Министерство общего и профессионального образования
Российской Федерации
Научно-информационный центр государственной аккредитации

Препринт № 4/97

В. Г. Наводнов

Математические модели САПР ПИМ

1997

Министерство общего и профессионального образования
Российской Федерации
Научно-информационный центр государственной аккредитации

Препринт № 4/97

В. Г. Наводнов

Математические модели САПР ПИМ

1997

УДК 001.891.573:658.512.011.56:378

Наводнов В.Г. Математические модели САПР ПИМ: Препринт № 4/97. — Йошкар-Ола: Научно-информационный центр государственной аккредитации, 1997. — 72 с.

Рассматриваются вопросы формализованного описания процедуры проектирования педагогических измерительных материалов.

Для специалистов-тестологов, занимающихся вопросами оценки уровня обученности.

Рецензенты: канд. физ.-мат наук, доцент Масленников А.С.,
канд. физ.-мат наук Петропавловский М.В.

Содержание

Введение

Важнейшей составляющей процедуры аккредитации образовательных организаций является установление соответствия содержания, уровня и качества подготовки выпускников требованиям ГОС (государственного образовательного стандарта). Один из возможных и широко апробированных подходов — тестирование учебных достижений. Тестовая диагностика в настоящее время широко используется на Западе. В США, например, существует несколько крупных (ETS, ACP, ...) и десятки мелких фирм, основная деятельность которых — определение уровня обученности учащихся. В течение десятилетий отрабатывались и внедрялись технологии проведения, обработки и анализа результатов тестирования. Разработано большое количество математических моделей, лежащих в основе этих технологий.

В последние годы и в нашей стране возрос интерес к внедрению тестовых технологий в практику работы образовательных организаций. Появились научные коллективы, ведущие исследования по построению математических моделей технологий тестирования, разработке методик создания тестов, разработке программного обеспечения для сопровождения тестовых мероприятий — В.С. Аванесов [1]–[2], В.П. Беспалько [4], А.Н. Майоров [11], В.Г. Наводнов, В.Ж. Куклин, А.С. Масленников, Б.А.Савельев, А.В. Ельцын, М.В. Петропавловский [7]–[10], [12]–[13], Б.У. Родионов, А.О. Татур [14], М.Б. Челышкова [15], [16] и др. При этом необходимо отметить принципиальное отличие российских педагогических измерительных процедур, а именно:

- использование заданий открытого типа;
- многовариантность;
- отсутствие инфраструктуры, позволяющей проводить тестирование в один и тот же день и час по всей стране.

Объективное оценивание уровня обученности студентов требует использование большого и разнообразного количества педагогических измерительных материалов (тестов, контрольных заданий, комплексных контрольных заданий, опросников и т. п.). Учитывая огромное количество образовательных программ, реализуемых в образовательных организациях, и необходимость проверки соответствия уровня обученности обучаемых по каждой образовательной программе требованиям ГОС, возникает задача создания системы автоматизированного проектирования педагогических измерительных материалов (САПР ПИМ), не привязанной к какой-либо конкретной предметной области.

Данная работа посвящена формализованному описанию баз тестовых заданий, методам калибровки заданий, математическим моделям и алгоритмам процедур проектирования педагогических измерительных материалов.

Автор выражает глубокую признательность сотрудникам Научно-информационного центра государственной аккредитации за полезное обсуждение работы.

1. Общая схема процедур педагогических измерений

Общую схему педагогических измерений (тестирования) можно представить в виде так называемого цикла Шухарта-Деминга (или APDC-цикла) (рис. 1.1):

1) постановка целей педагогических измерений;

2) разработка педагогических измерительных материалов;

3) педагогические измерения;

4) обработка, анализ, интеграция и мониторинг полученных результатов.

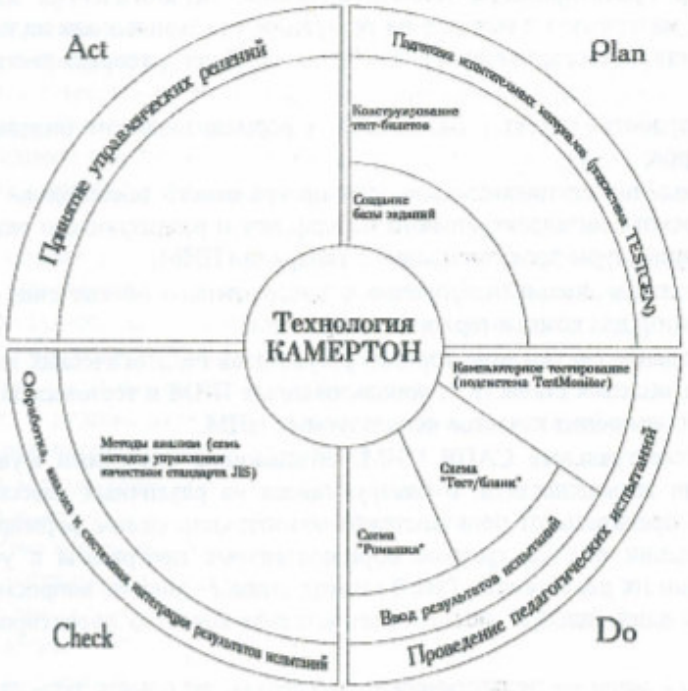


Рис. 1.1. APDC-цикл Шухарта-Деминга

Опишем кратко каждый из этапов (более полное описание можно найти в работах [7] и [8]).

Первый этап — постановка целей педагогических измерений. В зависимости от целей (текущая проверка знаний, итоговая проверка, вступительные экзамены в образовательную организацию, предметная олимпиада и т.п.) выбирается та или иная модель, ставится задача на получение той или иной информации.

Второй этап — разработка педагогических измерительных материалов. Основная идея состоит в проектировании ПИМ из банков тестовых заданий, подготовленных и откалиброванных заранее. Один из возможных подходов в решении этой задачи — создание системы автоматизированного проектирования педагогических измерительных материалов (САПР ПИМ), не привязанной к какой-либо конкретной предметной области.

Создание такой САПР ПИМ требует:

1) формализованного (математического, алгоритмического) описания процедур проектирования (конструирования) педагогических измерительных материалов и методов их генерации, основанных как на классической, так и на современной (Item Responce Theory) теориях тестирования;

2) разработки структур баз заданий и формализованного описания их параметров;

3) создания специализированного программного обеспечения ТестГен на основе интеллектуального интерфейса и реализующего разнообразные процедуры проектирования и генерации ПИМ;

4) создания специализированного программного обеспечения ТестЭкзаменатор для компьютерного тестирования;

5) создания систем мониторинга результатов педагогических измерений и накопления статистики использованных ПИМ и тестовых заданий с целью повышения качества используемых ПИМ.

Интеллектуальные САПР ПИМ, обладающие широкими функциональными возможностями и базирующиеся на различных классах баз заданий, представляют пользователям возможность самим формировать базы заданий под конкретные образовательные программы с учетом специфики их реализации. Такой подход снимает многие вопросы, связанные с адаптацией,и обеспечивает высокое качество проектирования ПИМ.

Третий этап — педагогические измерения, которые с технологической точки зрения, можно проводить по одной из следующих схем (рис. 1.2):

Рис. 1.2. Технологические схемы педагогических измерений

1) *закрытая схема.* В этом случае тест-билет состоит из заданий, которые не требуют проверки со стороны преподавателей (экспертов), и результаты тестирования непосредственно вводятся в компьютер для обработки. Типичный пример заданий закрытого типа — задание с выбором одного правильного ответа из нескольких представленных. Большая часть широко известных западных тестов являются тестами такого типа;

2) *открытая схема.* В отличие от западной системы образования, российские педагоги отдают предпочтение проверке знаний учащихся при помощи открытых заданий. В этом случае, прежде чем ввести результаты в компьютер, требуется оценивание решения каждого задания педагогом (экспертом). Важно избежать субъективного влияния проверяющих. В течение ряда лет хорошо себя зарекомендовала система горизонтальной проверки (метод ромашки).

Суть метода состоит в том, что при разработке тест-билета создается карта проверки, в которой указывается система оценивания, а также количество и перечень учебных элементов, знание которых необходимо продемонстрировать, чтобы выполнить данное задание. В качестве примера на рис. 1.3 приведена схема оценивания решения логарифмического уравнения.

Для проведения оценивания:

- несколько проверяющих объединяются в группу;

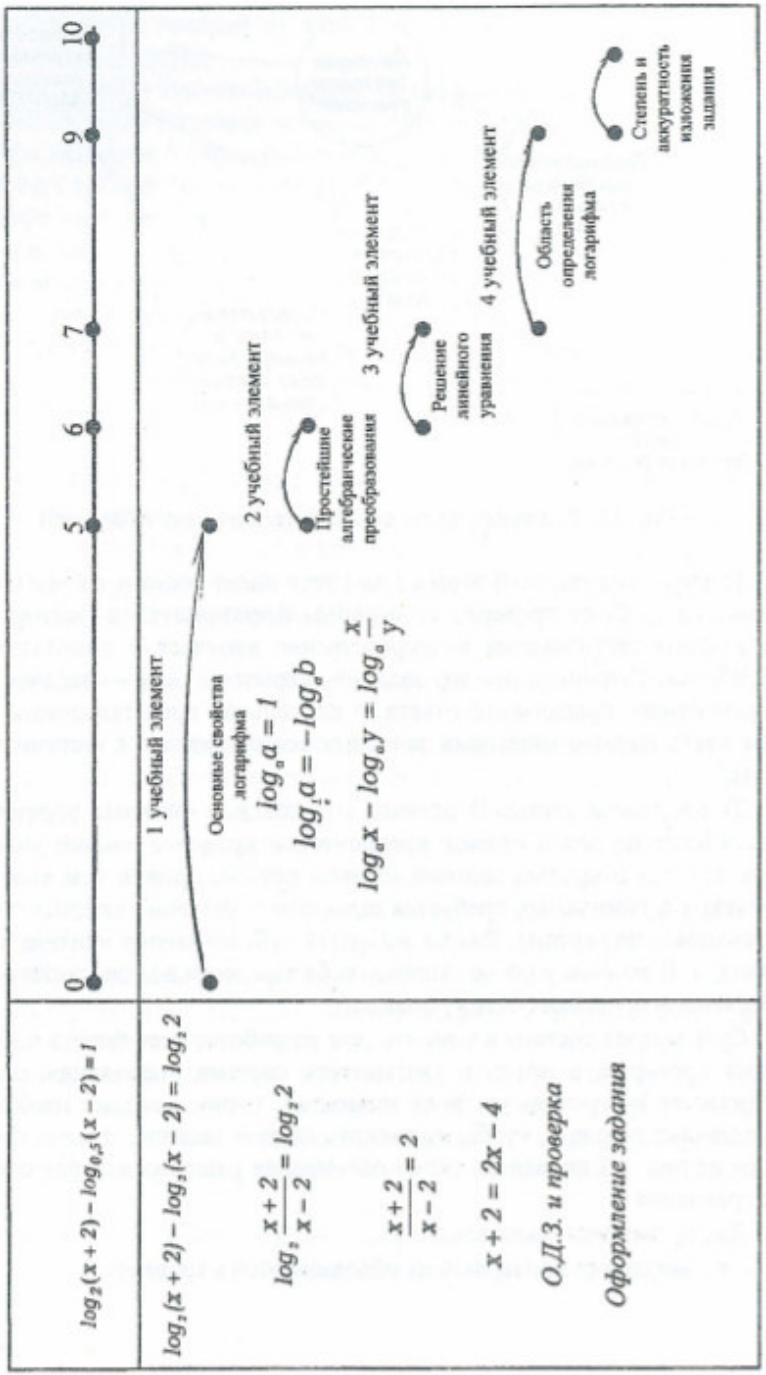


Рис. 1.3. Метод экспертного оценивания заданий открытого типа

- используется карта проверки с подробным описанием количества выставляемых баллов за выполненные этапы решения; как правило, используется многобалльная (например, 10-балльная) шкала;
- один педагог проверяет во всех работах задания, идентичные по своим параметрам. Хотя оценку проводят разные педагоги, но все тестируемые оказываются в равных условиях.

Результаты проверки заносятся в контрольный листок (рис. 1.4) и в компьютер для обработки.

Длительный опыт эксплуатации этого метода показал, что погрешность составляет 3–5%, чего вполне достаточно для получения надежных результатов при массовом тестировании.

Задание	*Перечень учебных элементов*	*Оценка учебных элементов*	*Оценка задания*	*Проверил*
1	*1.1*			
	1.2			
2	*2.1*			
	2.2			
	2.3			
	2.4			
3	*3.1*			
	3.2			
	3.3			
...				

Рис. 1.4. Контрольный листок

3) *компьютерное тестирование*. При наличии в образовательной организации определенного количества вычислительной техники и достаточной компьютерной подготовки обучаемых весьма эффективным является компьютерное тестирование, которое обладает рядом важных преимуществ:

- оперативностью проведения и обработки результатов, высокой технологичностью;
- использованием мультимедиа-возможностей представления тестовой информации, открывающим принципиально новые возможности в педагогическом измерении;

- игровой мотивацией, позволяющей повышать достоверность результатов и привлекательность процесса тестирования;
- возможностью реализации адаптивных алгоритмов тестирования.

Для реализации компьютерного тестирования в Научно-информационном центре государственной аккредитации разработаны оболочки ТестЭкзаменатор (для DOS) и ТестЭкзаменаторWin (для операционной системы Windows'95), цель которых — представление тестируемому тест-билетов, сгенерированных системой ТестГен [12], и передача результатов тестирования в систему обработки.

Четвертый этап. Обработка, анализ, интеграция и мониторинг полученных результатов. Это заключительный этап цикла, который должен обеспечить:

- решение задач, поставленных при определении целей тестирования;
- интеграцию информации для передачи в систему мониторинга результатов тестирования;
- уточнение (определение) статистических характеристик использованных заданий и передачу этой информации в базу тестовых заданий.

Для обработки, анализа и интеграции результатов тестирования разработан программный комплекс КАМЕРТОН (см. [7]).

2. Математические модели проектирования измерительных материалов

2.1. Основные понятия теории тестирования. Графические методы представления

Появление методов измерения учебных достижений обучаемых (в современном понятии этого слова) связано с работами американских исследователей в конце XIX века. Хотя предыстория этого явления весьма обширна и уходит корнями в глубокую древность (см., например, замечательные исследования по этому вопросу в [17]). В начале XX века были заложены основы теории тестирования, которые активно развивались до начала 70-х годов. Этот период развития теории принято называть классическим, а разработанную теорию — классической теорией тестирования. В 1968 г. Ф. Лорд и М. Новик [32, 33, 35] сформулировали основные постулаты математической модели классической теории тестирования.

Под тестом T в классической теории тестирования понимается структурированная система заданий и соответствующая ей процедура проверки этих заданий, обеспечивающая однозначность интерпретации полученных результатов тестирования.

В связи с возрастающим использованием современной компьютерной техники при определении уровня обученности студентов и ее широким внедрением в практику работы образовательных организаций возникает задача переосмысления методов и средств классической теории тестирования, формализации процедур и методов, создания технологии тестирования, рассчитанной на массового пользователя.

При этом важным аспектом использования основных понятий классической теории тестирования, с целью измерения уровня обученности испытуемых, является визуализация результатов измерения. Как показал многолетний опыт эксплуатации технологии КАМЕРТОН [7], это формы «семи методов управления качеством» японского промышленного стандарта JIS: гистограмма, наложение (стратификация) гистограмм, диаграмма Парето, контрольная карта Шухарта и др.

Пусть тест-билет T, составленный из L заданий, был представлен для тестирования группы из N испытуемых. Результаты тестового испытания удобно представить в виде таблицы:

Тестируемые (испытуемые)	Задания				
	1	*2*		*l*	*L*
1				:	
2				:	
				:	
				:	
n	..	..	..	..	u_{nl}
N					

$$\sum_{l=1}^{L} u_{nl} = r_n$$

$$\sum_{n=1}^{N} u_{nl} = s_l,$$

где u_{nl} — оценка выполнения l-го задания n-м испытуемым.

Для дихотомических заданий, то есть заданий, оцениваемых в бинарной шкале (верно/неверно):

$$u_{nl} = \begin{cases} 1, & \text{если } l\text{-е задание выполнено } n\text{-м испытуемым верно;} \\ 0, & \text{в противном случае.} \end{cases}$$

Для политомических заданий, то есть заданий, оцениваемых в многобалльной шкале ($[0,m]$): $0 \le u_{nl} \le m$.

Матрицу $U = \left[u_{n,l}\right]_{1,1}^{N,L}$ размером $N \times L$ будем называть *матрицей ответов*, сумму $\sum_{l=1}^{L} u_{nl} = r_n$ — *первичным результатом тестирования n-го испытуемого*, $\frac{r_n}{N}$ — *коэффициентом*, а $\frac{r_n}{N} \cdot 100\%$ — *процентом выполнения тест-билета.*

Для анализа и принятия решений по итогам педагогических измерений результаты удобно представить в наиболее наглядной форме. Простыми и удобными формами являются рейтинг-лист и гистограмма.

Под рейтинг-листом будем понимать список испытуемых, упорядоченный в порядке убывания полученных ими баллов (результатов тестирования). Такой упорядоченный список можно представить как для отдельных групп, так и для целых потоков.

На рейтинг листе можно отметить границу-минимум процента выполнения данного тест-билета. Это позволяет сразу определить как число испытуемых, не выполнивших тест-билет, так и их фамилии, и группы. Четко представлены лидеры тестирования и виден максимальный процент выполнения тест-билета.

