РЪКОВОДСТВО ЗА СЪСТАВЯНЕ НА ТЕСТОВЕ*

Фернандо Картрайт
Джери Мусио[1)]

Резюме. Образованието се превърна в основен фактор, определящ социалното и икономическото развитие на индивидите и нациите. Затова реформата на образователната система е обикновено сред ключовите национални приоритети на правителствата.

Националните програми за външно оценяване на учениците са основният инструмент за измерване на качеството на образованието и за текущо наблюдение на ефекта от образователните реформи. Това налага да бъде гарантирано, че разработените тестове и изпитни програми са на най-високо професионално равнище.

Целта на настоящото ръководство е да подпомогне всеки екип, отговорен за разработването и провеждането на тестове, както и при анализа и обявяването на резултатите. Текстът е фокусиран върху практическото приложение на теорията за разработване на тестове въз основа на опит от провеждането на различни програми за оценяване в Канада и в други държави.

Главните теми, които са разгледани, са следните: планиране на оценяването, разработване на тест, създаване на тестови задачи, съставяне на тестове и провеждане на пилотни тестове, анализ на тестовете и тестовите задачи и интерпретиране на тестовите резултати.

Keywords: education, assessment, item development, test design, testing

5. Анализ на тестовете и тестовите задачи

Анализът на задачите и тестовете представлява специфичен анализ на данни и затова трябва да се спазват общите принципи за управление и анализ на данни.

5.1. Форматиране на данни

Съществуват различни широко разпространени таблични формати за създаване на бази данни: ACCESS, SQL, EXCEL, SPSS и текстови файлове с определени граници. Ако информацията е поместена в SPSS файл, то данните

* Продължение от книжка 1 и 2, 2013 на сп. „ Стратегии на образователната и научната политика“

са вече правилно форматирани. Ако данните не са подредени в правилната структура, софтуерът няма да може да осъществи анализа. За да е сигурно, че данните могат да бъдат анализирани правилно, структурата им трябва да отговаря на определени характеристики и принципи за форматиране.

1. Използването на pivot таблици е за предпочитане пред стандартните. Ако данните са в база данни, тогава нужната информация може да бъде извлечена чрез командата PIVOT, вградена в повечето системи за обработка на бази данни.

Неправилно

Променлива	**Ученик**	**Оценка**
Променлива 1	Ученик 1	A
Променлива 2	Ученик 1	B
Променлива 1	Ученик 2	C
Променлива 2	Ученик 2	D
Променлива 1	Ученик 3	E
Променлива 2	Ученик 3	F

Правилно

Ученик	**Променлива 1**	**Променлива 2**
Ученик 1	A	B
Ученик 2	C	D
Ученик 3	E	F

2. Данните с отговорите на учениците се подреждат така, че променливите са по колони, а учениците – по редове.

Неправилно

Променлива	**Ученик 1**	**Ученик 2**	**Ученик 3**
Променлива 1	A	C	E
Променлива 2	B	D	F

Правилно

Ученик	Променлива 1	Променлива 2
Ученик 1	A	B
Ученик 2	C	D
Ученик 3	E	F

3. Данните за параметрите на задачите се подреждат така, че параметрите са по колони, а задачите – по редове.

Неправилно

Параметър	Задача 1	Задача 2	Задача 3
Параметър 1	A	C	E
Параметър 2	B	D	F

Правилно

Задача	Параметър 1	Параметър 2
Задача 1	A	B
Задача 2	C	D
Задача 3	E	F

4. Наименованията на променливите се изписват в най-горната част на колоната.

Неправилно

Ученик 1	A	B
Ученик 2	C	D
Ученик 3	E	F

Правилно

Ученик	Променлива 1	Променлива 2
Ученик 1	A	B
Ученик 2	C	D
Ученик 3	E	F

5. Наименованията на променливите започват с букви, а не с цифри и не съдържат специални символи като &. -, /, $,#,@,%.

Неправилно

Ученик	**1Променлива**	**Променлива#2**	**Променлива3**
Ученик 1	A	B	C
Ученик 2	D	E	F
Ученик 3	G	H	I

Правилно

Ученик	**Променлива 1**	**Променлива 2**	**Променлива 3**
Ученик 1	A	B	C
Ученик 2	D	E	F
Ученик 3	G	H	I

6. Празните редове и колони вляво, над и в самата таблица се премахват.

