Применение современной теории тестирования IRT в системе контроля измерительных свойств диагностических материалов

Белобородов Владимир Николаевич — кандидат физико-математических наук, доцент кафедры общей физики НИЯУ МИФИ, заведующий отделом стандартизации инструментария диагностики МЦКО, v-belob@mail.ru

Татур Александр Олегович — кандидат физико-математических наук, начальник отдела развития инструментария оценки качества образования ГАОУ ДПО МЦКО, главный научный консультант ФГБНУ «ФИПИ», tatur@bk.ru

Ключевые слова: вероятность, способность испытуемого, трудность задания, характеристическая функция, информационная функция, ошибка определения способности, якорные задания.

С целью повышения качества диагностических материалов Московский центр качества образования проводит непрерывный контроль качества вариантов и заданий. На этапе подготовки комплектов вариантов диагностические материалы проходят экспертизу специалистами-предметниками и тестологами. При сборке вариантов диагностических материалов из отдельных заданий учитываются статистические характеристики этих заданий.

После проведения диагностических мероприятий с количеством участников несколько тысяч человек проводится математическая обработка полученных результатов с целью определения статистических параметров, как вариантов, так и отдельных заданий. Используется параллельно два подхода для анализа статистических параметров.

Во-первых, извлекаются параметры обработки в рамках классической теории тестов[1]. В парадигме этой теории основными параметрами для заданий являются: процент выполнения задания, дискриминативность (дифференцирующая способность) задания, коэффициенты корреляции исходов выполнения задания и набранного балла за вариант. В качестве основных параметров вариантов рассматриваются: средний процент выполнения заданий (средний набранный первичный балл), стандартное отклонение балла (дисперсия), асим-

[1] Челышкова М.Б. Теория и практика конструирования педагогических тестов. — М.: «Логос», 2002. — 432 с.

метрия, эксцесс. В качестве меры самосогласованности вариантов используются коэффициенты надежности.

Во-вторых, определяются параметры обработки в рамках современной теории тестирования[2]. В рамках данной теории способность (уровень подготовки) испытуемого и уровень трудности задания определяются на одной шкале (интервальной). Взаимная увязка начал отсчета шкал является в IRT проблемой даже в рамках одного комплекта априори параллельных вариантов. Такая увязка делается внутри комплекта введением в варианты общих заданий. Если при создании вариантов используются задания из базы, то есть с известными параметрами IRT, то их обычно называют якорными. В IRT объединены различные модели. В том числе могут применяться непараметрические модели. В МЦКО используется вариант модели, учитывающий политомические задания, то есть многобалльные задания (Partial Credit Model — модель с промежуточными ступенями выполнения заданий)[3].

Анализ заданий в современной теории тестирования

В основе большого семейства моделей IRT лежит простейшая модель Раша, которая связывает вероятность правильного выполнения дихотомического (однобалльного) задания с параметрами трудности задания и способности испытуемого его выполнить[4]

$$P = \frac{1}{1+\exp(\beta-\theta)}. \qquad (1)$$

Здесь β — трудность задания, а θ — способность испытуемого его выполнить. При неограниченном убывании способности θ эта величина стремится к нулю. А при неограниченном росте способности θ вероятность P стремится к единице. При этом, как следует из формулы (1), конкретное значение вероятности определяется разностью способности и трудности задания. Эти величины определены на интервальной шкале действительных (как положительных, так и отрицательных) чисел. Это означает, что начало отсчёта этих величин не фиксировано. Для получения метрической шкалы, то есть шкалы отношений, следует увязывать начала отсчета для разных вариантов. Это делается с использованием общих или якорных заданий.

Следует отметить, что латентный параметр трудности задания β в моделях IRT не является синонимом или какой-либо однозначной функцией трудности задания в классической теории тестов, в которой эта величина, как правило, определяется как процент выполнения задания. Точно так же латентный параметр способность θ не имеет взаимнооднозначного соответствия с классическими параметрами успешного выполнения заданий тестов, такими как набранный первичный балл. Следует отметить, что модель Раша априори в большей мере подходит к заданиям с конструируемым ответом, чем к заданиям с выбором ответа. В заданиях с выбором ответа возможно угадывание. Его признаком можно считать то, что при неограниченном убывании способности θ вероятность P стремится к конечной величине, а не к нулю.

Вероятность (1) называется характеристической функцией задания, если рассматривается в зависимости от способности θ. Она изображена на рис. 1 для первого задания из диагностического варианта по физике, использованного в 2016 году.