№	Фамилия, имя, отчество	Группа	Итого баллов	Процент выполнения
1	КОРОСТЕЛЕВА О.	БУ-12	160,0	100,0
2	СНЫТКО Е.А.	БУ-12	160,0	100,0
3	БЕСПАЛОВА Е.В.	БУ-13	158,0	98,8
4	БОБКОВА Т.Н.	БУ-13	157,0	98,1
5	ПИРОГОВА И.А.	БУ-13	150,0	93,8
6	ЧИРКОВА Ю.	БУ-13	150,0	93,8
7	ЕФИМОВА Т.Н.	БУ-14	118,0	92,5
...	...	...	...	...
76	ТОЛМАЧЕВА	БУ-14	86,0	53,8
77	ЖУРБИНА Е.	БУ-14	82,0	51,3
78	КОНЬКОВА Ю.В.	БУ-11	80,0	50,0
79	ИВАНОВА О.С.	БУ-12	75,0	47,9
80	КУРЛОВ И.	БУ-12	70,0	43,8

Рис. 2.1. Пример рейтинг-листа по результатам тестирования

Если педагогическое измерение проводится на большом массиве испытуемых, то рейтинг-лист становится очень длинным, поэтому удобно для наглядного представления результатов в целом использовать другую форму представления — гистограмму.

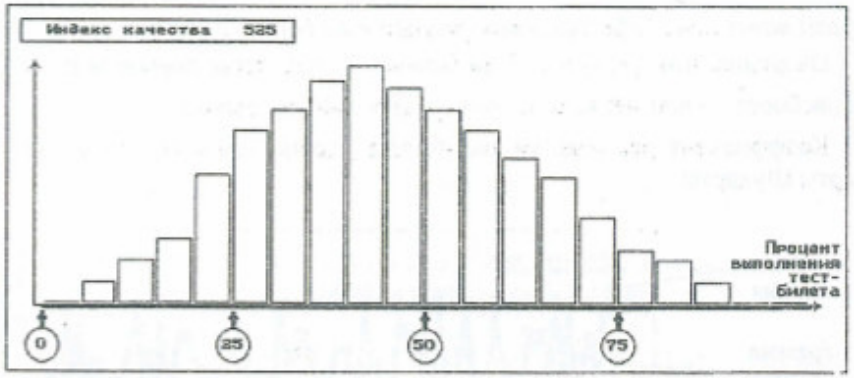


Рис. 2.2. Гистограмма распределения результатов тестирования по проценту выполнения тест-билета

По оси абсцисс откладывается процент выполнения тест-билета, а высота столбцов соответствует доле испытуемых, имеющих результат в заданном процентном интервале. Практика показывает, что в качестве шага удобно выбирать интервал в 5 или 10%.

Таким образом, гистограмма иллюстрирует плотность распределения результатов педагогических измерений и позволяет показать соотношение размеров различных групп испытуемых, получивших низкие, средние или высокие баллы.

Результаты педагогических измерений представляют интерес не только с точки зрения анализа обученности испытуемых, но и с точки зрения качества разработки тест-билетов.

Часто перед тем, как перейти к анализу данных по результатам тестирования, проводят *выбраковку* — удаляют строки и столбцы, состоящие полностью из 0 или 1 (или M для политомических заданий), то есть удаляют задания, которые никто не смог выполнить или, наоборот, все выполнили. Аналогично с испытуемыми — не представляет интереса анализ результатов тестируемых, не справившихся ни с одним заданием (уровень тест-билета оказался выше его уровня подготовки) или, наоборот — абсолютно выполнивших все задания (уровень тест-билета оказался ниже его уровня подготовки).

Отношение

$$k_l = \frac{\frac{1}{m}\sum_{n=1}^{N} u_{nl}}{N}$$

будем *называть коэффициентом решаемости* l-го задания.

Очевидно, что $0 \le k_i \le 1$. Чем больше k_i, тем легче данное задание, и, наоборот — чем меньше k_i, тем труднее данное задание.

Коэффициент решаемости тест-билета удобно представлять в виде карты Шухарта:

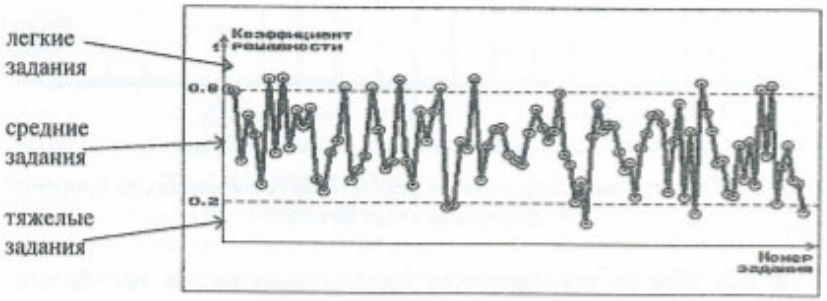


Рис. 2.3. Карта Шухарта коэффициентов решаемости тест-билета

где по оси абсцисс откладываются номера заданий в тест-билете, а по оси ординат — коэффициент решаемости.

Коэффициент селективности задания (другие названия — коэффициент чувствительности, дискриминационный индекс, *D*-индекс) используется как показатель дифференциации обучаемых. В классической теории тестирования разработаны десятки таких показателей (среди многих — бисериальный коэффициент r_{bis}, точечный бисериальный коэффициент r_{pb}, тетрахорический коэффициент r_{tet}, коэффициент φ, *ULI* (upper-lower-index)). Однако на практике эти показатели примерно одинаково эффективны.

Наиболее простым является коэффициент селективности, определяемый как upper-lower-index:

$$D_i = k_i^{'} - k_i^{''},$$

где $k_i^{'}$ — коэффициент решаемости i-го задания лучшей половины тестируемых; $k_i^{''}$ — коэффициент решаемости i-го задания худшей половины тестируемых.

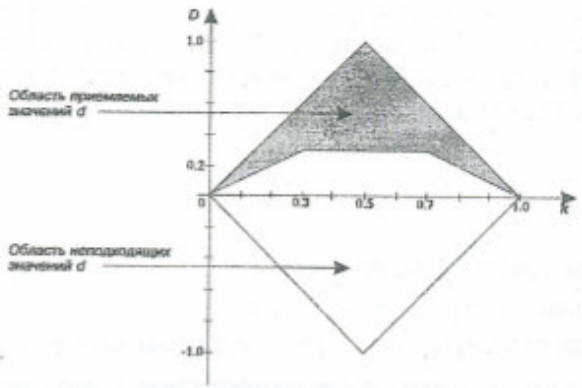


Рис. 2.4. Область зависимости между коэффициентом решаемости и коэффициентом селективности задания

Очевидно, что $-1 \le D_i \le 1$. Если задание правильно выполняет больше лучших, чем худших тестируемых, то $D>0$, в противном случае $D<0$. Если задание выполнит одинаковое количество лучших и худших, то $D=0$. Задание не дифференцирует тестируемых.

Обычно считается, что для заданий с коэффициентом решаемости $k \in [0,3;\ 0,7]$ коэффициент селективности должен быть не менее 0,25. Для заданий с $k \in [0,2;\ 0,3) \cup (0,7;\ 0,8]$ — D должен быть не менее 0,15.

Замечание: T. Kelley [29] показал, что оптимальный уровень селективности достигается, когда популяции требуемых делятся на лучших и худших не в соотношении 50%:50%, а в соотношении 27%:73%.

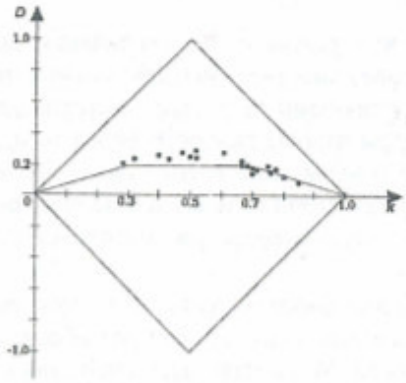


Рис. 2.5. Фазовая плоскость (D,k) тест-билета по элементарной математике–96 (Единый экзамен Республики Марий Эл в 1996 г.)

Точечно-бисериальный коэффициент — часто используемый коэффициент селективности, представляющий собой упрощенную формулу коэффициента корреляции Пирсона между количеством тестируемых, выполнявших данные дихотомические задания, и общим результатом, а именно:

$$r_{ipb} = \frac{\mu_i - \mu_x}{\sigma_x}\sqrt{\frac{k_i}{1-k_i}},$$

где μ_i — среднее значение результатов тестирования среди тестируемых, ответивших корректно на i-е задание,

μ_k — среднее значение результатов тестирования всех тестируемых,

σ_x — среднее квадратичное отклонение результатов тестирования всех тестируемых,

k_i — коэффициент решаемости i-го задания.

2.2. Латентное пространство. Характеристические кривые

Цель любого тестирования (в том числе и педагогического) — оценка определенных характеристик личности, которые явно не наблюдаемы и поэтому не поддаются прямому измерению (принято называть такие характеристики скрытыми или латентными). Для оценки латентных характеристик используются косвенные методы, в частности анализ ответа на поставленные вопросы, или анализ реакций на определенные стимулы. Вообще говоря, латентная переменная не есть какая-либо врожденная черта. Она может и должна меняться во времени вместе с личностью, например: способность читать, складывать числа, умение вычислять интегралы и т. п.

Важно отметить, что результаты тестирования зависят от многих факторов: условий проведения тестирования, мотивации испытуемых, их опыта работы с аналогичными тестовыми материалами и т.п. В этой работе мы не учитываем данные факторы. Более того, считаем, что при проведении повторного испытания тестируемый покажет те же самые результаты (значение латентной переменной не изменится). Это предположение существенно при анализе рассматриваемых математических моделей тестирования.

Таким образом, будем рассматривать латентную переменную θ как абстрактное математическое понятие, которое обозначает исследуемую характеристику личности. Множество всех возможных значений латентной переменной представим в виде одномерного или многомерного пространства, которое называется *латентным пространством* Ω.

Модели, в которых пространство Ω является одномерным, называются *гомогенными*.

Модели, в которых при анализе результатов тестирования рассматривается многомерное пространство Ω, называется *гетерогенным*.

Один из классических примеров латентного пространства — модель Бине-Симона [3], где в качестве латентной переменной рассматривался так называемый «умственный возраст».

В качестве другого примера укажем на подход Б.У. Родионова, А.О. Татура [14], которые представляют структуру латентного пространства в виде плоскости. По одной из осей откладывается объем учебной информации (знания), которой владеет тестируемый, а по другой оси — степень владения этой информацией (умения).

Примером использования многомерного латентного пространства (более 10 независимых латентных переменных) является широко известный психологический опросник MMPI (см. [3]).

Выбор структуры латентного пространства зависит от структуры информации, которую желает получить тестолог в результате тестирования. Построение латентного пространства — непростая задача, тесно соприкасающаяся с психологией и теорией познания.

Всюду далее, если не оговорено противное, будем предполагать, что латентное пространство одномерно.

Пусть Ω — латентное пространство и $\theta \in \Omega$. Для каждого дихотомического задания (задание называется дихотомическим, если ответ оценивается в бинарной шкале — верно/неверно) через $P=P(\theta)$ обозначим вероятность правильного ответа испытуемого, для которого θ есть истинное значение латентной переменной. Для политомических заданий (т. е. заданий, оцениваемых в многобалльной шкале $[0, m]$) $P(\theta)$ отождествим с оценкой в отнормированной шкале оценивания $[0,1]$.

Монотонная неубывающая функция $\pi_i : \Omega \to [0,1]$, описывающая вероятность выполнения задания тестируемым с различным уровнем латентной переменной $\theta \in \Omega$ $\pi_i(\theta_j) = P(\theta_j) = P(u_{ij} = 1 : \theta_j)$, называется характеристической функцией i-го задания. Примеры характеристических функций заданий приводятся на рис. 2.6.

Характеристическая функция задания является монотонной, неубывающей и $\lim\limits_{\theta \to +\infty} \pi(\theta) = 1$, $\lim\limits_{\theta \to -\infty} \pi(\theta) = 0$.

Испытуемый с большей латентной переменной θ_2 имеет большую вероятность ответа, чем испытуемый с латентной переменной θ_1 $(\theta_1 < \theta_2)$.

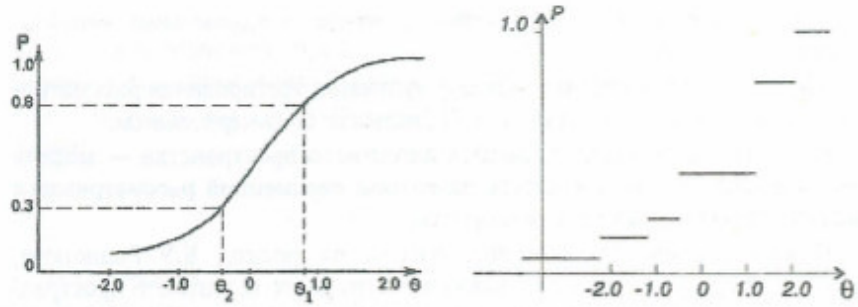


Рис. 2.6. Примеры характеристических кривых заданий

Введение характеристических кривых является краеугольным камнем Item Response Theory. Фактически первыми, кто обратил внимание на наличие зависимости вероятности правильного ответа от степени развития ребенка (как латентной переменной) были Binet, Simon [19], которые изучали зависимость между возрастом ребенка и его способностью выполнять различные задания (рис. 2.7).

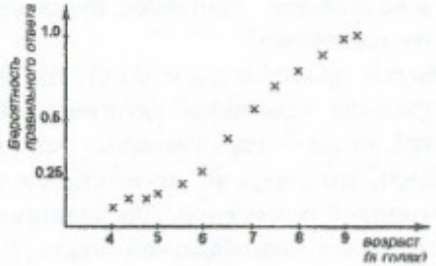


Рис. 2.7. Зависимость вероятности правильного ответа от «умственного возраста» ребенка

Основная идея введения характеристических функций заданий состоит в том, что вероятность правильного ответа на задание и ошибка измерения связаны с латентной переменной θ функциональной зависимостью. Это принципиальное отличие от классической теории тестирования, в которой коэффициент решаемости задания не зависит от латентной переменной тестируемого, и является константой. Можно сказать, что коэффициент решаемости — это усредненная характеристика задания.

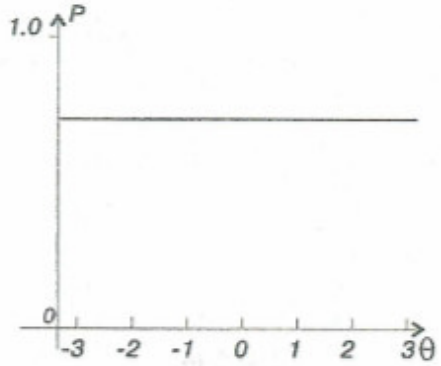


Рис. 2.8. Характеристические кривые в классической теории тестирования (константы)

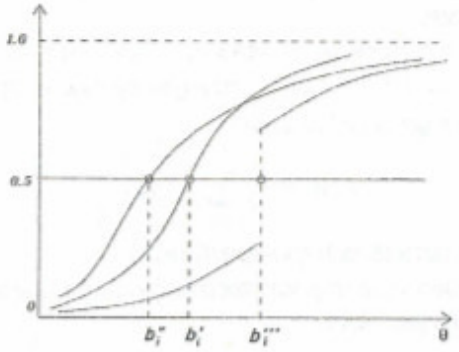


Рис. 2.9

Значение переменной θ, в котором функция $\pi_i(\cdot)$ равна *0,5*, называется *трудностью задания* и обозначается b_i : $\pi_i(b_i) = 0{,}5$. Если такого значения переменной не существует (характеристическая функция терпит разрыв), то трудностью задания b_i называется точка, в которой $\lim\limits_{\theta \to b_i - 0} \pi_i(\theta) < 0{,}5$, но $\lim\limits_{\theta \to b_i + 0} \pi_i(\theta) > 0{,}5$.

Множество заданий называется равномерно трудным, если для каждой пары заданий из этого множества их характеристические функции π_i и π_k удовлетворяют условию:

$$\pi_i(\theta) \le \pi_k(\theta), \ \forall \theta \in \Omega.$$

При этом задание с характеристической функцией $\pi_k(\cdot)$ называется более трудным, чем задание с характеристической функцией $\pi_i(\cdot)$.

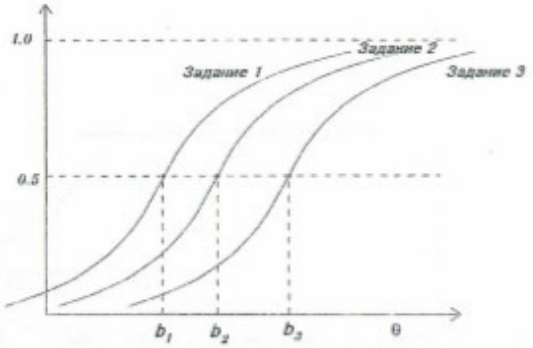


Рис. 2.10

Класс характеристических функций в модели Раша (см. п. 2.3) является примером множества характеристических функций с равномерно трудными заданиями.

Пусть тест T составлен из N заданий с характеристическими функциями $\pi_i(\cdot)$ $(i=1, ..., N)$. Функцию, равную среднему арифметическому характеристических функций заданий

$$\pi(\theta) = \frac{1}{N}\sum_{i=1}^{N}\pi_i(\theta),$$

называют характеристической функцией теста T.

Типичный пример характеристической функции (характеристической кривой) приведен на рис. 2.11.

