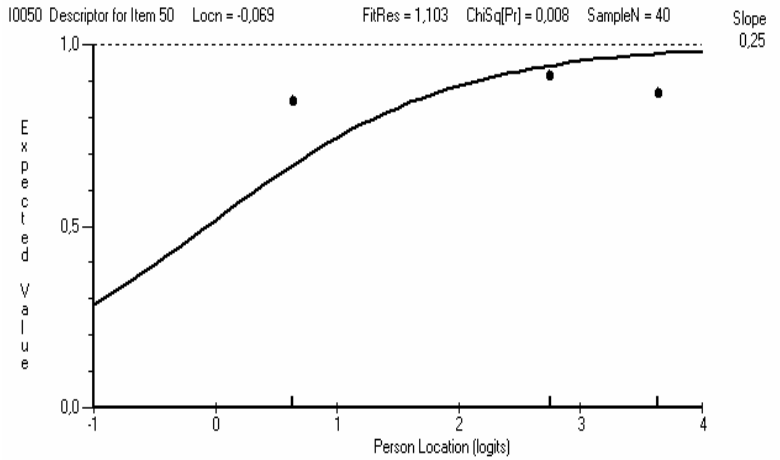


Рис. 5.8.9. ICC для задания №50 с χ^2_{prob} = 0.008, входящего в состав четвертой группы.

Как отмечалось выше, параметр $\chi^2_{probability}$ позволяет судить о степени соответствия экспериментальных данных модели Раша. По величине $\chi^2_{probability}$ все экспериментальные данные были распределены по четырем группам:

группа №1 $\chi^2_{probability} \geq 0.8$, 11 заданий;
группа №2 $0.6 \leq \chi^2_{probability} < 0.8$, 13 заданий;
группа №3 $0.05 \geq \chi^2_{probability} < 0.6$, 39 задания;
группа №4 $\chi^2_{probability} < 0.05$, 9 заданий.

В таблице 5.8.1 приведено распределение тестовых заданий по всем четырем группам. В ней данные представлены следующим образом. Допустим, нас интересует - в какую группу попадает задание №45? На пересечении столбца «40» и строки «5» находится число «1» - первая группа, следовательно, 45-е задание имеет $\chi^2_{probability} \geq 0,8$.

Таблица 5.8.1. Распределение тестовых заданий по группам $\chi^2_{probability}$.

No Задания	0	10	20	30	40	50	60	70
0	-	3	2	4	4	4	2	2
1	3	3	1	3	3	3	3	3
2	3	1	3	3	3	3	3	3
3	1	3	3	3	4	1	3	-
4	2	3	3	3	2	3	3	-
5	3	2	4	2	1	1	3	-
6	3	3	2	3	2	1	3	-
7	3	4	1	2	4	1	4	-
8	4	3	1	2	3	3	3	-
9	1	3	2	3	2	3	3	-

Каждое задание характеризуется своей мерой трудности. Этот параметр можно охарактеризовать проекцией точки перегиба логистической кривой на ось θ. Для определения трудности задания следует на графике провести горизонтальную прямую с ординатой P=0.5 до пересечения с характеристической кривой (ICC), затем опустить перпендикуляр на ось θ. Отметим, что в RUMM сразу проводится вычисление этого значения θ (Location), которое показано на графиках ICC.

В таблице 5.8.2 приведено распределение заданий по степени трудности. В таблице строка «id» обозначает номер задания, а строка

«Lcn» - обозначает значение θ, для которого вероятность правильного ответа равна P=0,5.

Из таблицы 5.8.2 видно, что задания теста с удовлетворительной равномерностью покрывают диапазон θ от -2,4 до +2,1 логитов.

Обычно считается, что тест должен перекрывать диапазон от -3 до +3 логитов. Это означает, что в анализируемом тесте не хватает очень легких и очень трудных заданий. В существующем виде тест больше предназначен для испытуемых со средними способностями.

Таблица 5.8.2. Распределение заданий по уровню их трудности.