Как следует из формулы (1), вероятность верного выполнения задания равна 0,5, если трудность задания равна способности: $\beta = \theta$. Это позволяет по характеристической функ-

---

[2] Rasch G. Probabalistic Models for Some Intelligence and Attainment Tests. — Chicago: Mesa Press, 1980. — 199 p. Карданова Е.Ю. Моделирование и параметризация тестов: основы теории и приложения. — М.: Федеральный центр тестирования, 2008. — 296 с.

[3] Masters G.N. A Rasch model for partial credit scoring // Psychometrika. — 47. — P. 149–174.; Линда Крокер, Джеймс Алгина. Введение в классическую и современную теорию тестов. — М.: Логос. — 2010. — 668 с.

[4] Rasch G. Probabalistic Models for Some Intelligence and Attainment Tests. — Chicago: Mesa Press, 1980. — 199 p.

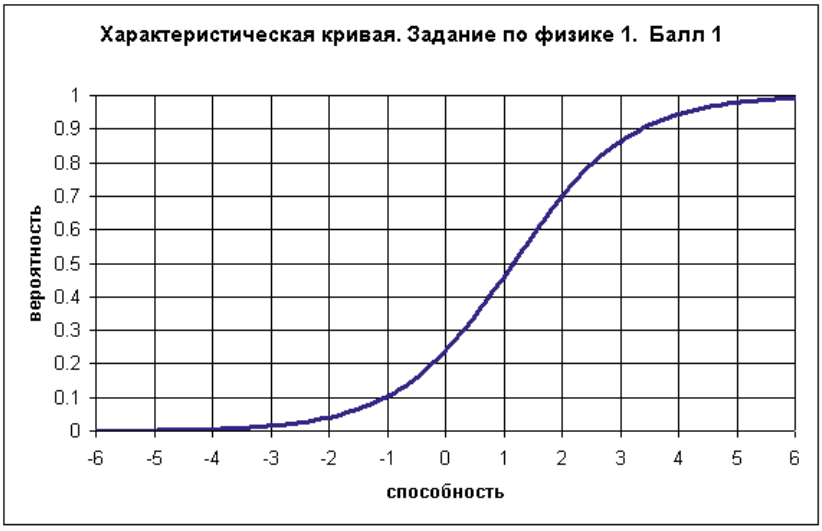


Рис. 1. Вероятность верного ответа в зависимости от способности в модели Раша для дихотомического задания 1

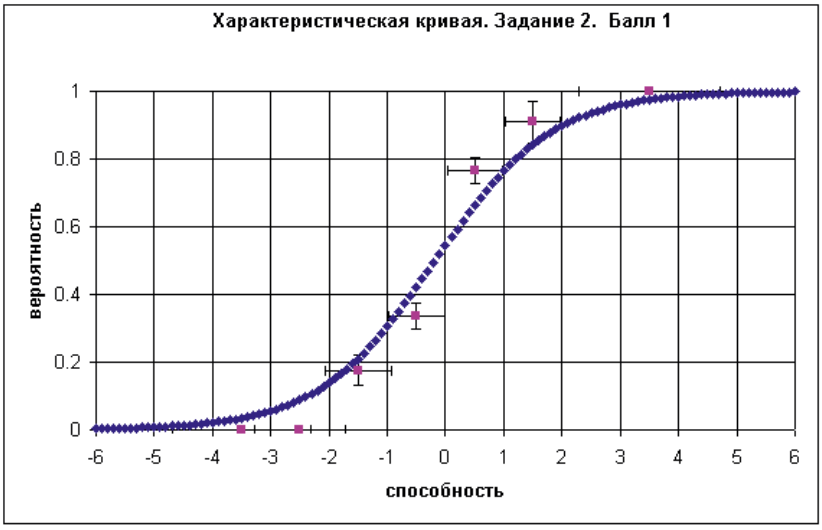


Рис. 2. Характеристическая кривая однобалльного (дихотомического) задания 2

ции определить трудность задания. Для рассматриваемого задания она, как видно из рис. 1, приблизительно равна 1,2.

В модели IRT на уровне одного задания удаётся отследить соответствие используемой модели полученным результатам.

Реальный интерес представляет сравнение теоретических характеристических кривых и экспериментальных значений. Чтобы определить соответствующие экспериментальные вероятности, следует сгруппировать испытуемых на нескольких интервалах, так чтобы образовалось несколько групп (12), для которых будут вычислены соответствующие частоты и доли испытуемых, выполнивших задание правильно.

Характеристическая кривая задания, на которой отображаются экспериментальные значения, изображена на рис. 2. Погрешности определяются разбросами способностей и полученных баллов на 12 интервалах.