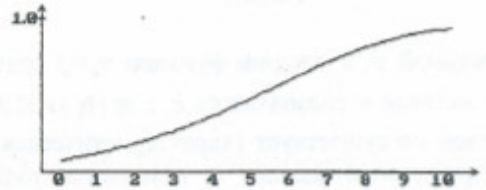


Рис. 2.11. Пример характеристической функции тест-билета Единого экзамена по математике в 1995 г. в Республике Марий Эл (число тестируемых —1011)

Характеристическая функция π осуществляет преобразование латентного пространства Ω в шкалу результатов тестирования. Это преобразование, вообще говоря, не является линейным. Например, если все задания, составляющие тест, имеют одну и ту же характеристическую функцию, то характеристическая функция теста совпадает с характеристической функцией заданий. Если тест составлен из двух подмножеств заданий — легких и тяжелых, то на характеристической кривой будут ярко выражены три участка с различной степенью кривизны. Такой тест будет хорошо дифференцировать «слабых» и «сильных» учащихся, но плохо дифференцировать «средних». Кривая будет пологой в середине интервала изменения θ, но крутой на концах. Чем круче кривая, тем больше степень расслаивания.

Характеристическая функция π является полезным инструментом при конструировании теста. Можно сказать, что π осуществляет преобразование исходного распределения тестируемых в распределение результатов тестирования.

На рис. 2.12 — 2.14 приведены примеры характеристических функций тестов с различной степенью расслаивания в предположении нормального распределения тестируемых.

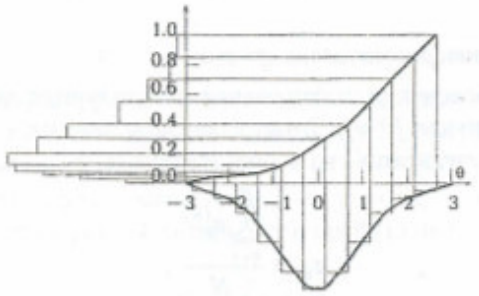


Рис. 2.12. Пример слабо расслаивающей «слабых» тестируемых характеристической функции трудного тест-билета

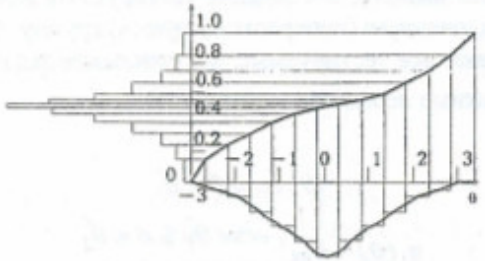


Рис. 2.13. Пример слабо расслаивающей характеристической функции тест-билета средней трудности

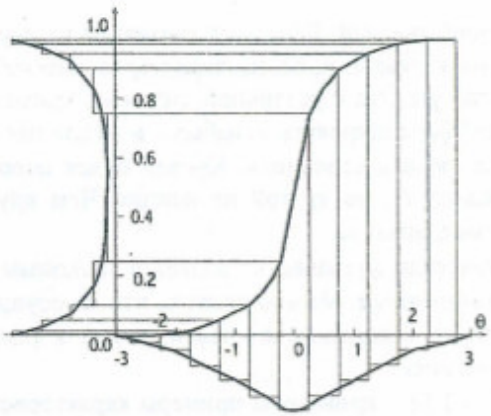


Рис. 2.14. Пример сильно расслаивающей «средних» тестируемых характеристической функции тест-билета

2.3. Модели характеристических кривых

Пусть при проведении тестирования используются задания, которые оцениваются в шкале *[0,m]*. Тогда с каждым заданием свяжем *(m+1)*-мерный вектор результатов его использования $(z_0, z_1, ..., z_m)$,

$$z_{ki} = \frac{\sum_{n=1}^{N} u_{ni}^{(k)}}{N},$$

где $u_{ni}^{(k)}$ — результат выполнения *n*-м тестируемым *i*-го задания, оцененный в *k* баллов, *N* — количество тестируемых.

По результатам выполнения задания тестируемые могут быть разбиты на *m+1* упорядоченную (непересекающуюся) группу: $G_0, G_1, ..., G_m$. В группу G_k попадают все тестируемые, выполнившие задание на *k* баллов и имеющие латентные переменные $\theta \in [\theta_k^*, \theta_{k+1}^*)$.

Функция

$$\pi_i(\theta) = \begin{cases} 0, \text{ если } \theta < \theta_1^* \\ \frac{1}{m}, \text{ если } \theta_1^* \le \theta < \theta_2^* \\ \dots \\ 1, \text{ если } \theta > \theta_m^* \end{cases}$$

является характеристической функцией задания.

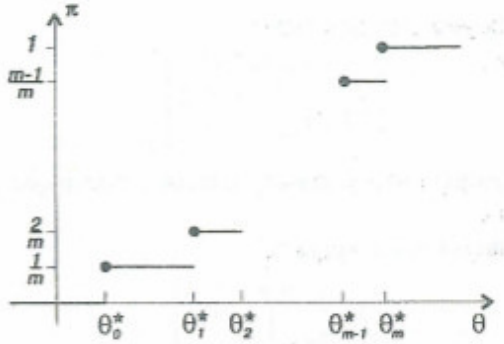


Рис. 2.15

При использовании параметрических моделей задача состоит в подборе кривой из заданного класса, «наилучшим» образом описывающей данную параметрическую характеристическую кривую.

2.3.1. Нормальные модели

С исторической точки зрения, естественно ожидать, что первый класс характеристических кривых, который был рассмотрен, основывался на идее нормального распределения. F. Lord [31]–[33], используя идеи R. Fergusson, рассмотрел параметрические модели, в основе описания характеристических функций которых лежала функция

$$\Phi(\theta) = \int_{-\infty}^{\theta} e^{-\frac{t^2}{2}} dt .$$

Говорят, что рассматривается трехпараметрическая нормальная модель, если для описания характеристических функций используется следующий класс кривых $\Phi_{a,b,c}(\theta)$:

$$\left\{ c + \frac{1-c}{\sqrt{2\pi}} \int_{-\infty}^{a(\theta-b)} e^{-\frac{t^2}{2}} dt \right\}_{a,b,c} .$$

Числа *a, b, c* называются параметрами модели.

Параметр *a* называется дифференцирующей способностью задания.

Параметр *b* — трудностью задания и совпадает с точкой на шкале, в которой значение функции равно *0,5*. Функция в этой точке имеет точку перегиба.

Параметр *c* называется коэффициентом угадывания.

Если используется класс кривых

$$\left\{ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a(\theta-b)} e^{-\frac{t^2}{2}} dt \right\}_{a,b}$$

(т.е. $c=0$), то говорят, что рассматривается двухпараметрическая нормальная модель.

Если используется класс кривых

$$\left\{ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\theta-b} e^{-\frac{t^2}{2}} dt \right\}_{b}$$

(т.е. $c=0$, $a=1$), то говорят, что рассматривается однопараметрическая нормальная модель.

В настоящее время изучение нормальных моделей представляет лишь теоретический интерес, поскольку их практическое использование затруднено вычислительными трудностями и наличием более практичных логистических моделей.

2.3.2. Логистические модели

Для описания характеристических функций A. Birnbaum [20] предложил использовать более простые функции, получившие название логистических:

$$LGT(z) = \frac{e^z}{1+e^z}.$$

Дело в том, что, с практической точки зрения, функции $\Phi(z)$ и $LGT(1.7z)$ отличаются на всей числовой оси не более чем на 1% их значений, но, с математической точки зрения, существенно более просты в работе. Более того, опыт работы показал, что наибольшее количество новых идей и приложений item response theory связано именно с логистическими функциями.

По аналогии с нормальными моделями различают одно-, двух- и трехпараметрические модели. A. Birnbaum рассмотрел двух- и трехпараметрические модели. Однопараметрическую модель исследовал G. Rasch [37].

Опишем наиболее часто встречающиеся и используемые логистические модели характеристических кривых.

Название моделей	Класс функций
однопараметрическая	$\left\{ \frac{1}{1+e^{-d(\theta-b)}} \right\}_{b}$
двухпараметрическая	$\left\{ \frac{1}{1+e^{-da(\theta-b)}} \right\}_{a,b}$
трехпараметрическая	$\left\{ c + \frac{1-c}{1+e^{-da(\theta-b)}} \right\}_{a,b,c}$

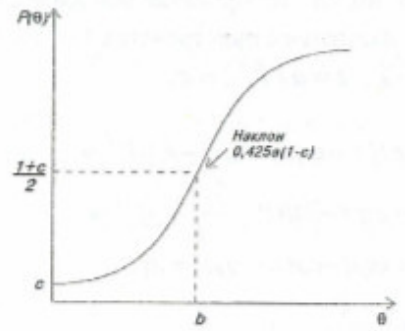


Рис. 2.15. Характеристическая функция трехпараметрической логистической модели

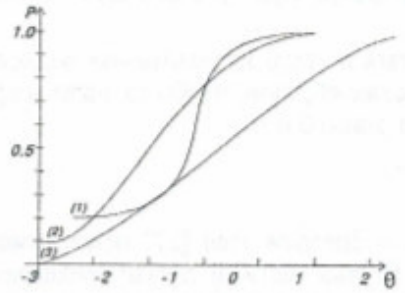


Рис. 2.16. Примеры трех логистических кривых с различными параметрами: *(1) a=5 b=0 c=0,2; (2) a=1,5 b=-1 c=0,1; (3) a=0,5 b=1 c=0*

Параметры трехпараметрической логистической модели называются:

- a — дифференцирующая способность задания,
- b — трудность задания,
- c — коэффициент угадывания.

Константа d обычно принимается равной 1,7.

Чем больше a, тем круче характеристическая кривая, т. е. больше дифференцирующая способность задания расслаивать тестируемых.

Чем больше b, тем больше трудность задания.

Коэффициент угадывания c обычно рассматривается в заданиях закрытого типа, где вероятность угадывания правильного ответа довольно существенна.

Предложение. Характеристическая функция задания

$$\pi(\theta) = c + \frac{1-c}{1+e^{-da(\theta-b)}}$$

трехпараметрической модели инвариантна относительно линейного преобразования $\theta \to \hat{\theta}$ латентного пространства Ω : $\hat{\theta} = l\theta + k$.

При этом $\hat{b} = lb + k$, $\hat{a} = a/l$, $\hat{c} = c$.

Доказательство:

$$\pi(\hat{\theta}) = \hat{c} + (1-\hat{c})[1 + exp(-d\hat{a}(\hat{\theta} - \hat{b}))]^{-1} =$$

$$= \hat{c} + (1-\hat{c})[1 + exp(-da l(\theta - \frac{\hat{b}-k}{l}))]^{-1} =$$

$$= c + (1-c)[1 + exp(-da(\theta - b))] = \pi(\theta),$$

где $a = \hat{a}l$, $b = \frac{\hat{b}-k}{k}$.

Из этого предложения следует, что шкала значений θ до некоторой степени произвольна, поскольку любые две шкалы связаны простыми линейными соотношениями. При этом вид характеристической функции не изменяется.

Один из возможных и часто используемых подходов в выборе шкалы — подбор l и k таким образом, чтобы медиана и среднее отклонение были соответственно равны 0 и 1.

2.3.3. Модель Раша

Датский математик Джордж Раш [37] независимо от исследований Лорда и Бирнбаума изучил частный случай трехпараметрической логистической модели Бирнбаума, в которой предполагал:

1) все задания имеют одинаковый коэффициент селективности;

2) коэффициент угадывания пренебрежимо мал.

В этом случае характеристическая функция задания может быть записана в виде:

$$p_i(\theta) = \frac{1}{1+e^{-da(\theta-\beta_i)}}, \qquad (1)$$

где α — средний коэффициент селективности заданий.

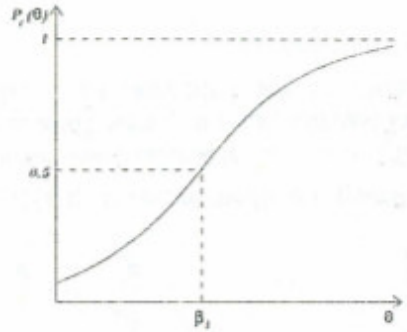


Рис. 2.17. Характеристическая функция модели Раша

В этом случае точкой перегиба характеристической функции является значение $\theta = \beta_i$. Значение функции в этой точке равно 0,5. Таким образом, в этой модели испытуемый со значением латентной переменной $\theta = \beta_i$ ответит корректно на это задание с вероятностью, равной 0,5.

2.3.4. Модель Гутмана

Guttman [26], [27] предложил в качестве характеристических функций дихотомических заданий использовать функции вида

$$P_i(\theta) = \begin{cases} 1, \theta \geq b_i \\ 0, \theta < b_i \end{cases}.$$

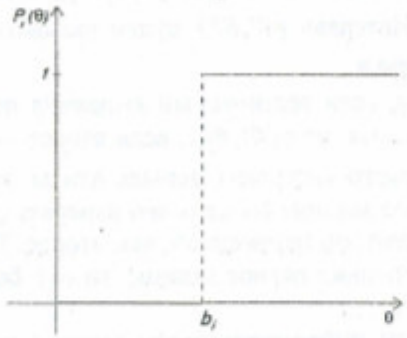


Рис. 2.18. Характеристическая функция модели Гутмана для дихотомических заданий

Можно сказать, что модель Гутмана является предельным случаем двухпараметрической логистической модели при $a \to \infty$.

2.4. Информационные функции

Как отмечалось ранее, степень расслаивания тестируемых зависит от крутизны характеристической кривой. Рассмотрим в качестве примера два различных задания u_1 и u_2 с характеристическими функциями π_1 и π_2 с различной крутизной (т.е. производная $\pi_1' \geq \pi_2'$):

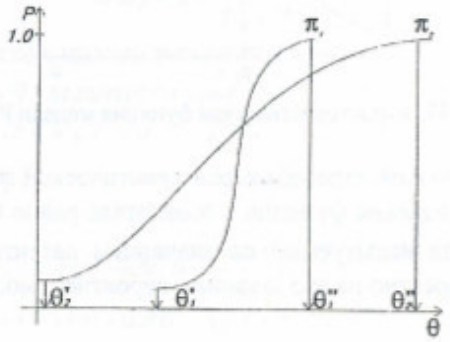


Рис. 2.19. Характеристические функции с различной степенью крутизны

Нетрудно видеть, что кривая π_1 различает тестируемых с латентными переменными из интервала (θ_1', θ_1'') и не различает их левее θ_1' и правее θ_1''. Аналогично, кривая π_2 различает тестируемых лишь с латентными переменными из интервала (θ_2', θ_2''). Поскольку кривая π_1 круче, чем кривая π_2, то длина интервала (θ_1', θ_1'') будет меньше длины интервала (θ_2', θ_2''). Интервал (θ', θ'') будем называть носителем задания u и обозначать *supp u*.

С другой стороны, если тестируемый выполнил первое задание, то его латентная переменная $\theta^* \in (\theta_1', \theta_1'')$, если второе — то $\theta^* \in (\theta_2', \theta_2'')$. Поскольку длина первого интервала меньше длины второго, то можно утверждать, что первое задание более точно измеряет (либо не измеряет вообще) значение латентной переменной, чем второе. Другими словами, если тестируемый выполнил первое задание, то оно более информативно, чем второе.

Зависимость между информативностью задания и латентной переменной θ описывается информационной функцией задания. Более точно, информационной функцией задания называется функция

$$J(\theta, u_i) = \frac{[\pi_i'(\theta)]^2}{\pi_i(\theta) \cdot [1 - \pi_i(\theta)]}, \qquad (2)$$

где $\pi_i(\theta)$ — характеристическая функция задания u_i, $\pi_i'(\theta)$ — производная этой функции.

В качестве примера рассмотрим вид информационных функций заданий, если их характеристические функции описываются логистическими кривыми. Подставляя в (2) конкретные значения $\pi_i(\theta)$ для одно-, двух- и трехпараметрических логистических моделей, получим:

Модель	Вид информационной функции
однопараметрическая	$I(\theta, u_i) = 1.7\, p_i'(\theta) = \dfrac{(1.7)^2 e^{-1.7(\theta - b_i)}}{\left(1 + e^{-1.7(\theta - b_i)}\right)^2}$
двухпараметрическая	$I(\theta, u_i) = 1.7 a_i p_i'(\theta) = \dfrac{(1.7)^2 a_i e^{-1.7 a_i(\theta - b_i)}}{\left(1 + e^{-1.7 a_i(\theta - b_i)}\right)^2}$
трехпараметрическая	$I(\theta, u_i) = 1.7 \dfrac{a_i}{1 - c_i} \dfrac{p_i - c_i}{p_i} p_i'(\theta)$

Нетрудно видеть, что для одно- и двухпараметрической логистической модели информационная функция принимает максимальное значение при аргументе, равном b_i.

В трехпараметрической логистической модели экстремум информационной функции достигается [28] в точке

$$\theta = b_i + \frac{1}{1.7 a_i} \ln \frac{1 + \sqrt{1 + 8c_i}}{2}.$$

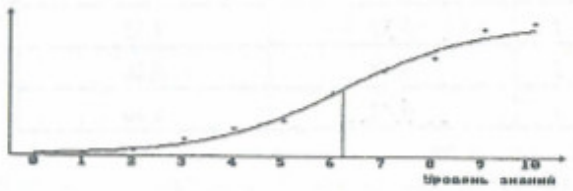


Рис. 2.20. Характеристическая кривая файла заданий (двухпараметрическая модель, a=0,44, b=6,20)

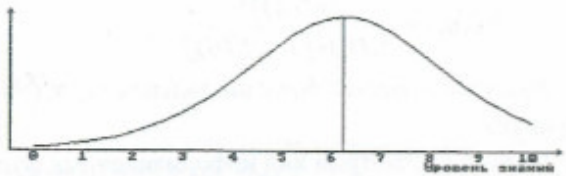


Рис. 2.21. Информационная функция файла заданий
(двухпараметрическая модель, a=0,44, b=6,20)

Информационной функцией тест-билета, состоящего из n дихотомических заданий, называется функция

$$J(\theta) = \sum_{i=1}^{n} J(\theta, u_i) = \sum_{i=1}^{n} \frac{[\pi_i'(\theta)]^2}{\pi_i(\theta)[1-\pi_i(\theta)]}. \qquad (3)$$

Информационная функция была введена A. Birnbaum в 1968 г. для оценивания эффективности каждого задания и тест-билета в целом. Основная идея — минимизировать ошибки измерения.