№	1	2	3	4	5	6	7	8	9	10
id	27	12	28	56	24	67	57	45	21	11
Lcn	-2,365	-2,324	-2,122	-2,122	-1,687	-1,645	-1,414	-1,405	-1,388	-1,303
№	11	12	13	14	15	16	17	18	19	20
id	69	25	5	17	60	70	49	59	58	48
Lcn	-1,303	-1,192	-1,183	-1,176	-0,953	-0,909	-0,871	-0,815	-0,796	-0,672
№	21	22	23	24	25	26	27	28	29	30
id	18	32	20	4	65	46	36	26	72	23
Lcn	-0,589	-0,579	-0,563	-0,5	-0,494	-0,472	-0,455	-0,452	-0,379	-0,273
№	31	32	33	34	35	36	37	38	39	40
id	40	38	50	19	64	39	35	54	66	29
Lcn	-0,272	-0,159	-0,069	-0,008	0,03	0,049	0,055	0,117	0,215	0,257
№	41	42	43	44	45	46	47	48	49	50
id	14	31	41	71	33	55	47	13	16	22
Lcn	0,27	0,305	0,31	0,447	0,475	0,484	0,496	0,5	0,5	0,687
№	51	52	53	54	55	56	57	58	59	60
id	9	10	3	43	37	53	6	2	68	61
Lcn	0,693	0,728	0,731	0,742	0,743	0,902	1,016	1,025	1,103	1,104
№	61	62	63	64	65	66	67	68	69	70
id	52	42	7	62	44	1	51	15	63	34
Lcn	1,208	1,259	1,264	1,285	1,364	1,416	1,42	1,594	1,855	1,955
№	71	72	73	74	75	76	77	78	79	80
id	30	8	-	-	-	-	-	-	-	-
Lcn	2,091	2,21	-	-	-	-	-	-	-	-

Перейдем к обсуждению качества тестовых заданий на основании полученных характеристических кривых (рис. 5.8.2-5.8.9).

Из графиков видно, что экспериментальные данные для всех заданий расположены в области от 0 до 4 логитов.

Задания, входящие в первую группу (например, с ICC, показанными на рис. 5.8.2 и 5.8.3) имеют отличное согласие с моделью Раша и оставляются в тесте.

Задания, входящие во вторую группу (например, с ICC, показанными на рис. 5.8.4 и 5.8.5) имеют хорошее согласие с моделью Раша и также оставляются в тесте.

Задания, входящие в третью группу (например, с ICC, показанными на рис. 5.8.6 и 5.8.7) имеют удовлетворительное согласие с моделью Раша. Такие задания можно оставить в тесте. Отметим, что эти задания желательно дополнительно проанализировать с точки зрения их содержания. Желательно собрать дополнительную статистику на предмет выявления отклонений в процедуре тестирования.

Задания, входящие в четвертую группу (например, с ICC, показанными на рис. 5.8.8 и 5.8.9) не согласуются с моделью Раша. Такие задания следует исключить из теста.

Из рис. 5.8.6 видно, что задание №33, характеризующееся значением $\chi^2_{probability} = 0.349$, имеет удовлетворительное согласие с моделью Раша, но имеет аномальный участок - сильные испытуемые отвечают хуже, чем испытуемые со средним уровнем знаний. В.Аванесов связывает подобные эффекты с нарушениями формальных, организационных и этических требований. В связи с тем, что аномальный эффект проявляется лишь частично, а $\chi^2_{probability} > 0.05$, то это задание можно временно оставить в изучаемом наборе заданий, имея в виду дальнейшую проверку теста в целом.

На рис.5.8.8 приведена логистическая кривая для задания №30 с $\chi^2_{probability} = 0.014$. Это задание имеет сверхвысокую дифференцирующую способность, то есть имеет малый диапазон перекрытия по уровню знаний испытуемых. Экспериментальные данные показывают, что слабые испытуемые практически не могут дать верный ответ на это задание. С другой стороны, средние и сильные испытуемые на это задание отвечают гораздо лучше, чем того требует модель Раша. Как указывалось выше, ввиду несоответствия модели Раша, подобные задания исключаются из теста.

Пример логистической кривой для задания №50 ($\chi^2_{probability} = 0.008$) с практически полным отсутствием дифференцирующей способности приведен на рис. 5.8.9. Это задание

почти не различает слабых, средних и сильных испытуемых.

Это довольно легкое задание (Location = -0,069), но сильные испытуемые показывают такую же вероятность успеха, как средние и слабые, что противоречит модели Раша. Кроме того, это задание плохо соответствует другим заданиям и по всем этим причинам должно быть удалено из теста.

Таким образом, анализ результатов тестирования на основе подхода Rasch measurement позволяет оптимизировать содержание теста и превращать его в инструмент для измерения уровня знаний испытуемых. Особенно удобно это делать с применением программного средства RUMM

Литература к главе 5

[1] Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen, 1960, Danish Institute of Educational Research. (Expanded edition, Chicago, 1980, The University of Chicago Press).

[2] Нейман Ю.М., Хлебников В.А. Введение в теорию моделирования и параметризации педагогических тестов. -М.: Прометей, -169 с.

[3] Аванесов В.С. Применение тестовых форм в Rasch Measurement // Педагогические измерения, 2005, №4. -С.3-20.

[4] Маслак А.А. Измерение латентных переменных в социально-экономических системах: Монография. -Славянск-на-Кубани: Изд.центр СГПИ, 2006, -333 с.

[5] Wilson M. Constructing Measures: An Item Response Modeling Approach. - Mahwah, New Jersey: Lawrence Erlbaum associates, 2005. -228 p.