Рис. 3. Характеристическая кривая двухбалльного задания 3 для балла 1

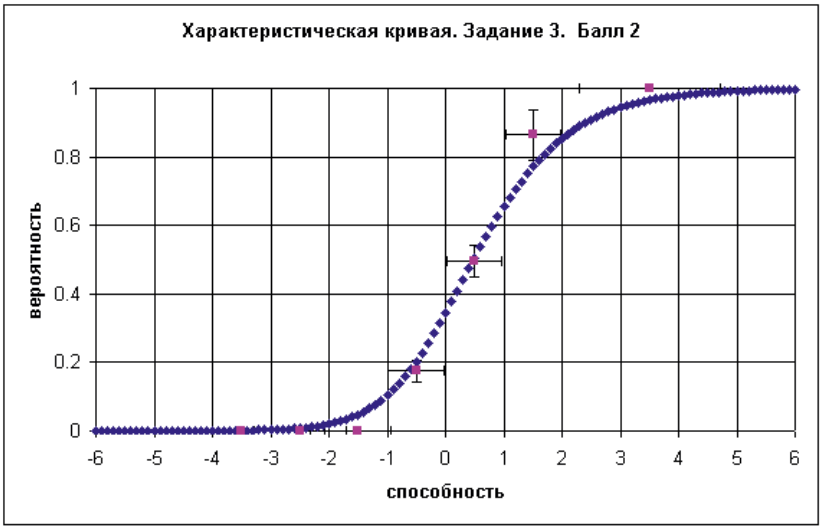


Рис. 4. Характеристическая кривая двухбалльного задания 3 для балла 2

В пределах погрешностей экспериментальные и теоретические значении для задания 2 совпадают за исключением точек нулевого и стопроцентного выполнения задания (10 и 3 испытуемых соответственно из 378 человек, выполнявших вариант) (см. рис. 2).

Модель Partial Credit позволяет анализировать результаты и для политомических (многобалльных) заданий. Причём характеристические кривые получаются для отдельных шагов выполнения заданий (набранных баллов). В случае политомического задания формула (1) определяет условную вероятность правильного выполнения шага в задании, если предыдущие шаги сделаны правильно.

На рис. 3 изображена зависимость вероятности получения единичного балла в духбалльном задании. Видно, что в отличие от однобалльного задания эта вероятность сначала растёт, а потом убывает. Низка вероятность получения балла 1 как в группе самых слабых испытуемых, так и в группе самых сильных испытуемых. Первые в основном получают 0 баллов, а последние — 2 балла, то есть максимальный балл в данном задании.

В задании 3 зависимость процента набравших 1 балл и 2 балла от способности в пределах погрешностей совпадает с теорией в большинстве точек (рис. 3 и рис. 4). Превышение теоретических значений наблюдается для низких значений способностей при балле 1. Чтобы понять причину этого, необходимо рассмотреть само задание.

Текст задания 3

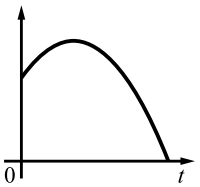
Мячик бросают с начальной скоростью $\vec{\upsilon}_0$ под углом α к горизонту с балкона высотой h (см. рисунок). Сопротивлением воздуха пренебречь. Графики А и Б представляют собой зависимости физических величин, характеризующих движение мячика в процессе полёта, от времени t.

Установите соответствие между графиками и физическими величинами, зависимости которых от времени эти графики могут представлять.

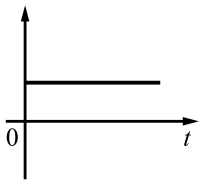
К каждой позиции первого столбца подберите соответствующую позицию второго и запишите в таблицу выбранные цифры под соответствующими буквами.

ГРАФИКИ

А)



Б)



ФИЗИЧЕСКИЕ ВЕЛИЧИНЫ

1) координата x мячика
2) проекция скорости мячика на ось x
3) кинетическая энергия мячика
4) координата y мячика

Ответ:

А	Б

Видно, что задание имеет конечное число ответов. Формально имеется 25 различных ответов, включая пустые, полных из них 16. Верный ответ один (42). Вероятность дать полный неправильный ответ равна $(3/4)^2 = 9/16$. Таким образом, вероятность угадать правильный или частично правильный ответ будет равна $1–9/16 = 7/16$. А вероятность дать именно частично правильный ответ равна $P = 7/16–1/16=6/16=3/8=0{,}375$. С этой вероятностью гадающие испытуемые получат за это задание 1 балл. Как видно из рис. 3, приблизительно такая вероятность наблюдается при способностях в окрестностях значений $\theta \approx –2{,}5$ и $\theta \approx –1{,}5$.