Если в классических моделях тестирования стандартная ошибка измерения не зависит от θ и определяется «в среднем», то в моделях IRT ошибка измерения является функцией от θ:

$$SE(\theta) = \frac{1}{\sqrt{J(\theta)}}.$$

Следует отметить, что термин «информационная функция» не совсем удачен. Корректнее эту функцию можно было бы назвать функцией точности измерения.

В качестве примера рассмотрим информационную функцию тест-билета, состоящего из трех дихотомических заданий, характеристические функции которых описываются логистическими кривыми с параметрами:

n	a_i	b_i
1	0,64	1,51
2	0,44	6,20
3	0,42	6,96

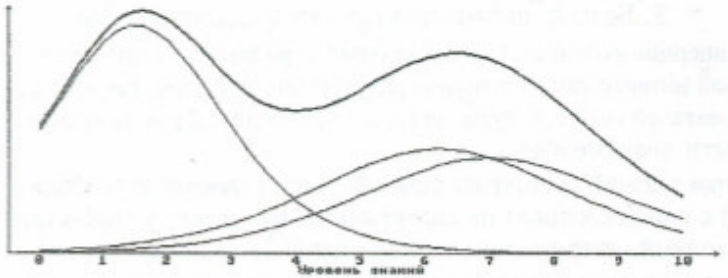


Рис. 2.22. Информационная функция из 3 заданий

Заметим, что «вклад» каждого задания в эффективность измерения не зависит от «вклада» других заданий, составляющих тест-билет. Это существенное отличие IRT от классической теории тестирования, где «вклад» каждого задания в надежность и/или валидность тест-билета зависит от заданий, составляющих этот тест-билет.

Равенство (3) делает возможным эффективную процедуру конструирования тест-билета из отдельных откалиброванных заданий (т. е. заданий, статистические характеристики которых получены в результате пилотных испытаний).

Два тест-билета могут быть сравнимы с точки зрения их информационных функций.

Отношение информационной функции тест-билета T_1 к информационной функции тест-билета T_2 называют *относительной эффективностью (relative efficiency)* тест-билета T_1 по отношению к T_2.

Процедура сравнения информационных функций различных версий тест-билета (добавляя или убирая задания различной степени трудности) может оказаться весьма полезной при его конструировании.

Информационные функции заданий и тест-билета в целом — важнейшее понятие в Item Responce Theory. Во-первых, с помощью этих понятий определяется стандартная ошибка измерения для каждого задания в зависимости от значения латентной переменной θ. (В этом существенное отличие современных подходов от классической теории тестирования, в которой стандартная ошибка измерения определяется «в среднем» для всех тестируемых.) Во-вторых, информационные функции позволяют оценивать «вклад» каждого отдельного задания. Добавляя или удаляя задания, можно отслеживать эффективность тест-билета в целом как «измерительного устройства». Это свойство информационных функций делает их весьма удобным инструментом при конструировании тест-билетов.

3. Базы заданий для проектирования ПИМ

Концепция создания банков заданий в последние годы представляет большой интерес как со стороны разработчиков тестов, так и со стороны пользователей (школы, вузы, органы управления образованием, службы занятости, военные и др.).

Банки заданий состоят из заданий (точнее файлов «однородных» заданий) с описывающими их характеристиками и могут рассматриваться как исходный материал для создания тест-билетов.

Процесс создания тест-билета с помощью имеющегося банка заданий представляет собой более легкую процедуру, чем процесс разработки «своего собственного» набора заданий и формирования структуры теста. Более того, наличие специального программного обеспечения для ЭВМ (например, ТестГен 3.0) позволяет сделать процедуру создания тест-билета весьма технологичной.

Особую значимость базы заданий приобретают при конструировании мультимедиа-тестов, поскольку основная (и довольно большая) стоимость разработки заключена в подготовке и калибровке именно мультимедиа-заданий.

Можно предположить, что в ближайшем будущем значение банков заданий будет возрастать. Чем больше создано банков заданий, тем более эффективной и качественной становится процедура создания тест-билета.

3.1. Технологические, экспертные и статистические параметры заданий

Для описания баз тестовых заданий и проектирования тест-билетов удобно подразделять параметры заданий на три категории: технологические, экспертные и статистические.

К *технологическим* относят параметры, описывающие задания с технической (не содержательной) точки зрения. Как правило, эти характеристики допускают однозначное толкование. Примеры таких параметров:

1) тип заданий:

- закрытое — задание, форма представления которого допускает дихотомическую оценку правильности выполнения — верно/неверно. Проверка таких заданий не требует привлечения педагогов;
- открытое — задание, форма представления которого требует для проверки привлечения педагогов: задания открытого типа обычно проверяются по многобалльной шкале (чаще всего *(10+1)*-балльной):

$$open(u_n) = \begin{cases} 1, \text{если задание открытое} \\ 0, \text{если задание закрытое} \end{cases}$$

2) форма представления информации:

- текст $u_i \in G_{text}$,
- графика $u_i \in G_{picter}$,
- фото $u_i \in G_{foto}$,
- аудио $u_i \in G_{audio}$,
- видео $u_i \in G_{video}$;

3) время представления задания (для аудио- и видео- тестовых заданий) τ_n;

4) возможность компьютерного представления $u_i \in G_{comp}$;

5) место (в мм), необходимое для выполнения задания;

6) требуемый объем памяти компьютера (внешние накопители) для хранения задания (файла заданий);

7) количество заданий в файле.

К *экспертным* относят параметры, значения которых описываются экспертами. Примеры таких параметров:

- количество учебных элементов p_i,
- время выполнения задания t_n,
- сложность (уровень усвоения знаний) α_i,
- ступень абстракции β_i,
- степень осознанности усвоения знаний γ_i,
- экспертная трудность.

К *статистическим* относят параметры тестовых заданий, которые вычисляются по результатам пилотных испытаний (предтестирование). По мере проведения педагогических измерений статистика может накапливаться и могут вноситься соответствующие коррективы. Примеры статистических параметров:

1) коэффициент решаемости задания k_i;

2) коэффициенты селективности задания D_i и r_{ipb};

3) коэффициент привлекательности дистракторов;

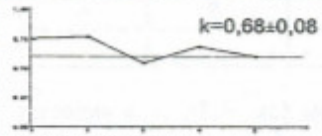
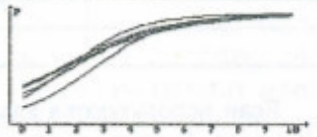
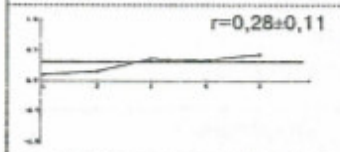
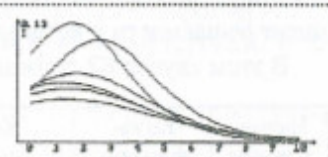
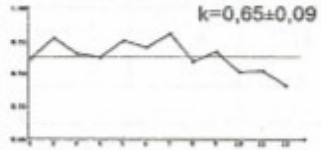
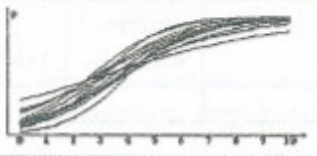
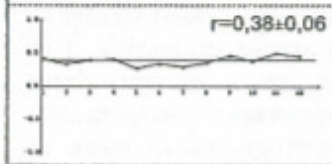
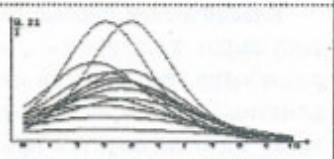
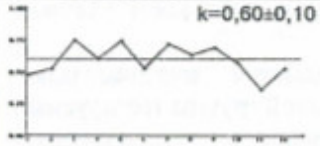
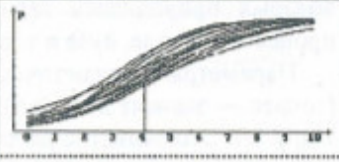
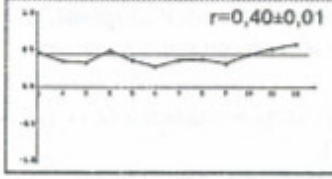
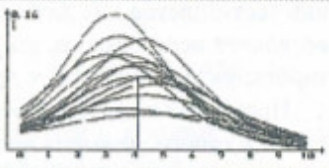
4) параметры логистических кривых:

- дифференцирующая способность задания a_i,
- трудность задания b_i,
- коэффициент угадывания c_i.

Таблица 1. Пример описания банка заданий по математике

Название набора (файла) заданий/ адрес на сервере T:\E_m1_96 \Basa_odn\	Типовой пример	Технологические характеристики			Экспертные характеристики			
		тип задания (открытый, закрытый)	количество задач в наборе	год использования	число контролируемых элементов и их описание p		среднее время выполнения задания, мин	уровень усвоения (по бесталько)
1. Пропорция. Выполнение арифметических действий. 01.txt	Вычислить x, если $\frac{5-0{,}8\div\frac{3}{8}}{2{,}5-2\frac{1}{3}}=\frac{7x}{2{,}5+3\frac{1}{3}}$	O	5	1996	1	- действия с дробями	10	1
2. Алгебраические выражения. Преобразование алгебраических выражений. 02.txt	Упростить: $\left(\frac{a^3+1}{a+1}-a\right)\div+(1-a^3)\div\frac{2a}{1+a}$	O	12	1996	4	- сокращение алгебраических дробей, - приведение алгебраических дробей к общему знаменателю, - действия над алгебраическими дробями, - использование формул сокращенного умножения	10	1
3. Неравенства. Решение системы линейных неравенств с одной переменной. 03.txt	Найти среднее арифметическое целых решений системы неравенств: $\begin{cases}\frac{3x-2}{4}+\frac{x}{2}<\\ <\frac{x-3}{2}-3\end{cases}$	O	12	1996	4	- использование свойств числовых неравенств, - решение линейных неравенств, - нахождение общего решения системы линейных неравенств, - умение относить число к заданному числовому множеству	10	1

(Единый экзамен в Республике Марий Эл, 1996)

Статистические характеристики		
классические	Item Responce Theory	
k — коэффициент решаемости заданий r — коэффициент селективности заданий	a — дифференцирующая способность b — уровень трудности задания w — весовой коэффициент	графические характеристики: характеристическая и информационная кривые
 k=0,68±0,08	a=0,29 b=1,40 w=0,49	
 r=0,28±0,11		
 k=0,65±0,09	a=0,39 b=3,42 w=0,66	
 r=0,38±0,06		
 k=0,60±0,10	a=0,34 b=4,07 w=0,58	
 r=0,40±0,01		

3.2. Структура базы заданий (БЗ)

При разработке структуры БЗ возникает вопрос о том, какие параметры заданий должны быть описаны. Ответ на этот вопрос естественно зависит от тех моделей, которыми пользуется составитель при проектировании тест-билета.

Так, для проектирования тест-билетов с помощью процедуры В.П. Беспалько (см. 4.2) структура БЗ должна иметь следующий вид:

Наименование задания	Перечень учебных элементов	Параметры		
		α	β	γ

Если используются классические процедуры (см. 4.3), то в качестве основных (статистических) параметров должны быть указаны коэффициент решаемости и коэффициент селективности.

В этом случае БЗ должны иметь следующую структуру:

Наименование задания	Коэффициент решаемости k_i	Коэффициент селективности D_i	Коэффициент решаемости дистракторов			
			$k_i^{(1)}$	$k_i^{(2)}$	$k_i^{(3)}$	$k_i^{(4)}$

Классические параметры существенно (и часто нелинейно) зависят от популяции тестируемых, на которых была получена статистика. При разработке тест-билетов классические параметры (они интуитивно более понятны) удобны на этапе первичного конструирования тест-билета либо когда БЗ используется примерно для той же категории испытуемых, на которых проводилась калибровка заданий, при стабильном учебном процессе в школе, вузе и т. д.

Параметры логистических кривых являются инвариантными (точнее — зависят линейно) при переходе от одной группы тестируемых к другой. Это свойство делает их весьма практичными при проектировании тест-билетов для любой популяции тестируемых, с одной стороны, и позволяет использовать широкий спектр современных процедур проектирования тест-билетов, с другой.

Примером комплексного использования параметров является база заданий Единого экзамена по математике (табл. 1).

3.3. Калибровка заданий. Экспертные методы

Процедуру описания характеристик заданий называют *калибровкой*. Одним из наиболее простых экспертных методов калибровки является ***метод комиссии***.

Суть этого метода состоит в открытой дискуссии с целью выработки единого мнения. Каждый из членов комиссии опытных педагогов (экспертов) аргументированно обосновывает свою точку зрения по характеристикам представленных тестовых заданий. Решения о присвоении той или иной характеристики тестовому заданию достигается либо путем консенсуса, либо путем голосования.

Достоинством метода является открытое аргументированное обсуждение, когда растет уровень информированности экспертов и происходит изменение первоначальной точки зрения экспертов. Недостаток метода — возможность «давления» со стороны более авторитетных экспертов, что, однако, не гарантирует компетентности в оценке характеристик. Более того, активность ряда экспертов может не коррелировать с их компетентностью.

Другим широко распространенным методом в проведении экспертизы является ***метод Делфи***.

Суть метода состоит в создании условий, обеспечивающих наиболее продуктивную работу экспертов.

Оценивание происходит в несколько этапов. На первом этапе аналитические группы готовят анкеты эксперта и необходимую сопроводительную информацию.

Таблица 2. Пример анкеты для оценивания (калибровки) экспертных характеристик тестовых заданий

Задания	Экспертные характеристики					
	Перечень учебных элементов	Экспертная трудность	Время выполнения задания t (в мин.)	Уровень усвоения знаний α	Ступень абстракции β	Степень осознанности усвоения знаний γ
...						

Анкета передается или рассылается экспертам. Для этой цели удобно использовать возможности Internet (или электронной почты) и электронные шаблонные формы (например, в формате Excel), которые позволят оперативно производить обработку полученных результатов.

Обработка состоит в следующем:

- определение экспертов, предоставивших «крайние» оценки характеристик;
- усредненное мнение экспертов;
- представление разброса экспертных оценок.

На втором этапе экспертам представляются усредненные оценки экспертов и (анонимное) обоснование экспертов, предоставивших «крайние» оценки. После получения этой дополнительной информации эксперты высылают свои новые откорректированные оценки. После обработки вновь полученной информации проводится третий этап, аналогичный второму. Процедура завершается, когда оценки экспертов стабилизируются. В некоторых случаях процедура охватывает четыре-пять этапов.

Как показывает опыт, метод Делфи является достаточно надежным инструментом для получения оценок экспертных характеристик тестовых заданий.

3.4. Калибровка заданий. Статистические методы

Пусть по результатам тестирования получена матрица ответов D размером $N \times L$, причем произведена выбраковка строк и столбцов, целиком состоящих из нулей и единиц (m — для политомических заданий).

Нахождение статистических классических параметров не вызывает трудностей и проводится по формулам:

коэффициент решаемости задания: $k_i = \frac{1}{N}\sum_{n=1}^{N}\frac{u_{ni}}{m} \quad (i=\overline{1,n})$,

коэффициент селективности задания: $D_i = k_i^{'} - k_i^{''}$,

где $k_i^{'}$ — коэффициент решаемости i-го задания лучшей половины тестируемых; $k_i^{''}$ — коэффициент решаемости i-го задания худшей половины тестируемых.

Вычисление статистических параметров Item Responce Theory представляет собой более сложную задачу. Рассмотрим несколько методов ее решения. В качестве первого рассмотрим относительно простой метод PROX, предложенный L. Cohen в 1976 г. (см. [48]) для модели Раша.