Неправилно

	Ученик	Променлива 1		Променлива 2	Променлива 3
	Ученик 1	A		B	C
	Ученик 2	D		E	F
	Ученик 3	G		H	I

Правилно

Ученик	**Променлива 1**	**Променлива 2**	**Променлива 3**
Ученик 1	A	B	C
Ученик 2	D	E	F
Ученик 3	G	H	I

7. Различните видове информация се поместват в различни видове таблици.

Неправилно

Отговори			
Ученик	Задача 1	Задача 2	Задача 3
Ученик 1	A	B	C
Ученик 2	D	E	F
Ученик 3	G	H	I
Характеристики на задачите			
Параметър 1	J	K	L
Параметър 2	M	N	O
Параметър 3	P	Q	R

Правилно

Таблица 1. Отговори

Ученик	**Задача 1**	**Задача 2**	**Задача 3**
Ученик 1	A	B	C
Ученик 2	D	E	F
Ученик 3	G	H	I

Таблица 2. Характеристики на задачите

Задача	**Параметър 1**	**Параметър 2**	**Параметър 3**
Задача 1	J	M	P
Задача 2	K	N	Q
Задача 3	L	O	R

Забележка: имената на таблиците с данни НЕ се виждат в таблиците.

8. Некоректните и пропуснатите отговори за всяка задача се отбелязват винаги с едни и същи кодове.

Неправилно

Ученик	Задача 1	Задача 2
Ученик 1	8	7
Ученик 2	C	.
Ученик 3	9	D

Забележка: 7, 8 = некоректен отговор, оценен като грешен;
9, . = пропуснат отговор, НЕ оценен като грешен.

Правилно

Ученик	**Задача 1**	**Задача 2**
Ученик 1	8	8
Ученик 2	C	9
Ученик 3	9	D

Забележка: 8 = некоректен отговор, оценен като грешен;
9 = пропуснат отговор, НЕ оценен като грешен.

5.2. Статистическа обработка на задачите чрез класическата теория на тестовете

Класическата теория на тестовете описва ефективността на задачите в рамките на определена извадка и определен тест. При пилотните тестове класическата теория е полезна, макар извадката от задачи и извадката от ученици да не са представителни за финалния тест и за учениците като цяло. Класическата теория се нуждае от няколко степени на свобода, за да предостави надеждни резултати. В контекста на разработването на задачи и тестове са особено полезни следните три показателя:

– трудност на задачата
– способност на задачата да разграничава (дискриминативна сила)
– анализ на дистракторите.

5.2.1. Трудност на задача

Трудността (или процентът верни отговори) на задачата описва частта на отговорилите вярно. По принцип най-добри са тестовите задачи с трудност около 0,50 (тоест решени вярно от 50% от учениците), но е почти невъзможно да се съставят задачи само с такава трудност. Задачите с трудност от 0,40 до 0,80 като цяло предоставят добра информация за разграничаване на учениците.

Тъй като съществува вероятност учениците случайно да посочат верния отговор при задачи с избираем отговор, идеалната трудност всъщност е по-висока от 0,50. При съставянето на задачи с избираем отговор се използват следните препоръчителни граници за трудността:

Тип задача	Препоръчителна трудност на задачата
Задача с 5 възможни отговора	0,40 – 0,75
Задача с 4 възможни отговора	0,45 – 0,80
Задача с 3 възможни отговора	0,50 – 0,85
Задача с 2 възможни отговора (вярно/невярно)	0,55 – 0,90

Тези правила не са абсолютни. Някои задачи могат да покриват важни дялове от учебната програма или да имат за цел да окуражат учениците. Тези задачи трябва да останат, въпреки че не попадат в границите на трудност, определени за всеки тип задача.

5.2.2. Способност на задачата да разграничава (дискриминативна сила)

Дискриминативната сила се отнася за способността на задачата да предизвиква различни отговори от ученици с различни способности. Ако всички ученици посочват един и същи отговор на една задача, независимо от своите способности, умения и ниво на успеваемост, тази задача не би била полезна – поради неспособността си да разграничи учениците според нивото на способностите им. Задачи с висока дискриминативна сила са ценени, тъй като доказват, че ученици с висок общ резултат на теста като цяло успяват правилно да разберат задачата, докато онези с нисък общ резултат най-често не се справят с тази задача. Класическата дискриминативна сила на задача се изчислява като разлика между броя верни отговори на учениците с висок общ резултат

минус броя верни отговори на учениците с нисък общ резултат. Класическата дискриминативна сила трябва да е над 0,25.