Степень соответствия заданий модели IRT контролируется с помощью величины хи-квадрат[5].

Анализ вариантов диагностических материалов в современной теории тестирования

Для анализа вариантов из набора заданий в IRT может быть проведено соотнесение распределения трудностей заданий (их шагов) и распределения испытуемых на одной шкале.

На диаграмме на рис. 5 в средней части расположена вертикальная шкала с делениями через 0,25. Левее шкалы указаны номера заданий (с шагами их выполнения в скобках) на уровне, соответствующем их трудности. Правее шкалы расположена столбиковая диаграмма распределения испытуемых по способностям. Видно, что основная масса заданий (первых шагов их выполнения) расположена вблизи локальных максимумов в распределении испытуемых. В область высоких значений попал первый шаг задания 15, которое

[5] Карданова Е.Ю. Моделирование и параметризация тестов: основы теории и приложения. М.: Федеральный центр тестирования, 2008, 296 с.

Таблица 1

Распределение испытуемых и заданий на шкале способностей-трудностей для варианта 2 по физике

Способность-трудность	Количество испытуемых	Задание (шаг)
-5.125	1	
-3.625	2	
-2.875	2	
-2.625	0	12(1)
-2.375	9	10(1)
-1.875	16	
-1.625	23	1, 9(1)
-1.375	27	
-1.125	40	3, 11(2)
-0.875	33	5
-0.625	34	8(1), 9(2)
-0.375	31	10(2)
-0.125	39	2, 4, 6, 12(2)
0.125	28	7
0.375	24	14, 15(2)
0.625	50	
0.875	18	
1.125	13	13
1.375	1	
1.875	2	
2.125	2	
2.875	3	
3.375	0	15(1)
3.875	1	

оказалось в целом трудным для испытуемых. В таблице 1 помещены численные значения данных для диаграммы на рис. 5.

Модель IRT предполагает наличие локальной независимости вероятности выполнения различных заданий с известными параметрами трудности при заданной способности. Это означает, что вероятность определяется на основе одномерного распределения способностей и трудностей. Из этого также следует, что сумма математических ожиданий баллов по заданиям варианта даст среднее значение балла, который будет получен испытуемым за вариант при данной способности. Такая сумма называется характеристической функцией варианта. Она равна математическому ожиданию набираемого испытуемым балла при определенной способности. Если характеристическую функцию варианта (теста) разделить на максимальный балл за тест, то получится нормированная на единицу характеристическая функция. Формально это вероятность правильного выполнения варианта, но реально эта величина равна доле верно выполненных заданий (доле полученного балла от максимально возможного при наличии в вариантах политомических заданий). На рис. 6 изображена характеристическая функция варианта 1 по физике.

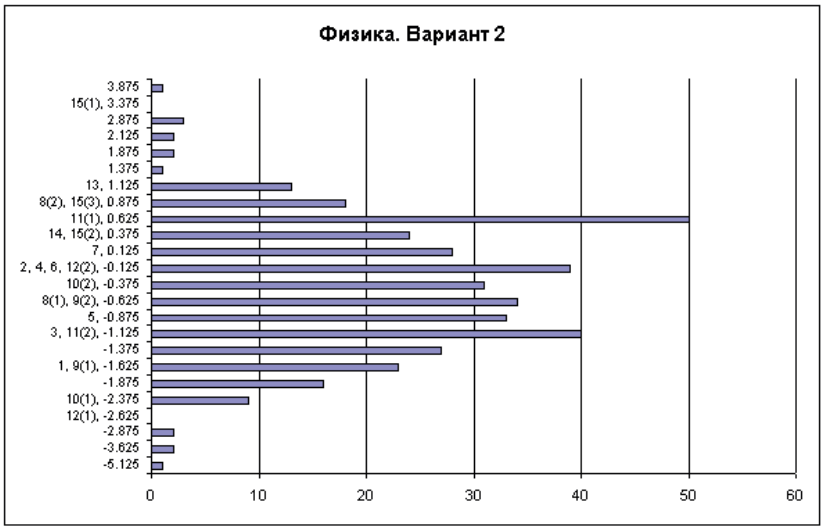


Рис. 5. Диаграмма расположения заданий и испытуемых на шкале способностей-трудностей