Метод PROX для модели Раша

Предполагается, что тест составлен из дихотомических заданий, латентные переменные испытуемых и трудность заданий в тест-билете распределены нормально:

$$\theta_n \approx N(M,\sigma^2) \quad и \quad \delta_l \approx N(H,\omega^2)$$

Алгоритм вычисления:

1. Вычисляются коэффициенты решаемости заданий и коэффициенты выполнения испытуемыми тест-билета:

$$k_l = \frac{1}{N}\sum_{n=1}^{N} u_{nl} \quad (l=\overline{1,L}), \quad r_n = \frac{1}{L}\sum_{l=1}^{L} u_{nl} \quad (n=\overline{1,N}).$$

2. Вычисляется (в логитах) трудность заданий и латентные переменные испытуемых (см. рис. 3.1):

$$x_l = ln\left(\frac{1}{k_l} - 1\right), \quad y_n = ln\left(\frac{1}{r_n} - 1\right).$$

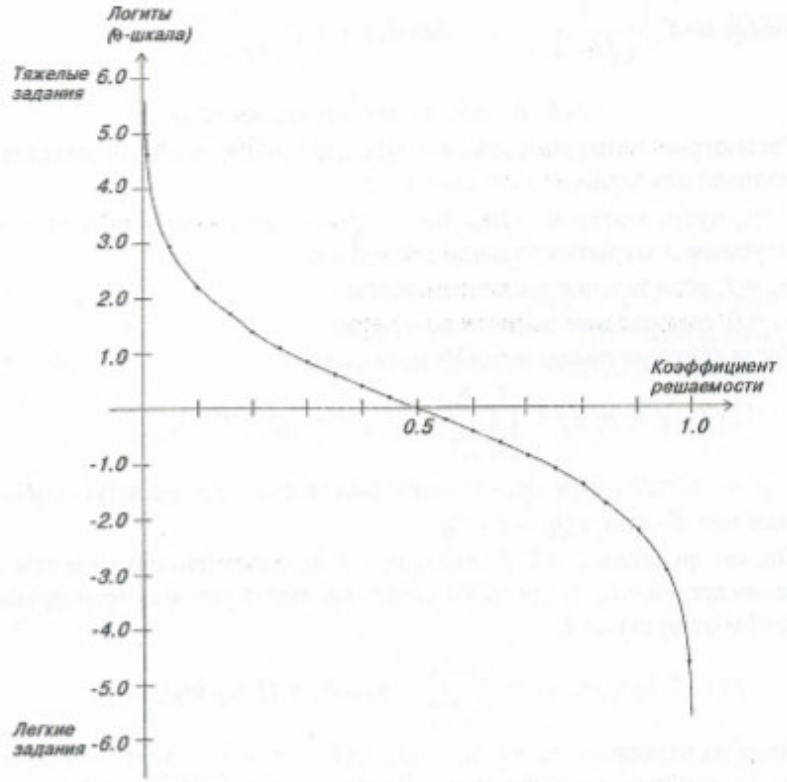


Рис. 3.1. Пример зависимости латентной переменной и коэффициента решаемости

3. Вычисляются средние значения и дисперсии:

$$\overline{X} = \frac{1}{L}\sum_{l=1}^{L} x_l\,, \qquad \overline{Y} = \frac{1}{N}\sum_{n=1}^{N} y_n\,, \qquad D = \frac{1}{(1.7)^2 (L-1)} \cdot \sum_{l=1}^{L} (x_l - \overline{X})^2\,,$$

$$B = \frac{1}{(1.7)^2 (N-1)} \cdot \sum_{n=1}^{N} (y_n - \overline{Y})^2\,, \qquad G = B \cdot D\,.$$

4. Вычисляются поправочные коэффициенты:

$$X = \sqrt{\frac{1+D}{1-G}}\,, \qquad Y = \sqrt{\frac{1+B}{1-G}}\,.$$

5. Оценка трудности заданий β_i и латентных переменных θ_n:

$$\beta_i = Y(x_i - \overline{X}) \quad (i = \overline{1, L})\,, \qquad \theta_n = X \cdot y_n \quad (n = \overline{1, N})\,.$$

6. Находятся стандартные ошибки:

$$SE(\beta_i) = Y\sqrt{\frac{1}{k_i(1-k_i)}}\,, \qquad SE(\theta_n) = X\sqrt{\frac{1}{r_n(1-r_n)}}\,.$$

Метод наибольшего правдоподобия

Рассмотрим метод наибольшего правдоподобия, который достаточно эффективен для заданий закрытого типа.

Итак, пусть вектор $u_i = \{u_{i1}, u_{i2}, \ldots, u_{in}\}$ — результат выполнения i-м испытуемым n закрытых заданий тест-билета,

где $u_{ij} = 1$, если задание выполнено верно,

$u_{ij} = 0$, если задание выполнено неверно.

Тогда функция правдоподобия имеет вид:

$$L(U \mid q, a; b; c) = \prod_{j=1}^{N}\prod_{i=1}^{n} [(P_{ij})^{u_{ij}} + (Q_{ij})^{(1-u_{ij})}],$$

где Pij — вероятность правильного выполнения i-м испытуемым j-го задания тест-билета и $Qij = 1 - Pij$.

Так как функции L и $\ln L$ достигают максимума при одном и том же значении аргумента, то для вычислительных целей удобно рассматривать логарифм от функции L:

$$\ln L(U \mid q, a; b; c) = \sum_{j=1}^{N}\sum_{i=1}^{n} [u_{ij} \ln P_{ij} + (1-u_{ij}) \ln Q_{ij}]$$

Одно из основных предположений IRT, что все задания тест-билета являются локально независимыми. Предположение о локальной независимости является существенным. Оно означает, что при данном уровне знаний ответ на каждое задание тест-билета не зависит от результатов выполнения остальных его заданий.

Значение $\overline{\theta}$, $\overline{\alpha}$, $\overline{\beta}$, $\overline{c}$, при котором функция правдоподобия достигает максимума, принимают в качестве объективных оценок θ, α, β, c и называют оценками наибольшего правдоподобия.

Неизвестные оценки наибольшего правдоподобия для параметров испытуемых находятся из необходимого условия экстремума функции $\ln L_i$; по каждой из переменных θ, α, β, c. Система уравнений для определения величины θ_i в группе из N испытуемых имеет вид:

$$\frac{\partial \ln L_i\{u_i \mid \theta_i\}}{\partial \theta_i} = 0\,, \quad \text{где } i=1, 2, \ldots, N.$$

Уравнения системы являются нелинейными, и их решение сопряжено с определенными вычислительными трудностями. Но каждое j-е уравнение зависит только от переменной θ_j, следовательно, значения θ_j можно определять независимо.

Система уравнений для определения характеристик тест-билета из n заданий в группе имеет вид:

$$\frac{\partial \ln L_i\{u \mid \alpha_{ij}, \beta_j, c_j\}}{\partial \alpha_j} = 0\,, \quad \text{где } j=1, 2, \ldots, n,$$

$$\frac{\partial \ln L_i\{u \mid \alpha_{ij}, \beta_j, c_j\}}{\partial \beta_j} = 0\,, \qquad \frac{\partial \ln L_i\{u \mid \alpha_{ij}, \beta_j, c_j\}}{\partial c_j} = 0\,.$$

Решение систем правдоподобия проводится по очереди. Сначала полагают известными значения параметра α_j, β_j, c_j, а θ рассматривают как переменную. Затем значения θ_j переопределяют, принимая за новые θ_j, и находят оценки α_j, β_j, c_j, доставляющие максимум функции $\ln L_j$. На втором этапе переопределяют значения α, β, c. Процесс продолжается до тех пор, пока абсолютные значения разностей в результате итераций не станут меньше 0,01:

$$|(\overline{\theta})_{k+1} - (\theta_i)_k| < 0{,}01; \; |(\overline{\beta_i})_{m+1} - (\beta_i)_m| < 0{,}01.$$

Конечно, для реализации этого метода нужны специальные программы. Важным предварительным моментом является выбор хорошего начального приближения при оценивании θ_i и α_j, β_j, c_j, $i=1, 2, \ldots, N$; $j=1, 2, \ldots n$. Начальная оценка уровня i-го испытуемого находится по формуле: $\theta_i^0 = \ln (p_i / q_i)$ $i=1, 2, \ldots N$, где N — число испытуемых, p_i — доля правильных ответов i-го испытуемого на все задания теста, q_i — доля неправильных, т.е. $q_i = 1 - p_i$. Аналогично начальная оценка параметров j-го задания находится по формуле: $\beta_j^0 = \ln (p_{ji} / q_{ji})$ $j=1, 2, \ldots n$, где n — число заданий, p_j — доля правильных ответов всех испытуемых на j-е задания теста, q_j — доля неправильных, т.е. $q_j = 1 - p_j$.

Для нахождения максимального значения функции $ln\ L$ можно использовать любой метод безусловной оптимизации функций нескольких переменных. Анализ показывает, что численные методы нулевого порядка дают плохую и очень медленную сходимость (например, метод покоординатного спуска). Метод Ньютона, использующий матрицу Гессе, также дает неудовлетворительный результат из-за плохой обусловленности матрицы Гессе. Практические расчеты показали, что квазиньютоновский метод Бройдена дает хорошую и устойчивую сходимость.

Метод наименьших квадратов[1]

Для заданий открытого типа достаточно эффективным является метод наименьших квадратов. Алгоритм вычисления параметров j-го задания:

1. В случае заданий открытого типа баллы за задание приводятся к диапазону [0,1].

2. Вычисляются коэффициенты выполнения испытуемыми тест-билета:

$$r_n = \frac{1}{L}\sum_{l=1}^{L} u_{nl} \quad (n = \overline{1,N}).$$

3. Множество всех испытуемых делится на 11 подмножеств $G_0 ... G_{10}$ таким образом, что n-й испытуемый относится к i-й группе, если $0.1i \le r_i < 0.1(i+1)$.

4. Находятся коэффициенты решаемости l-го задания для группы G_i:

$$k_{li} = \frac{1}{M(G_i)}\sum_{n \in G_i} u_{nl} \quad (l = \overline{1,L}),$$

где $M(G_i)$ — мощность множества G_i.

Таким образом, для каждого задания l получаем таблично заданную функцию $K_l(i)$.

5. Считаем, что значение функции $K_l(i)$ находится по формуле:

$$K_l(i) = c + \frac{1-c}{1+e^{-da_l(i-b_l)}}.$$

Исходя им этого, подбираем коэффициенты a_l, b_l, c_l, минимизируя функцию:

$$\sum_{j=0}^{10}\left[\left(c + \frac{1-c}{1+e^{-da_l\left(\frac{j}{10}-b_l\right)}}\right) - k_l(j)\right]^2 \to min.$$

[1] Метод предложен А. В. Ельцыным.

Минимизацию можно производить любым численным методом, например, методом покоординатного спуска.

3.5. Выравнивание заданий в файле

Одна из основных и трудоемких задач при формировании банка файлов тестовых заданий — соблюдение принципа однородности, который предполагает группирование в одном файле заданий, близких по своим технологическим, экспертным и статистическим характеристикам. Важным инструментом для этой цели является введение *мер близости*, которые позволяют определить, насколько «близки» или «далеки» друг от друга тестовые задания.

В качестве примера рассмотрим векторную меру близости $v = (v_k, v_л, v_э)$,

где v_k — составляющая классических статистических характеристик,

$v_л$ — составляющая латентных (IRT) статистических характеристик,

$v_э$ — составляющая экспертных характеристик.

Пусть $u_i = (k_i, D_i, a_i, b_i, p_i, \alpha_i, \beta_i, t_i)$ — тестовое задание с описывающими его статистическими и экспертными характеристиками (параметрами):

k_i — коэффициент решаемости,

D_i — коэффициент селективности (D-индекс),

a_i — дифференцирующая способность,

b_i — трудность,

p_i — количество учебных элементов в задании,

α_i — уровень усвоения,

β_i — ступень абстракции,

t_i — время выполнения.

Тогда u_i можно идентифицировать с точкой в векторном пространстве

$$l^8 = l_k^2 \oplus l_л^2 \oplus l_э^4,$$

где l_k^2 — подпространство классических статистических параметров,

$l_л^2$ — подпространство латентных статистических параметров,

$l_э^4$ — подпространство экспертных параметров.

Для каждой пары тестовых заданий u_i и u_j в пространстве l_k^2, $l_л^2$ и $l_э^4$ введем обычные l^q -нормы с весами:

$$\nu_k(u_i,u_j)=\left[w_k\left|k_i-k_j\right|^q+w_D\left|D_i-D_j\right|^q\right]^{\frac{1}{q}},$$

$$\nu_n(u_i,u_j)=\left[w_a\left|a_i-a_j\right|^q+w_b\left|b_i-b_j\right|^q\right]^{\frac{1}{q}},$$

$$\nu_э(u_i,u_j)=\left[w_p\left|p_i-p_j\right|^q+w_\alpha\left|\alpha_i-\alpha_j\right|^q+W_\beta\left|\beta_i-\beta_j\right|^q+W_t\left|t_i-t_j\right|\right]^{\frac{1}{q}},$$

где $w_k \geq 0$, $w_D \geq 0$, $w_a \geq 0$, $w_b \geq 0$, $w_p \geq 0$, $w_\alpha \geq 0$, $w_t \geq 0$ — весовые коэффициенты.

Тогда $\nu=[\nu_k^q+\nu_n^q+\nu_э^q]^{\frac{1}{q}}$ является интегральной нормой параметров в l^8.

Одна из целей предтестирования — получение статистической информации для принятия решений о приемлемости, неприемлемости и «подозрительности» заданий (т. е. могут обладать определенными техническими недостатками). «Подозрительные» задания не обязательно должны исключаться из банка заданий. Они лишь должны быть подвергнуты дополнительному, более тщательному анализу.

Анализ несоответствия параметров конкретного задания, содержащего его файла заданий, общим параметрам удобно проводить «сверху вниз».

На первом этапе выделяются задания, резко отличающиеся по интегральной норме:

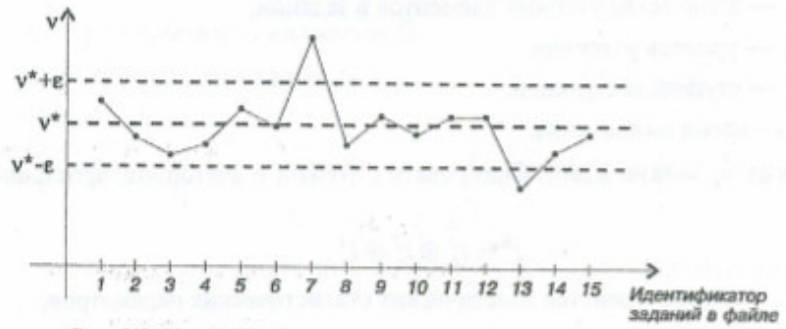


Рис. 3.2. Карта Шухарта интегральной нормы заданий файла, состоящего из 15 заданий. Хорошо видно, что «резко выделяются» седьмое и тринадцатое задания

На втором этапе происходит определение составляющих характеристик, по которым произошел «выброс»:

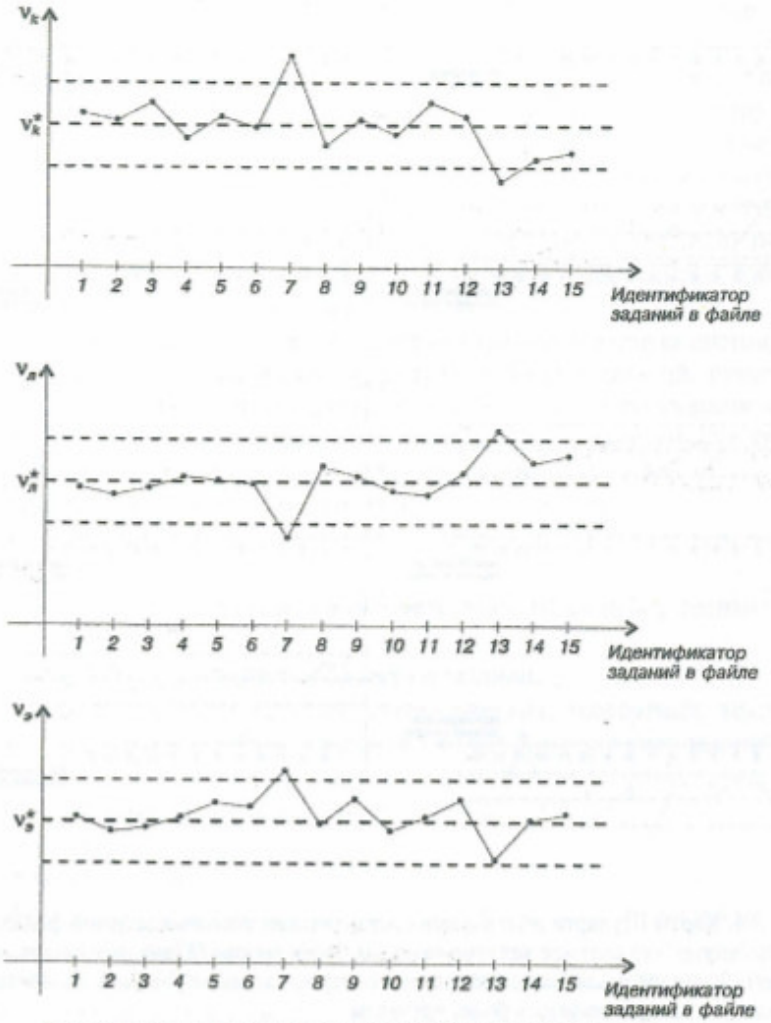


Рис. 3.3. Карты Шухарта норм заданий файла по классическим, латентным статистическим и экспертным составляющим

На третьем этапе находятся конкретные параметры заданий, резко отличающиеся от средних (рис. 3.4).