[6] Челышкова М.Б. Теория и практика конструирования педагогических тестов: Учебное пособие. – М.: Логос, 2002. -432 с.

[7] Михеев О.В. Математические модели педагогических измерений // Педагогические измерения, 2004, №2. -С.75-88.

[8] Smith E.V., Conrad J.M., Chang K., Piazza J. An introduction to Rasch measurement for scale development and person assessment //Journal of

Nursing Measurement, 2002, 10, 189-206. (Перевод Н.Пракиной //Педагогические измерения, 2006, №1. -С.65-81.

[9] Birnbaum A. Some Latent Trait Models and Their Use in Inferring and Examinee's Ability. In Lord F.M., Novick M. Statistical Theories of Mental Test Scores. Addison-Wesley Publ. Co. Reading, Mass, 1968. -P.397-479.

[10] Беспалько В.П. Программированное обучение. Дидактические основы. –М., 1970. -300 с.

[11] Stenner A.J., Wright B.D. & Linacre J.M. From P-values and RAW Score Statistics to Logits //Rasch Measurement Transactions, 1994, RMT, 8:1. -p.338.

[12] Колпаков А.В., Колпакова А.В., Захаров А.А. Численный метод получения логитов из первичного балла // Вопросы тестирования в образовании, 2002, №3. -С.125-128.

[13] Овчинников В.В. Оценивание учебных достижений учащихся при проведении централизованного тестирования. -М.: Центр тестирования МО РФ, 2001. -27 с.

[14] Ким В.С. Анализ результатов тестирования в процессе Rasch measurement //Педагогические измерения, N4, 2005. –С.39-45.

[15] Ким В.С. Измерение латентных параметров испытуемых и тестовых заданий. - Мат. IX Всерос. научно-практ. конф. «Теория и практика измерения латентных переменных в образовании» (21-23 июня 2007 г.). -Славянск-на-Кубани: Изд.центр СГПИ, 2007. -С.70-71.

[16] Hambleton R.K. Application of Item Response Theory . -Vancouver: Educ.Res.Inst. B.C. , 1983.

[17] Аванесов В.С. Знания как предмет педагогического измерения // Педагогические измерения. №3, 2005г. Там же: Аванесов В.С. Применение тестовых форм в Rasch Measurement /Сб. тр. Научно - метод конф. Славянского-на-Кубани госпединститута. 2005

[18] Baker F.B. The Basics of Item Response Theory. –ERIC, 2001. -172 p

[19] Аванесов В.С. Основы научной организации педагогического контроля в высшей школе. М., 1989. - 167 с.

[20] Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen, 1960, Danish Institute of Educational Research. (Expanded edition, Chicago, 1980, The University of Chicago Press).

[21] Ким В.С. Анализ результатов тестирования в процессе Rasch measurement //Педагогические измерения, N4, 2005. –С.39-45.

[22] Ким В.С. Измерение латентных параметров испытуемых и тестовых заданий. - Мат. IX Всерос. научно-практ. конф. «Теория и практика измерения латентных переменных в образовании» (21-23 июня 2007 г.). - Славянск-на-Кубани: Изд.центр СГПИ, 2007. -С.70-71.

[23] http://www.rummlab.com/, Andrich, D., Sheridan, B., Lyne, A. & Luo, G. (2000) RUMM: A windows-based item analysis program employing Rasch unidimensional measurement models (Perth: Murdoch University).

СОДЕРЖАНИЕ

Монография

Владимир Сергеевич Ким
Тестирование учебных достижений

Редактор С.С. Кушнир
Корректор В.В. Елизенцева
Верстка С.С.Кушнир

Подписано в печать 20.12.07. Формат 60×90/16
Бумага офсетная. Печ.л.13.
Тираж 200 экз. Заказ 354.

Издательство УГПИ, 692508, г. Уссурийск, ул. Тимирязева, 33

Отпечатано участком оперативной полиграфии
Уссурийского государственного педагогического института
692508, г. Уссурийск, ул. Некрасова, 25

Ким Владимир Сергеевич,
кандидат физико-математических наук, доцент, начальник управления по информатизации Уссурийского государственного педагогического института, директор межвузовского центра контроля качества знаний.
Профессиональные интересы сосредоточены в области тестирования учебных достижений, компьютерного моделирования в учебном процессе.
Имеет более 100 опубликованных работ.

Монография посвящена теоретическим и практическим проблемам тестирования учебных достижений. Содержит краткий обзор развития тестирования в России и за рубежом. Рассмотрены понятия надежности и валидности теста, структуры и формы тестового задания. Обсуждаются вопросы обработки результатов тестирования, как на основе классической теории тестов, так и с применением Item Response Theory к анализу качества тестовых заданий.

Монография предназначается преподавателям, учителям, аспирантам и всем, кто интересуется тестированием учебных достижений.