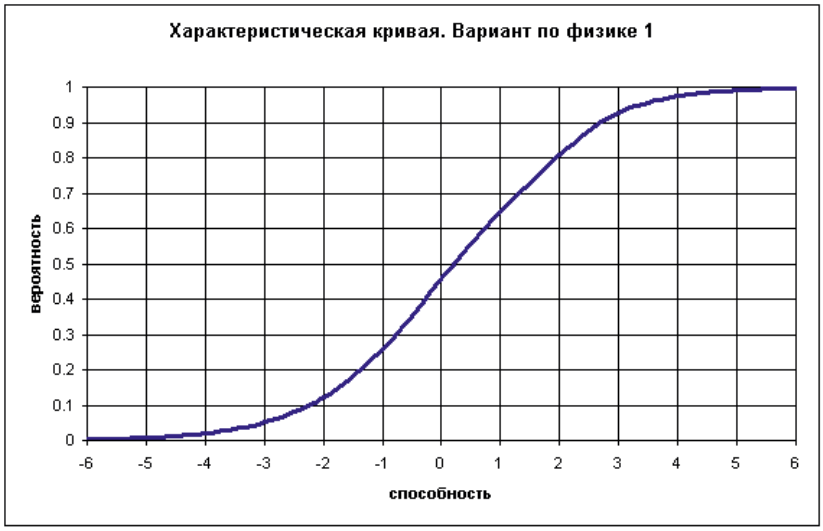


Рис. 6. Характеристическая кривая варианта 1

Из рис. 6 следует, что при способности $\theta = 0{,}3$ ожидается выполнение варианта приблизительно наполовину ($P = 0{,}5$), то есть набранный первичный балл будет приблизительно равен половине от максимально возможного.

Теоретическую характеристическую кривую, изображенную на рис. 6, целесообразно дополнить экспериментальными данными с соответствующими доверительными интервалами.

На рис. 7 изображены три величины. Теоретические и экспериментальные значения вероятности выполнения варианта в виде отношения получаемого балла к максимальному баллу (точки с областями погрешностей), а также распределение испытуемых по способностям (треугольники). Квазинепрерывная кривая — теоретическая характеристическая функция. Погрешности на теоретической кривой определяются свойствами используемой модели IRT[6]. Экспериментальные и теоретические значения совпадают в пределах погрешностей. Из рис. 7 видно, что рабочей частью шкалы способностей приблизительно

[6] Белобородов В.Н. Надежность тестов. — М.: НИЯУ МИФИ. — 2012, 36 с.

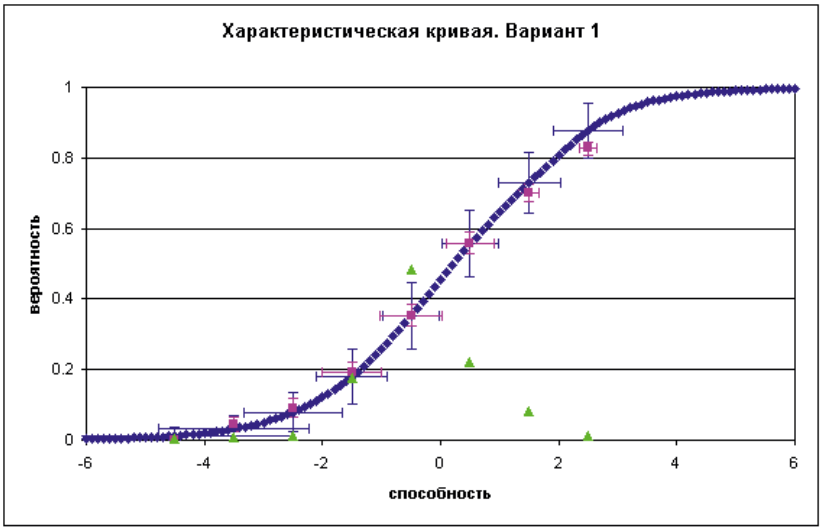


Рис. 7. Характеристическая кривая варианта 1 с экспериментальными точками и распределением испытуемых по способностям