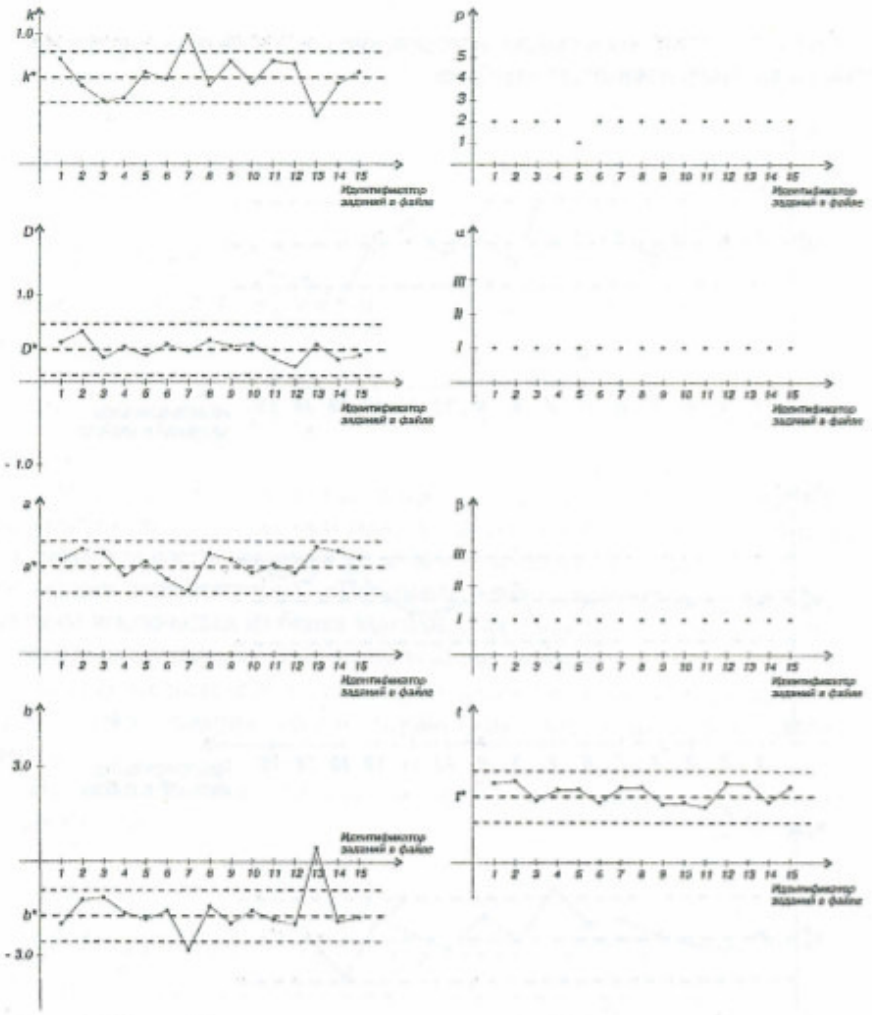


Рис. 3.4. Карты Шухарта всех восьми составляющих тестовых заданий файла. Хорошо видно, что седьмое задание является более легким (k_7 значительно выше среднего, b_7 значительно ниже среднего) и содержит меньше учебных элементов. Тринадцатое задание является более трудным

4. Модели и алгоритмы проектирования ПИМ

Хотя теория тестирования насчитывает сто лет, лишь в последние годы наметился подход к более фундаментальному, почти математическому, описанию процедуры проектирования тестов. Цель настоящего параграфа — выделить и описать различные модели процедур проектирования тестов, как получивших широкое распространение в практике тестирования, так и новых моделей для их использования в процедурах самооценки и аттестации.

В каждой модели дается описание целевой функции и характеристик заданий, используемых как основные.

Первый шаг в этом направлении — четкое определение теста. Если в западной научной литературе под тестом понимается любой инструмент для измерения латентных характеристик испытуемых, то в российской научной литературе с понятием теста связывается лишь определенный тип педагогических измерений. В связи с этим дадим несколько определений.

Под *педагогическими измерительными материалами (ПИМ)* будем понимать любой тип педагогических измерений: контрольные работы, комплексные контрольные задания, тесты и т. п.

Под *тест-билетом* будем понимать структурированную систему заданий и соответствующую ей процедуру проверки заданий, обеспечивающую однозначную интерпретацию результатов педагогических измерений. Если описаны экспертные, статические и технологические характеристики тест-билета, то будем его называть *откалиброванным*.

Под *тест-комплектом* будем понимать совокупность:

- описание процедуры педагогического измерения и условий ее применения;
- структура тест-билета и соответствующий ей набор тестовых заданий;
- ключи и/или критерии проверки заданий;
- описание характеристик (статистических, экспертных, технологических) тест-билета в целом и отдельных тестовых заданий;
- интерпретационную систему (шкалу) и систему описания (толкования) результатов оценивания.

4.1. Базовая модель

Иногда эту процедуру называют неформальной (интуитивной, естественной, учительской), а созданные тест-билеты — неформальными (учительскими). Это объясняется тем, что тест-билеты по данной процедуре готовятся педагогами, исходя из своего большого опыта и для своих нужд. Их (тест-билеты), как правило, отличает высокая степень валидности, но, к сожалению, невысокая степень надежности и других статистических характеристик. Данная процедура не требует специальных знаний теории тестирования и довольна проста в освоении.

Опыт показывает, что можно резко повысить эффективность создания учительских тест-билетов, если пользоваться некоторыми простыми инструментами. В качестве таких инструментов удобно использовать:

- карту Шухарта коэффициентов решаемости (трудности) заданий;
- гистограмму (пилотных) результатов тестирования;
- диаграмму Парето коэффициентов решаемости заданий.

Коэффициент решаемости k — статистический параметр задания, который определяется в результате пилотных испытаний (предтестирование) на группе испытуемых, близкой по уровню тех, для кого готовятся тест-билеты. Для дихотомических заданий k определяется по формуле:

$$k = \frac{M}{N},$$

где N — количество решавших задание,

M — количество решивших задание.

Для политомических — по формуле:

$$k = \frac{1}{N}\sum_{l=1}^{N} l \cdot M_l,$$

где N — количество решавших задание,

M_l — количество решивших данное задание с оценкой, равной l,

l — оценка задания.

Ясно, что коэффициент решаемости задания $0 \le k \le 1$. Если k близко к нулю, то задание тяжелое — его почти никто не решает. Если k близко к единице, то задание легкое — его выполняют почти все. С точки зрения составителя тест-билета, эти задания не представляют интереса, поскольку практически ничего не проверяют.

Структуру решаемости заданий в тест-билете удобно представить в виде так называемой карты Шухарта (см. рис. 4.1): по оси абсцисс откладываются номера заданий, составляющих тест-билет, а по оси ординат — их коэффициенты решаемости. Пунктирные линии разделяют карту Шухарта на три зоны: зона легких заданий, зона средних заданий и зона тяжелых заданий. Большая часть заданий должна находиться в средней зоне. Однако небольшая доля заданий может находиться в нижней (часть заданий должны выполнить почти все) и верхней (задания для наиболее способных испытуемых) зонах.

Второй удобный инструмент визуализации — гистограмма результатов пилотных испытаний (рис. 4.2).

Гистограмма — удобный инструмент при сравнении результатов тестирования, полученных на различных группах испытуемых с помощью одного и того же тест-билета (рис. 4.3).

Таким образом, в базовой модели в качестве основной характеристики выступает коэффициент решаемости, а в качестве целевой функции (если можно так сказать) — форма гистограммы результатов тестирования.

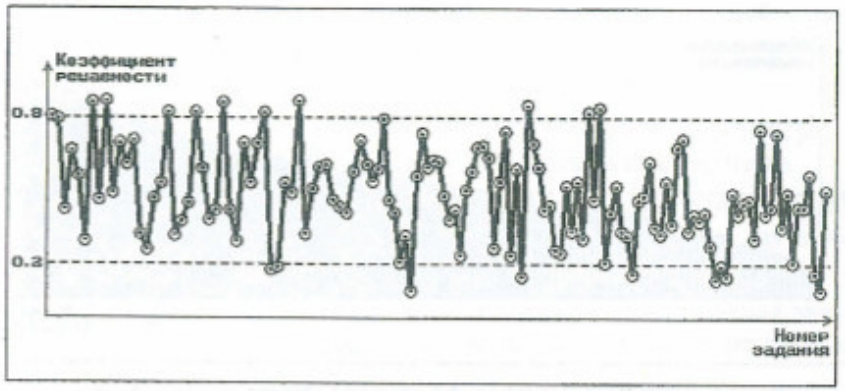


Рис. 4.1. Карта Шухарта коэффициентов решаемости тест-билета по экономике, состоящего из 130 заданий

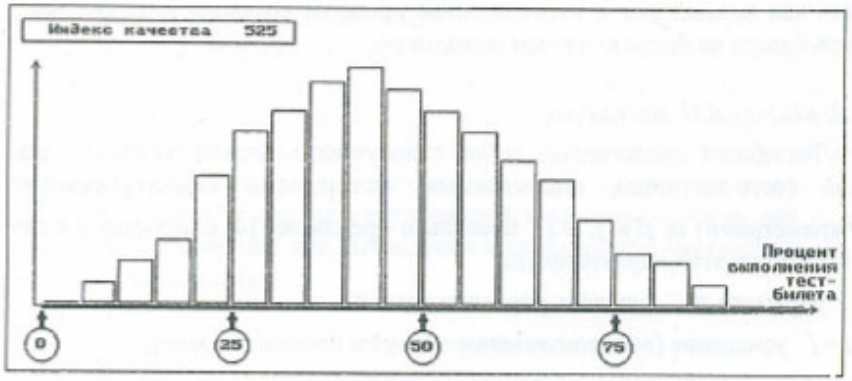


Рис. 4.2. Гистограмма результатов тестирования

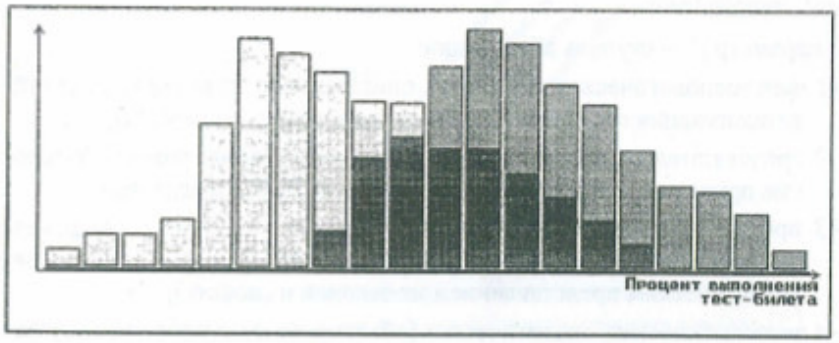


Рис. 4.3. Сравнение гистограмм

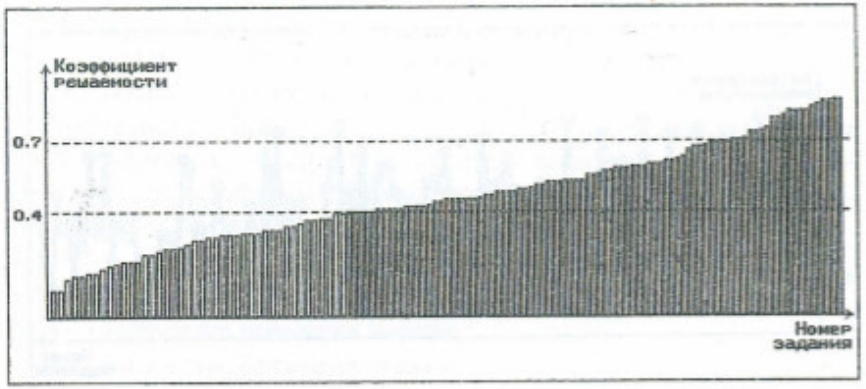


Рис. 4.4. Диаграмма Парето коэффициентов решаемости тест-билета по экономике, состоящего из 130 заданий

Базовую (естественную, интуитивную) процедуру можно рассматривать как первый шаг в итерационном процессе создания качественного тест-билета по более сложным методикам.

4.2. Модель В.П. Беспалько

Тест-билет представляет собой совокупность блоков тестовых заданий (тест-лестница), описываемых экспертными характеристиками (параметрами) α, β и γ. В.П. Беспалько предлагает [4] следующую классификацию этих характеристик:

параметр α — уровень усвоения знаний:

$\alpha=1$ узнавание (воспроизведение знаний с помощью извне);

$\alpha=2$ воспроизведение усвоенных знаний в типовых заданиях;

$\alpha=3$ применение знаний в практической деятельности;

$\alpha=4$ творчество;

параметр β — ступень абстракции:

$\beta=1$ феноменологическая (внешнее, описательное изложение явлений; каталогизация объектов, констатация их свойств и качеств);

$\beta=2$ предсказательная, аналитико-синтетическая (элементарное объяснение природы и свойств объектов, закономерностей явлений);

$\beta=3$ прогностическая, аналитическая (объяснение явлений с созданием их количественной теории, моделирование основных процессов, аналитическим представлением их законов и свойств);

$\beta=4$ аксиоматическая, аналитическая (объяснение явлений с использованием высокой степени абстракции на базе сложного математического или логического формализма, обладающего большой обобщенностью описания);

параметр γ — степень осознанности усвоения знаний:

$\gamma=1$ первая степень осознанности (при аргументации выполнения заданий используется только информация из изучаемого предмета).

$\gamma=2$ вторая степень осознанности (при аргументации выполнения заданий используется информация из близких по объекту изучения предметов).

$\gamma=3$ третья степень осознанности (при аргументации выполнения заданий используются широкие межпредметные связи из различных дисциплин).

Кроме параметров α, β и γ В.П. Беспалько выделяет параметр r — степень автоматизации деятельности [4].

Учащийся считается справившимся с данным уровнем, если он выполнил не менее 70% заданий из блока.

При конструировании тест-билета важное значение имеет определение (априорной) надежности тест-билета. Полагают, что она должна быть не менее 0,75. В.П. Беспалько предложил определять (априорную) надежность тест-билета, исходя из количества учебных элементов (существенных операций в его терминологии). Такой подход весьма удобен на первом этапе конструирования тест-билета, когда нет еще достаточной статистики для вычисления коэффициента надежности статистическими методами.

Рис. 4.5. Оси с параметрами уровня усвоения знаний α, ступени абстракции β и степени осознанности усвоения знаний γ

Под учебным элементом (существенной операцией) понимается элемент выполнения задания, без использования которого невозможно получить верное решение.

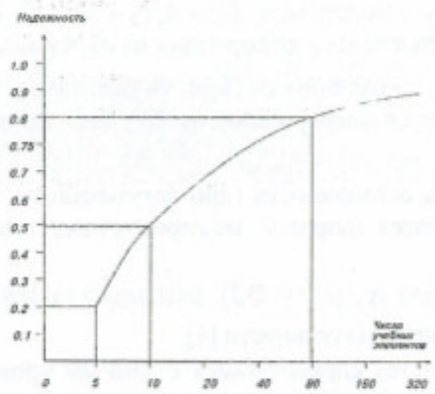


Рис. 4.6. Зависимость априорной надежности тест-билета от количества учебных элементов

Из графика видно, что для (априорной) надежности тест-билета не менее 0,75 необходимо, чтобы он состоял из 40–60 учебных элементов.

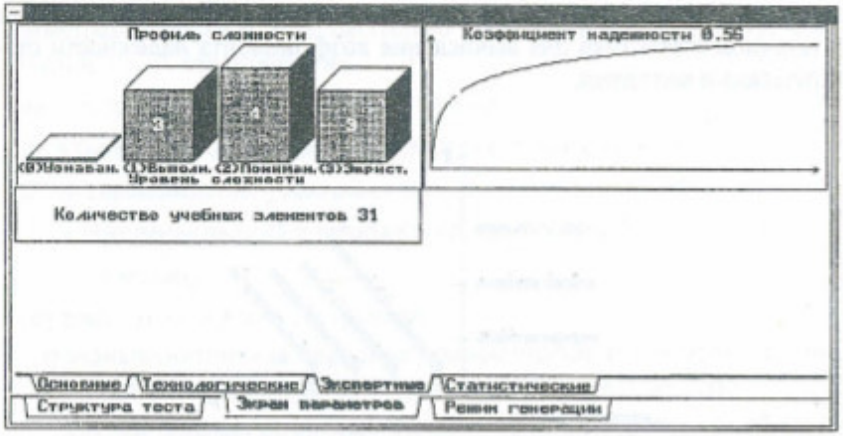


Рис. 4.7. Экран ТестГен 3.0 с экраном основных характеристик модели В.П. Беспалько: диаграмма уровней α, β и γ и график зависимости надежности от количества учебных элементов

Таким образом, можно сказать, что в модели В.П. Беспалько [4]:

целевые функции: надежность $R \to max$,
валидность $V \to max$;

основные характеристики: уровень усвоения α,
ступень абстракции β,
степень осознанности усвоения знаний γ.

4.3. Классическая процедура

С точки зрения классического подхода, тест-билет, как и любой социологический инструмент измерения, должен удовлетворять свойствам высокой степени объективности, надежности и валидности.

Таким образом, в качестве критерия конструирования тест-билета выступают:

надежность $\to max$
валидность $\to max$

Объективность должна обеспечиваться процедурой организации тестирования.

В качестве основных статистических параметров заданий выступают

k — коэффициент решаемости задания (трудность),

D — коэффициент селективности.

Результат тестирования X можно представить в виде суммы:

$$X=T+E,$$

где T — истинное значение (true score) латентной переменной тестируемого, а E — ошибка измерений.

Распределение этих переменных, как правило, неизвестно.

Если тест-билет составлен из дихотомических заданий, то истинное значение:

$$T=\sum_{i=1}^{n} p_i(\theta),$$

если из политомических, то

$$T=\sum_{i=1}^{n}\sum_{k=0}^{m} k p_{i,k}(\theta).$$

Надежность теста ρ^2_{XT} определяется как квадрат корреляции между результатом и истинным значением тестирования и может быть записана как

$$\rho^2_{XT}=\frac{\sigma^2_{XT}}{\sigma^2_X\sigma^2_T}=\frac{\sigma^2_T}{\sigma^2_X}=1-\frac{\sigma^2_E}{\sigma^2_X},$$

где σ_{XT} — коэффициент корреляции X и T,

σ^2_X — дисперсия распределения результатов тестирования,

σ_T^2 — дисперсия распределения латентной переменной,

σ_E^2 — дисперсия распределения ошибки измерения.

Нетрудно видеть, что надежность теста равна единице, если $\sigma_E^2 = 0$.

Такое определение надежности представляет теоретический интерес, но на практике его использование невозможно. Одна из самых распространенных процедур практического нахождения надежности была разработана Kuder G., Richardson M. [30].

С другими процедурами можно ознакомиться по работам [3], [26].

Из различных формул, приведенных в [30], для вычисления коэффициента надежности тестов, составленных из дихотомических заданий, наиболее часто используется формула под номером 20 (получившая впоследствии название КР–20):

$$r = \frac{n}{n-1}\left(1 - \frac{\sum pq}{\sigma_t^2}\right).$$

Если тест-билет составлен из политомических заданий, то в качестве коэффициента надежности обычно используют коэффициент альфа:

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n}\sigma_i^2}{(\sum_{i=1}^{n}\sigma_i\rho_{iX})^2}\right],$$

где n — число заданий в тест-билете,

σ_i^2 — дисперсия i-го задания,

ρ_{iX} — коэффициент корреляции между i-м заданием и тест-билетом.

Таким образом, в классической модели:

целевая функция: коэффициент надежности $R \to max$.

основные характеристики: коэффициент решаемости k,
коэффициент селективности D.

В заключение еще раз отметим, что параметры k и D зависят от конкретной популяции тестируемых и могут эффективно использоваться лишь для аналогичной популяции.

4.4. Модель Лорда-Бирнбаума

В этой модели в качестве критерия построения тест-билета рассматривается максимизация количества «информации», которое можно получить о тестируемом (с уровнем $\theta = \theta_0$). Согласно A. Birnbaum, под

«количеством информации», которое получается при включении в тест-билет i-го задания, понимается величина, обратно пропорциональная стандартной ошибке измерения данного значения θ_0 с помощью i-го задания.

Процедуру можно представить из следующих шагов:

1) строится целевая информационная функция;

2) из банка заданий последовательно выбираются задания таким образом, чтобы коэффициент трудности b_i был как можно ближе к значениям θ, при которых достигается максимум целевой информационной функции;

3) после каждого добавления следующего задания (пере)вычисляется информационная функция тест-билета;

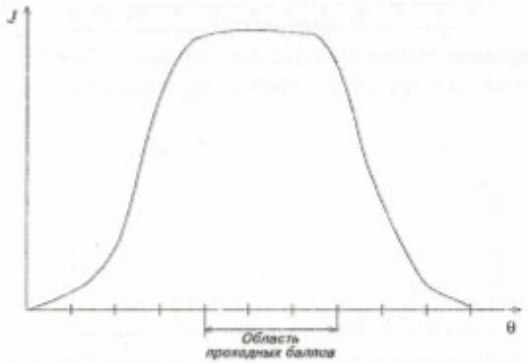


Рис. 3.8. Пример целевой информационной функции
для составления тест-билета для вступительных экзаменов в вуз

4) процедура продолжается до тех пор, пока информационная функция тест-билета не аппроксимирует с достаточной степенью целевую информационную функцию.

Весьма полезной при конструировании тест-билета может оказаться процедура сравнения информационных функций различных версий тест-билета (добавляя или убирая задания различной степени трудности).

В качестве примера рассмотрим информационные функции тест-билета Единого экзамена по математике в Республике Марий Эл (1995 г.) и различных его подтестов (рис. 4.9–4.12).

Таким образом, в модели Лорда-Бирнбаума:

целевая функция: информационная функция $J(\theta)$:

$$\left|J(\theta) - J_u(\theta)\right| \to min;$$

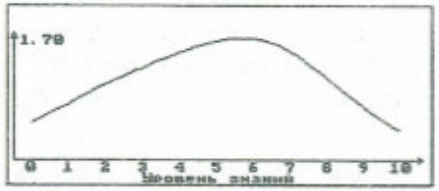


Рис. 4.9. Информационная функция тест-билета Единого экзамена по математике, 1995 г.

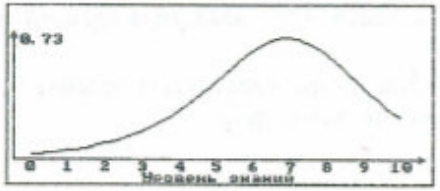


Рис. 4.10. Информационная функция подтест-билета Единого экзамена по математике, 1995 г. (только трудные задания)

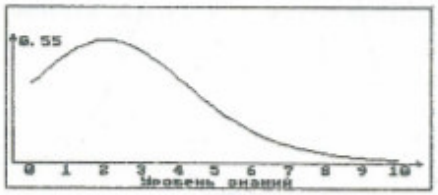


Рис. 4.11. Информационная функция подтест-билета Единого экзамена по математике, 1995 г. (только легкие задания)

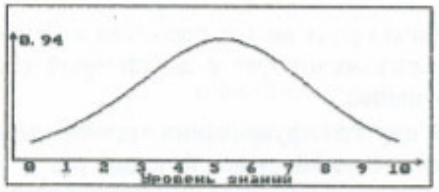


Рис. 4.12. Информационная функция подтест-билета Единого экзамена по математике, 1995 г. (задания средней трудности)

основные характеристики: параметры логистических кривых:
a — дифференцирующая способность,
b — трудность,
c — коэффициент угадывания.

4.5. Процедура В.С. Аванесова

Для минимизации процедуры определения уровня подготовленности тестируемого В.С. Аванесов предложил при конструировании тест-билета располагать тестовые задания в порядке возрастания их трудности: «Педагогический тест — это система фасетных (откалиброванных) заданий определенного содержания, возрастающей трудности, специфической формы, позволяющая качественно оценить структуру и эффективно измерить уровень знаний, умений, навыков и представлений» [2], что эквивалентно (рис. 4.13) убыванию коэффициентов решаемости (рис. 4.14).

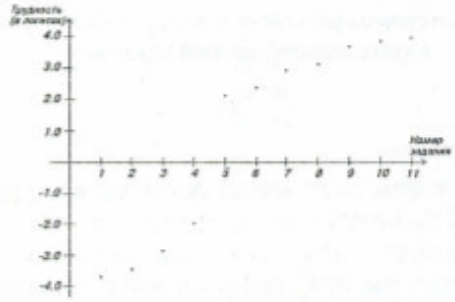


Рис. 4.13. Карта Шухарта коэффициентов трудности задания

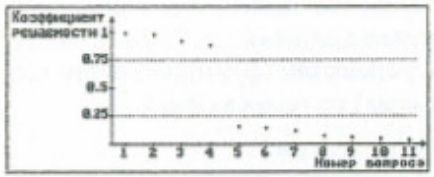


Рис. 4.14. Карта Шухарта с убывающими коэффициентами решаемости

Метод удобен для визуализации расположения в тест-билете заданий различной степени трудности. Однако, как показано на рис. 4.15, необходимо следить за тем, чтобы задания покрывали шкалу логитов достаточно плотно. В противном случае может образоваться большая группа тестируемых, результаты которых не различимы.

Таким образом, в модели В.С. Аванесова:

целевая функция: порядок расположения заданий u_i таким образом, чтобы $b_i \le b_{i+1}$;

основные характеристики: трудность задания b_i.

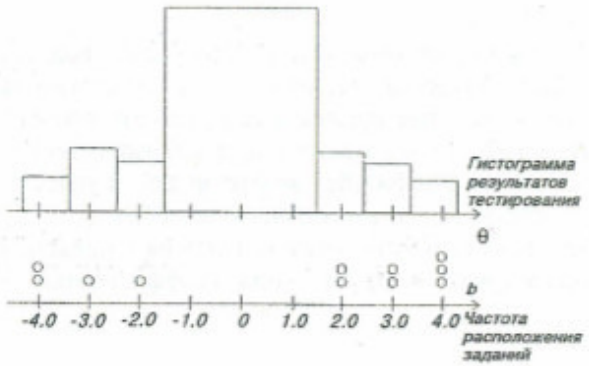


Рис. 4.15. Гистограмма результатов тестирования при отсутствии в тесте заданий средней трудности

4.6. Многоцелевая модель

В связи с появлением компьютерных систем конструирования тест-билетов (ТестГен 3.0) появляется возможность разработки многофункционального тест-билета с использованием различных моделей. Выбор конкретных характеристик при конструировании тест-билета зависит от целей тестирования, а именно — какую информацию желает получить тестолог в результате педагогических испытаний (рис. 4.16).

При конструировании тест-билета с помощью многоцелевой процедуры используется *секвенциальный подход*, т. е. последовательное приближение к целевым установкам (функциям) путем пошагового добавления (изменения, удаления) тестовых заданий.

Опишем алгоритм процедуры:

1) определение (задание) области валидности тестирования.

Этот этап заключается в построении дерева целей тестирования и выборе из банка заданий баз заданий, которые соответствуют выбранной области валидности и из которых будут выбираться тестовые задания;

2) выбор интерпритационной системы для последующего анализа результатов тестирования.

Если для анализа будет использоваться метод уровней, то предварительно производится вербальное описание каждого из уровней. При включении в тест-билет нового задания отмечается уровень, для описания которого оно будет использоваться (рис. 4.17);

3) выбор предполагаемого профиля гистограммы и определение набора $\{\delta_i\}$ «трудностей» тестовых заданий.

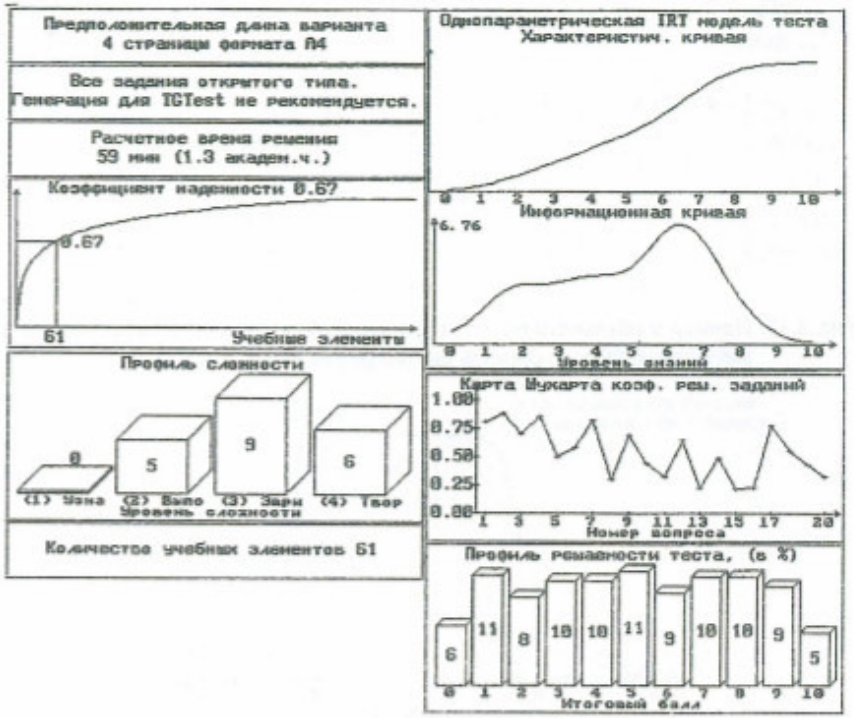


Рис. 4.16. Экран ТестГен 3.0 со «всеми» характеристиками в качестве основных

(а) Выбирается предполагаемый профиль гистограммы и выдвигается гипотеза о «среднем уровне обученности» M и размахе уровня обученности W (рис. 4.18).

При экспертном оценивании M и W может помочь опыт построения аналогичных тест-билетов, экспертная оценка уровня подготовленности учащихся и т. п. Бывает полезно предварительно ознакомиться с заданиями из базы заданий и, основываясь на известных параметрах заданий, предугадать возможный уровень ответов и, следовательно, возможный уровень заданий, с которыми справятся (не справятся) обучаемые.

(б) Полагают среднеквадратичное отклонение $\sigma = \frac{1}{4} W$.

В силу неравенства Чебышева интервал $(M\text{-}2\sigma;\ M\text{+}2\sigma)$ должен содержать 75% тестируемых, а если предполагаемое распределение нормально, то — более 95%.

Рис. 4.17. Пример вербального описания уровней обученности тестируемых по математике за курс средней школы (в соответствии с ГОС)

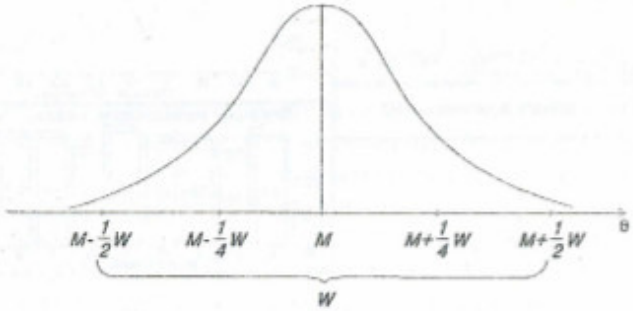
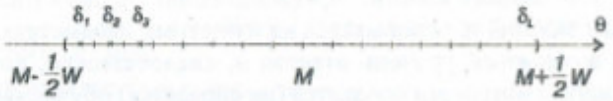


Рис. 4.18

(в) Вычисляется предполагаемая длина тест-билета $L = \frac{W}{S}$,

где S — стандартная ошибка измерений, оцененная каким-либо образом.

(г) Интервал $M - \frac{1}{2}W; M + \frac{1}{2}W$ разбивается на L равных интервалов:



Задания из базы заданий выбираются таким образом, чтобы их трудность принадлежала одному интервалу и так, чтобы каждому интервалу соответствовало по крайней мере одно задание.

4) определение целевой информационной функции и определение степени приближения к ней.

Предварительно задается целевая информационная функция. По мере добавления в тест-билет новых заданий производится визуальное сравнение двух функций.

Так, из рис. 4.19 видно, что в тест-билете не хватает заданий, уровень трудности которых лежит в интервале (δ', δ''). Поэтому на следующем шаге добавляется задание с $\beta_i \in (\delta', \delta'')$ и снова сравнивается информационная функция конструируемого тест-билета с целевой информационной функцией. При необходимости процедуру выполняют несколько раз.

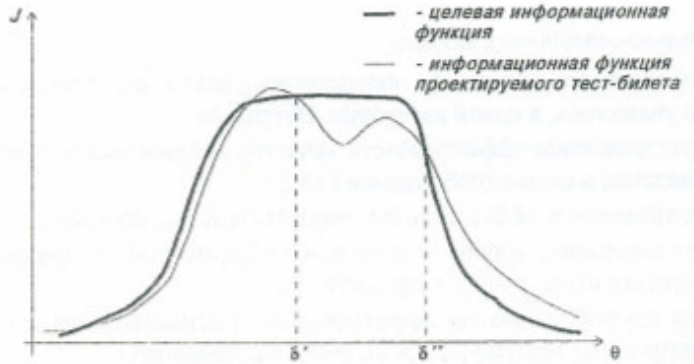


Рис. 4.19. Сравнение информационной функции конструируемого тест-билета с целевой информационной функцией

5) определение априорной надежности тест-билета.

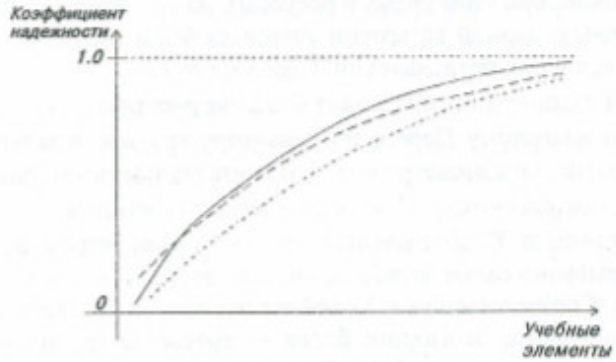


Рис. 4.20. Определение априорной надежности в зависимости от характеристик, составляющих тест-билет тестовых заданий

Как уже отмечалось в п. 4.2, при увеличении количества заданий (а следовательно, и учебных элементов) в тест-билете увеличивается его надежность. При конструировании тест-билета было бы весьма полезно, пользуясь такой зависимостью, оценивать надежность тест-билета еще до этапа его статистической апробации.

Большой ряд статистических испытаний показал, что не существует единой прямой зависимости между надежностью тест-билета и количеством учебных элементов. Эта зависимость существенно зависит от ряда других характеристик, описывающих тест-билет. На рис. 4.20 приведен ряд таких зависимостей для различных типов тест-билетов.

4.7. Модульно-матричные модели

Часто возникает задача об определении уровня подготовки не отдельного учащегося, а целой категории. Например:

- установление эффективности качества образования учебного заведения в целом требованиям ГОС;
- определение эффективности образовательных программ;
- установление уровня подготовки в образовательном учреждении при его аттестации и аккредитации.

В этих случаях довольно эффективными и дешевыми являются модульно-матричные процедуры проведения тестирования.

Пусть для проведения тестирования по какому-либо учебному предмету с помощью одной из процедур, описанных выше, разработан тест-билет T. Идея модульного метода состоит в том, чтобы разбить этот тест-билет на ряд подтест-билетов (модулей) T_1, T_2, ..., T_k таким образом, чтобы каждый учащийся отвечал не на весь тест-билет T, а лишь на его часть T_i (экономя при этом время и ресурсы), но так, чтобы в совокупности относительно данной категории учащихся были получены такие же результаты, как и при предъявлении целого теста T.

Алгоритм такого разбиения может быть следующим:

1) строим диаграмму Парето по параметру трудности заданий (или, что эквивалентно, по параметру коэффициента решаемости) (рис. 4.21);

2) задания разбиваются на блоки по n заданий в каждом;

3) подтест-билет T_1 составляется таким образом, чтобы из первого блока было выбрано самое легкое задание, из второго — второе по трудности и т. д. Соответственно в T_2 войдут задания: из первого блока — второе по трудности, из второго блока — третье по трудности и т. д. Такая компоновка позволяет формировать подтесты примерно одинаковой трудности;

4) полученная совокупность подтест-билетов T_1, ..., T_n является искомой.