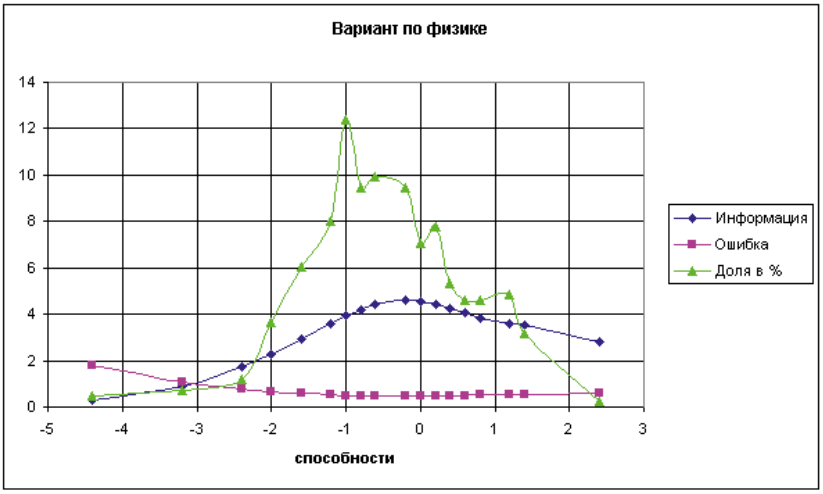


Рис. 8. Информационная функция (Информация), ошибка определения способности (Ошибка) и распределение испытуемых по способностям (Доля в %)

является интервал от минус 3 до плюс 3. Такой интервал соответствует рабочему интервалу разности способности и трудности для одного задания.

Графики информации, ошибки определения способности и доли испытуемых в соответствующей области способностей позволяют более детально соотнести распределение испытуемых по способностям с точностью определения способностей. Погрешность определения способности в IRT равна обратной величине от квадратного корня из информационной функции в данной точке[7], если значение информационной функции превышает 1. То есть чем выше информационная функция в данной точке, тем выше точность определения способности.

[7] Белобородов В.Н. Надежность тестов. — М.: НИЯУ МИФИ, 2012. — 36 с. Белобородов В.Н., Татур А.О. Измерения в физике и оценка уровня освоения её содержания. — М.: Физическое образование в вузах. — 2010, т. 16. — № 2. — с.83–94.

Рис. 9. Распределение по набранным первичным баллам

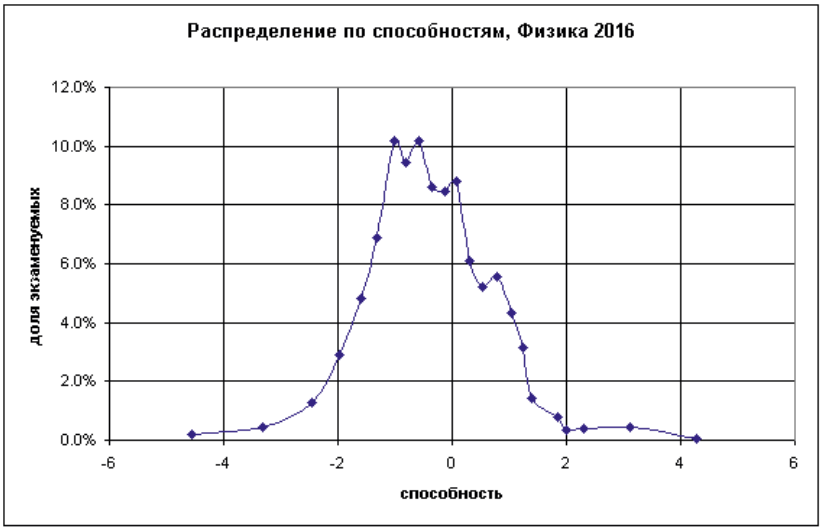


Рис. 10. Распределение по способностям

Из рис. 8 видно, что максимум информационной функции и минимум ошибки приблизительно лежат в области максимума распределения испытуемых по способностям.

Анализ комплектов вариантов диагностических материалов в современной теории тестирования

По комплекту вариантов могут быть проанализированы распределения по набранным первичным баллам и способностям (рис. 9 и рис. 10).

Общий вид этих распределений одинаков. Но есть отличия. Шкала первичных баллов конечная, а способностей — бесконечная. Как видно из рис. 10, испытуемые с представительностью выше 1% имеют способности в интервале приблизительно от минус 3 до плюс 2.

В IRT также имеется возможность анализировать зависимость средней ошибки определения способности от результата выполнения варианта (рис. 11). На средней части шкалы процентов выполнения (от 30% до 80%) эта погрешность не превосходит 0,5.

Ошибку определения способности в IRT следует сравнить с разбросом способностей по вариантам комплекта в зависимости от набранного балла.

Рис. 11. Ошибка (погрешность) определения способности

Рис. 12. Разброс способностей по вариантам (стандартное отклонение)

Видно (см. рис. 12), что разброс способностей по вариантам меньше ошибки определения способностей в несколько раз. Это означает, что выравнивание шкал по разным вариантам было бы превышением точности, которая обеспечивается используемой моделью IRT.

Определение измерительной эквивалентности вариантов может быть реализовано в IRT сравнением средних способностей по вариантам. Средние способности по вариантам комплекта по физике внесены в таблицу 2.

Также эту информацию можно проанализировать на диаграмме (рис. 13).

На рис. 13 видно, что разброс средних способностей по вариантам лежит приблизительно в доверительном интервале величины в две ошибки среднего (0,06). Вне этого интервала имеется различие только для вариантов 2 и 3 (см. табл. 2).

Степень параллельности вариантов может быть проанализирована на основе зависимостей способностей от процентов выполнения заданий вариантов (см. рис. 14). В модели IRT Partial Credit достаточной статистикой для определения способности является балл

Таблица 2

ВАРИАНТ	Количество испытуемых	Средняя способность	Ошибка средней способности
1	412	-0.32	0.03
2	399	-0.28	0.03
3	378	-0.35	0.03
4	364	-0.30	0.03
комплект	1553	-0.31	0.03

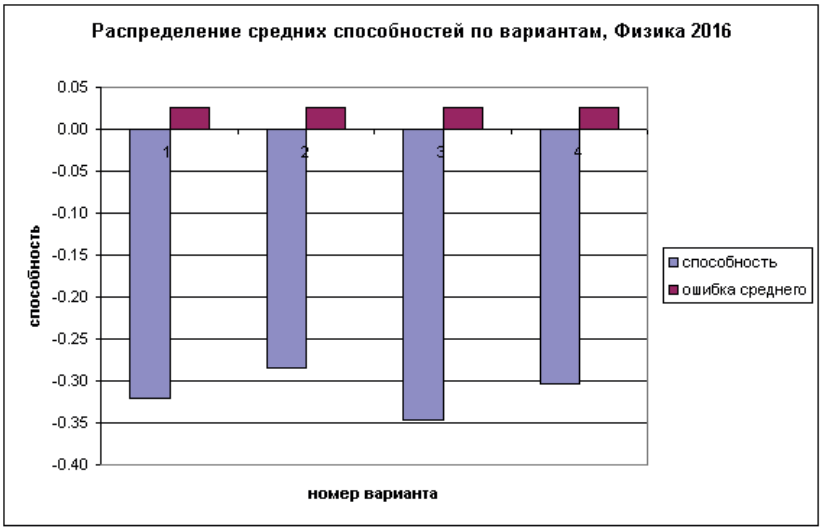


Рис. 13. Распределение средних способностей по вариантам и ошибки определения средних значений

Таблица 3

Способности в зависимости от балла за вариант

Балл	Способности по вариантам				Среднее	Стандартное отклонение
	1	2	3	4		
0	-4.40	-4.91			-4.65	0.36
1	-3.13	-3.62	-3.11	-3.45	-3.33	0.25
2	-2.35	-2.80	-2.32	-2.67	-2.53	0.24
3	-1.86	-2.28	-1.83	-2.18	-2.04	0.23
4	-1.49	-1.88	-1.45	-1.81	-1.66	0.22
5	-1.18	-1.54	-1.15	-1.49	-1.34	0.21
6	-0.91	-1.25	-0.87	-1.22	-1.07	0.20
7	-0.67	-0.99	-0.63	-0.97	-0.81	0.19
8	-0.45	-0.74	-0.40	-0.73	-0.58	0.18
9	-0.23	-0.51	-0.17	-0.50	-0.35	0.18
10	-0.01	-0.28	0.05	-0.27	-0.13	0.17
11	0.21	-0.05	0.27	-0.05	0.09	0.17
12	0.44	0.17	0.49	0.18	0.32	0.17

Окончание табл. 3

Балл	Способности по вариантам				Среднее	Стандартное отклонение
	1	2	3	4		
13	0.68	0.41	0.71	0.41	0.55	0.17
14	0.94	0.65	0.94	0.64	0.79	0.17
15	1.21	0.90	1.16	0.87	1.04	0.17
16	1.49	1.15	1.37	1.10	1.28	0.19
17		1.39	1.59	1.34	1.44	0.13
18	2.07	1.65	1.83	1.59	1.78	0.22
19	2.39	1.94		1.88	2.07	0.28
20		2.33		2.28	2.30	0.04
21		3.03	3.25	2.96	3.08	0.15
22		4.27			4.27	0