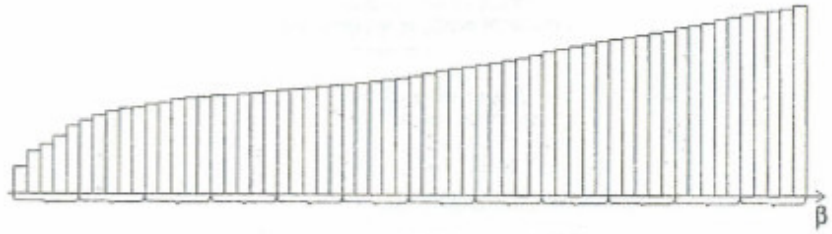


Рис. 4.21. Диаграмма Парето тест-билета из 60 заданий
с разбиением на 12 блоков по 5 заданий.
Заштрихованы задания подтест-билета T_1, состоящего из 12 заданий

4.8. Общая схема проектирования ПИМ

Выше были описаны различные алгоритмические процедуры проектирования тест-билетов. Рассмотрим теперь общую схему создания педагогических измерительных материалов. С технологической точки зрения, ее удобно представить в виде схемы (рис. 4.22).

Рассмотрим основные этапы создания ПИМ более подробно.

1. Постановка целей педагогических измерений. Перед началом разработки ПИМ необходимо четко определить, с какой целью проводится измерение, поскольку именно цель определяет содержательную и качественную сторону нижеследующих этапов. Целью педагогического измерения может быть:

- индивидуальная диагностика — оценка обученности отдельного учащегося;
- массовая диагностика — оценка обученности популяции учащихся (школьный класс, студенческая группа, образовательная организация в целом и т.п.);
- задача отбора (например, при приеме в учебное заведение или при переводе на следующий уровень обучения);
- задача селекции (например, определение победителей олимпиады или кандидатов на обучение по индивидуальной программе);
- задача оценки эффективности образовательной программы или методики обучения.

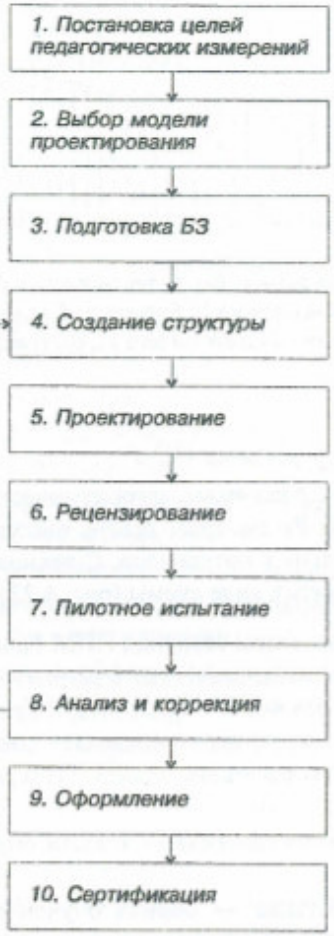


Рис. 4.22. Основные этапы создания ПИМ

2. Определение и выбор «подходящей» модели проектирования ПИМ.

Так, если цель педагогического измерения — оценка уровня обученности, то проектируемый ПИМ должен содержать задания (точнее — контролируемые учебные элементы) из наибольшего количества учебных тем. Статистическое распределение результатов измерений в этом случае близко к нормальному (рис. 4.23).

Рис. 4.23

Если цель измерений — отбор учащихся, то ПИМ должен содержать «норму-минимум», выполнение которой позволяет учащемуся, прошедшему отбор, успешно учиться дальше. Статистическое распределение результатов измерений в таком случае скорее всего будет иметь вид, близкий к представленному на рисунке. Иными словами, наибольшим будет число учащихся, выполнивших ПИМ на оценку «больше минимальной нормы». Информационная функция должна достигать максимума в районе границы «норма-минимум». Именно в этом диапазоне оценка должна быть наиболее точной, поскольку служит основой принятия решений. Если же форма гистограммы результатов измерений смещена влево, то это означает, что для данной категории учащихся ПИМ чрезмерно труден.

Рис. 4.24

Если цель измерений — селекционная, то ПИМ должен содержать задания с высоким уровнем усвоения ($\alpha>2$), ступенью абстракции ($\beta>2$) и степенью осознанности ($\gamma>2$). Поэтому лишь немногие выполнят достаточно большое число заданий. Максимальное значение статистического распределения в этом случае будет смещено влево.

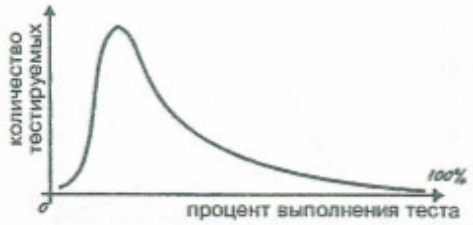


Рис. 4.25

Если целью измерений является оценивание образовательных программ, то в этом случае экономнее использовать выборочные методы (например, модульно-матричную модель).

3. Подготовка (выбор существующего или создание собственного) банка откалиброванных заданий. Наиболее трудоемкий (и, следовательно, дорогой) этап в процедуре разработки ПИМ. Созданию банка откалиброванных заданий посвящен разд. 3.

4. Создание структуры ПИМ. Определение валидности. Под структурой ПИМ понимают структурированный перечень названий файлов тестовых заданий с описывающими их характеристиками (и, в первую очередь, перечнем контролируемых учебных элементов). Обычно создание структуры производится вручную либо за экраном дисплея с помощью программного комплекса ТестГен:

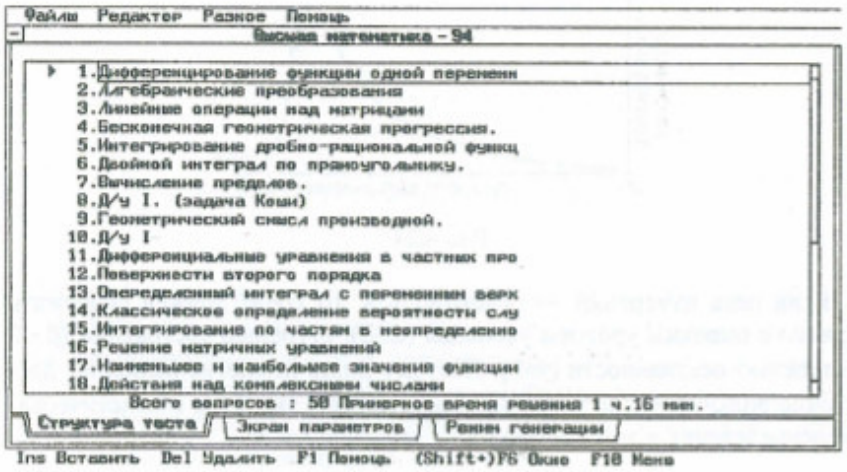


Рис. 4.26. Система ТестГен

После создания структуры можно экспертно определить содержательную валидность ПИМ. Напомним, что под валидностью понимается

соответствие между тем, что измеряется, и тем, что должно измеряться, т. е. степенью достижения ПИМ поставленных целей и отсутствием факторов, искажающих результаты измерения.

5. Проектирование ПИМ. Генерация необходимого количества вариантов после того, как определена структура производится проектирование ПИМ в соответствии с заданными целями.

6. Рецензирование и редактирование ПИМ являются следующим этапом проверки качества тестовых заданий и ПИМ в целом. Оно может проходить в форме простого обсуждения членами рабочей группы и в форме внешнего (независимого) рецензирования.

Рецензия внешних экспертов должна, как минимум, содержать оценку:

- соответствия ПИМ поставленным целям;
- правильности выделения контролируемых учебных элементов;
- технических качеств заданий: краткость, ясность, отсутствие двусмысленностей, искусственных сложностей, подсказок;
- единства стиля, единообразия формы и соответствия текстов заданий нормам языка;
- качества ПИМ в целом.

7. Пилотное испытание — важный этап в разработке ПИМ. Оно призвано определить его валидность и надежность, отбраковать неудачные задания. Пилотное (пробное) испытание проводится, как правило, на относительно небольших группах (хотя, с точки зрения накопления статистики, чем больше, тем лучше) для того, чтобы:

- определить первичные статистические характеристики заданий и ПИМ в целом;
- оценить возможность использования ПИМ для измерения тех характеристик, которые предусмотрены целями педагогического измерения;
- выполнить корректировку заданий на основе анализа результатов пробного испытания.

8. Анализ и коррекция. После пилотных испытаний может оказаться, что полученные результаты не (полностью) соответствуют поставленным целям. Может оказаться, что характеристики тестовых заданий, полученные на других выборках, не совпадают с полученными характеристиками. В этом случае необходима корректировка структуры ПИМ и набора заданий.

9. Оформление ПИМ включает в себя распечатку подготовленных вариантов, описания всех используемых характеристик, описания методики проведения педагогических измерений с использованием данного комплекса, методов оценивания заданий и методов интерпретации результатов измерений.

10. Сертификация ПИМ. Если предполагается, что ПИМ будет использоваться не только разработчиком в повседневной учительской практике, но и для нужд аттестации образовательных программ, то он должен удовлетворять ряду определенных (и достаточно жестких) требований, обеспечивающих возможность решать следующие задачи:

- проверять соответствие уровней обученности учащихся требованиям ГОС;
- обеспечивать возможность сравнения уровней обученности учащихся, проходящих подготовку по одинаковым образовательным программам в различных образовательных организациях.

Под сертификацией ПИМ будем понимать акт подтверждения (признания) того, что данный комплект ПИМ позволяет решать перечисленные задачи.

Поскольку отсутствует единая общепринятая методология разработки ПИМ, то проблема сертификации не может иметь однозначного решения. Один из возможных подходов — представление следующей информации, позволяющей экспертам принимать решение о пригодности данного комплекта ПИМ для решения задач аттестации:

- общие сведения (назначение, реквизиты разработчиков, инструкция по проведению, ...);
- характеристики ПИМ в целом;
- структура ПИМ и наборы заданий с описанными характеристиками (перечень контролируемых учебных элементов, расчеты параметров, характеризующих задания, ...);
- система оценивания заданий (ключи или критерии проверки);
- система толкования результатов измерений (шкала, формирование итоговой оценки, ...);
- результаты статистических испытаний (сведения о популяции тестируемых, на которой проводилась апробация; статистические результаты, представленные в стандартных статистических формах; расчет надежности теста);
- отзывы независимых экспертов; обоснование валидности.

Положительным моментом предложенного подхода является возможность формирования на ее основе Единого банка тестовых заданий (как централизованного, так и распределенного), позволяющего с помощью САПР ПИМ оперативно формировать качественные педагогические измерительные материалы.

Литература

1. Аванесов В.С. Математические модели педагогического измерения. — М.: Исследовательский Центр, 1994.
2. Аванесов В.С. Научные проблемы тестового контроля знаний. М.: Исследовательский Центр, 1994.
3. Анастази А. Психологическое тестирование. Т. 1, 2. — М.: Педагогика, 1982.
4. Беспалько В.П. Общая теория педагогических систем. — Воронеж, 1975.
5. Гласс Дж., Стенли Дж. Статистические методы в педагогике и психологии. — М., 1976.
6. Ингенкамп К. Педагогическая диагностика. М.: Педагогика, 1991.
7. Куклин В.Ж., Наводнов В.Г., Петропавловский М.В. КАМЕРТОН — технология проведения тестирования и анализа результатов. — Йошкар-Ола, 1995.
8. Куклин В.Ж., Мешалкин В.И., Наводнов В.Г., Савельев Б.А. О компьютерной технологии оценки качества знаний // Высшее образование в России. — 1993. — № 3. — С. 146–153.
9. О сертификации педагогических испытательных материалов / Куклин В.Ж., Масленников А.С., Наводнов В.Г., Савельев Б.А. // Квалиметрия человека и образование: методология и практика: Тез. докл. — М., 1996. — С. 134–135.
10. Куклин В.Ж., Масленников А.С., Наводнов В.Г. Технология разработки педагогических испытательных материалов // Прикладные исследования в электронике и новые технологии в обучении студентов. — Йошкар-Ола, 1996. — С. 44–46.
11. Майоров А.Н. Тесты школьных достижений: конструирование, проведение, использование. — Спб: Образование и культура, 1996.
12. TestGen — система формирования испытательных материалов / Ельцын А.В., Куклин В.Ж., Масленников А.С., Наводнов В.Г. — Йошкар-Ола, 1995.
13. Наводнов В.Г., Петропавловский М.В., Ельцын А.В. Автоматизированное проектирование педагогических измерительных материалов: Препринт № 2/97. — Йошкар-Ола, 1997.
14. Радионов Б.У., Татур А.О. Стандарты и тесты в образовании. — М., 1995.
15. Челышкова М.Б. Разработка педагогических тестов на основе современных математических моделей. — М.: Исследовательский Центр, 1995.

16. Челышкова М.Б., Савельев Б.А. Методические рекомендации по разработке педагогических тестов для комплексной оценки подготовленности студентов в вузе. — М., 1995.

17. Щербаков Э.Л. Оценка знаний. Эволюция и современное состояние. — Краснодар,1985.

18. Baker F.B. Item Responce Theory. Parameter Estimation Techniques. New York, 1992.

19. Binet A., Simon T.H., The Development of Intellegence in Young Children. Vineland, N-Y: The Training School, 1916.

20. Birnbaum A. Some Latent Trait Models. In Lord F.M., Novic M.R. Statistical Thoeries of Mental Test Scores. Reading, Mass.: Addison -Wesley, 1968.

21. Ebel R.L. The relation of item discrimination to the test reliability // Journal of Educational Measurement. — 1967, 4. — P. 125–128.

22. Fisher G.H., Molenaar I.W. (Eds) Rasch models. Foundations, Recent Developments and Applications. New York, Springer-Vorlag, 1995.

23. Glas C.A.W. The Rasch Model and Multistage Testing // Journal of Educational Statistics. — 1988, V13, N1. — P. 45–52.

24. Gruijter D.N.M., van der Kamp L. Item Theory. Statistical Models in Psychological and Educational Testing. Holland, 1984.

25. Gulliksen H., Theory of Mental Tests. N-Y: Wiley, 1950.

26. Guttman L. A basis for analyzing test-retest reliability. Psychometrika. — 1945, 10. — P. 255–282.

27. Guttman L. The quantification of a class of attributes: A theory and method of scale constraction. The prediction of personal adjustment. New York: Social Science Research Council. 1941. — P. 319–348.

28. Hambleton R.K., Swamimathan H. Item Responce Theory. Principles and Applications. Boston, 1985.

29. Kelly T.L. Selection of upper and lower groups for the validation of test items // Journal of Educational Psychology, 1939, 30. — P. 17–24.

30. Kuder G., Richardson M., The Theory of Test Reliability. Psychometrica, 2, 1937. — P. 151–160.

31. Lord F.M. A theory of test scores. Psychometrica Monograph No. 7, 17, 1952.

32. Lord F.M. Application of Item Response Theory to Practical Testing Problems. Hillsdale N - J.Lawrence Erlbaum Ass., Publ. 1980.

33. Lord F.M., Novic M.R. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison -Wesley, 1968.

34. Mislevy R.J., Stoching M. A Consumer's Guide to LOGIST and BILOG. Applied Psychological Measurement, 1989, 13. — P. 57–75.

35. Novick M.R. The Axioms and Principle Results of Classical Test Theory. Journal of Mathematical Psycholgy, 1966, N3. — P. 1–18.

36. Popham W.J. Criterion-Referenced Measurement. Englewood Cliffs (N.J.), 1978.

37. Rasch G. Probabilisyic Models for Some Intelligence and Attainment Tests. The Univ. Press. Chicago, 1980.

38. Samejima F.A. Weekly parallel test in latent trait theory with some criticisms of classical test theory. Psychometrica, 1977. V42. — P. 193–198.

39. Standards for Educational and Psychological Testing. American Psychological Association, 1985.

40. Thissen D., Steinberg L. A Taxonomy of Item Responce Models. Psichometrika, 1986. V. 51. — P. 567–577.

41. Thissen D., Wainer H. Some standard errors in item response theory. Psichometrika, 1982. V. 47. — P. 397–412.

42. Thorndike R.L. (Ed.) Educational Measurement (2nd ed.). Washington, D. C.,: American Council on Education, 1971.

43. Van der Linden W.J. Item banking met een dialoog gebaseerd op klassicke item- en testparameters. [Item Banking with a Dialogue Based on Classical Item and Test Parameters]. (Rapport 86-3, p. 1–25) Enschede, The Netherlands, University of Twente.

44. Wainer H., Brann H.J. (Eds) Test Validity. Hillsdale, New Jersay, 1988.

45. Wright B.D. Solving Measurement Problems with the Rasch Model. Journal of Educational Measurement, 1977, 14. — P. 97–116.

46. Wright B.D., Masters G.N. Rating Scale Analysis. Chicago: MESA Press, 1982.

47. Wright B.D., Panchapakesan N. A Procedure for Sample-free Item Analysis. Educational and Psychological Measurement, 1969, 29. — P. 23-48.

48. Wright B.D., Stone M.H. Best Test Design. Rasch Measurement. Chicago, 1979.

Владимир Григорьевич НАВОДНОВ

Математические модели САПР ПИМ

Препринт № 4/97

Редактор М.И. Шигаева
Компьютерный набор и верстка О.Г. Буркова

Лицензия № 020302 от 18.02.97.
Подписано в печать 25.02.97. Формат 60x84/16. Бумага тип. № 3.
Печать офсетная. Усл. печ. л. 4,2. Уч.-изд. л. 3,3.
Тираж 100 экз. Заказ № 1646. С-136.

Марийский государственный технический университет.
424024 Йошкар-Ола, пл. Ленина, 3

Отдел оперативной полиграфии
Марийского государственного технического университета.
424006 Йошкар-Ола, ул. Панфилова, 17