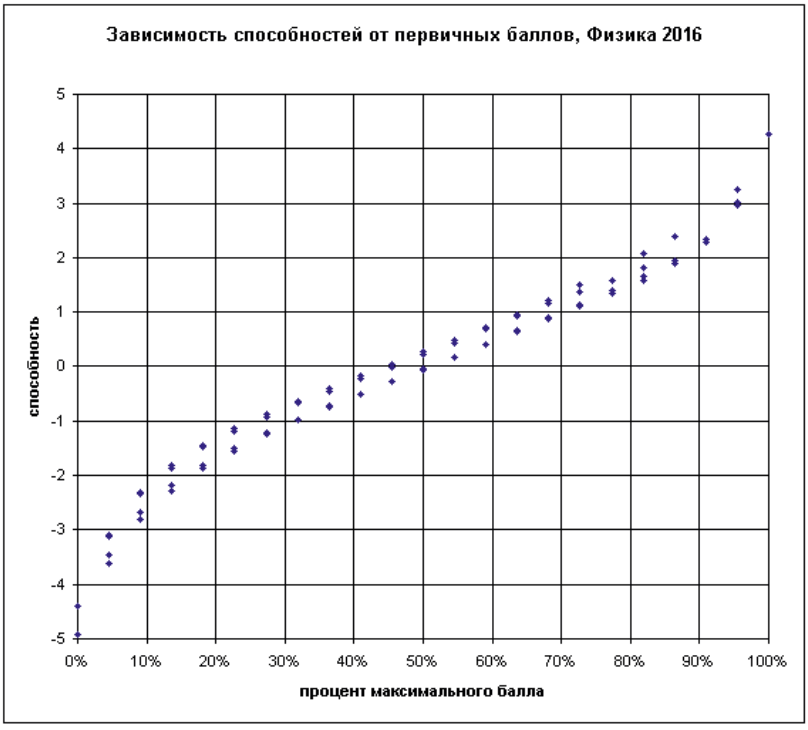


Рис. 14. Зависимости способностей от процентов выполнения заданий комплекта вариантов

за вариант. Для нескольких вариантов имеется некоторый разброс способностей при заданном балле.

Пустые клеточки в таблице 3 соответствуют отсутствию результатов с данным баллом в соответствующем варианте.

Из рис. 14 видно, что разбросы способностей при одинаковых процентах выполнения вариантов не превосходят 1, что соответствует ошибке измерения 0,5 (см. рис. 11).

Средние способности в пределах погрешностей (см. табл. 2) практически совпадают, а средние трудности вариантов (см. табл. 4) всё-таки отличаются на величину, превосходящую доверительный интервал по способностям 0,24-(–012)=0,36>0,06). Варианты 1 и 3 сложнее, чем варианты 2 и 4.

Таблица 4

Средние баллы, трудности и способности по вариантам

Вариант	Количество испытуемых	Средний балл	Средняя трудность β	Средняя способность θ
1	412	8.74	0.24	–0.32
2	399	10.09	–0.12	–0.28
3	378	8.45	0.20	–0.35
4	364	9.95	–0.11	–0.30

Таблица 5

Общие задания комплекта вариантов по физике

Вариант	Номер задания	Задание с типом	Процент выполнения
1	3	B1	48±2%
3	3	B1	48±3%
1	7	A3	41±2%
4	7	A3	42±2%
2	4	B2	38±2%
4	4	B2	38±2%
2	8	A4	50±3%
3	8	A4	42±3%

Так как варианты связаны четырьмя парами общих заданий, способности испытуемых и трудности заданий отложены на одной шкале, а не на четырех шкалах с различными началами отсчета (Таблица 5).

В правой колонке табл. 5 процент выполнения задания указан с погрешностью, которая определяется объёмом выборки испытуемых.

То, что средний уровень способностей испытуемых, выполнявших задания варианта 3, несколько ниже, чем средние уровни способностей испытуемых, выполнявших другие варианты, подтверждается тем, что процент выполнения общего задания № 8 (A4) в варианте 3 (42±3%) ниже, чем в варианте 2 (50±3%).

Заключение

Современная теория тестирования позволяет проводить анализ качества отдельных заданий, вариантов диагностических материалов и их комплектов. Получаемые на основе IRT данные могут быть представлены для анализа как в графическом, так и в табличном виде. Численные индикаторы в виде хи-квадрат позволяют выделить проблемы, как в выполнении заданий, так и в самих заданиях. Возможность вычислять в IRT ошибки исследуемых параметров намного упрощает задачу интерпретации результатов применения диагностических материалов.

Совместное исследование результатов выполнения различных вариантов диагностических материалов требует либо использования общих, либо якорных заданий, с помощью которых производится увязывание интервальных шкал различных вариантов. Для целей создания банка калиброванных заданий это является обязательным условием. Иначе в IRT анализ вариантов может рассматриваться только по отдельности. Но даже в этом случае анализ качества заданий и вариантов в рамках IRT остается вполне информативным и целесообразным. Связывание результатов по годам представляется возможным только с применением связывающих заданий из банка заданий.