Има много причини, поради които една задача може да има ниска или дори отрицателна дискриминативна сила. Възможни причини са:
– лош изказ;
– объркващи инструкции;
– некачествени дистрактори;
– грешки при формиране на извадката;
– грешен ключ на верните отговори;
– грешно кодиране.

5.2.3. Анализ на дистракторите

Този анализ разглежда как всеки възможен отговор (дистрактор или верен отговор) разграничава учениците според нивото на техните способности. Типичният анализ на една задача изглежда по следния начин:

Задача Q9	1*	2	3	4	8	9
Силна група (ученици с висок резултат)	78,4%	0,1%	14,7%	2,5%	0,0%	4,2%
Средна група (ученици със среден резултат)	52,5%	2,6%	30,8%	6,9%	1,1%	6,2%
Слаба група (ученици с нисък резултат)	18,1%	17,4%	38,1%	17,6%	0,7%	8,1%
Общо	50,5%	6,6%	27,4%	8,9%	0,6%	6,1%

Тази задача (наричана по-долу Q9) има 4 възможни отговора и два кода за липсващ отговор (8 и 9). Код 8 показва, че не е било възможно да се оцени отговорът на ученика поради нечетливост (избрани са 2 отговора или е налице друг оперативен проблем). Код 9 е индикатор за това, че отговорът е оставен празен (ученикът не е посочил отговор). Звездичката (*) до отговор 1 посочва, че това е верният отговор. Общият процент на учениците, посочили отговор 1, е равен на трудността на задачата, а именно 50,5% или 0,50.

Като цяло, една ефективна задача трябва да притежава следните характеристики:
– Колоната с верния отговор трябва да съдържа най-висок процент за силната групата, по-нисък за средната и най-нисък за слабата група.
– Колоните с дистракторите трябва да съдържат приблизително равен процент за трите групи като цяло.

– Процентът на избралите верния отговор за силната група трябва да е по-висок от процентите за същата група в колоните с дистракторите.

– Процентът на избралите верния отговор за слабата група трябва да е по-нисък от процентите за същата група в колоните с дистракторите.

– Процентът на кодовете за липсващ отговор за всички групи трябва да клони към 0;

– Ако е налице голям брой липсващи отговори, процентното разпределение трябва да е почти равномерно между трите групи.

Ако една задача не притежава тези характеристики, това обикновено е резултат от някоя от следните грешки:

Грешка 1. Грешен ключ или грешен код

Задача Q9	1	2	3*	4	8	9
Силна група	78,4%	0,1%	14,7%	2,5%	0,0%	4,2%
Средна група	52,5%	2,6%	30,8%	6,9%	1,1%	6,2%
Слаба група	18,1%	17,4%	38,1%	17,6%	0,7%	8,1%
Общо	50,5%	6,6%	27,4%	8,9%	0,6%	6,1%

Анализът на дистракторите по-горе илюстрира какво се получава, когато се отбележи за верен отговор на задача Q9 отговор 3, вместо отговор 1. При този грешен ключ за верния отговор дискриминативната сила на задачата е -0,23 (14,7%–38,1%). Накратко, при отрицателна дискриминативна сила за една задача вероятно има грешка в ключа. За да се идентифицира верният отговор и респективно да се коригира ключът, трябва да се открие отговорът, който най-добре съответства на изискванията за верен отговор сред данните в таблицата по-горе. В този случай отговор 1 е единственият, за който задача Q9 има положителна дискриминативна сила.

Грешка 2. Ниска дискриминативна сила – повече от един верен отговор

Задача Q9	1*	2	3	4	8	9
Силна група	55,2%	0,1%	37,9%	2,5%	0,0%	4,2%
Средна група	52,5%	2,6%	30,8%	6,9%	1,1%	6,2%
Слаба група	41,3%	17,4%	24,2%	17,6%	0,7%	8,1%
Общо	50,5%	6,6%	27,4%	8,9%	0,6%	6,1%

Този анализ на дистракторите представя резултатите от задача Q9, при положение че отговор 3 е избран за верен отговор заедно с отговор 1. Тази грешка се получава, когато авторите на задачите се опитват да повишат трудността чрез повишаване на достоверността на определени дистрактори. Учениците, за които условието или въпросът са двусмислени, трябва да разчитат на „здравия си разум“, а не на познанията или уменията си, за да достигнат до верния отговор.

Грешка 3. Ниска дискриминативна сила – не се измерват знания или умения по дадения предмет

Задача Q9	**1***	**2**	**3**	**4**	**8**	**9**
Силна група	38,6%	18,1%	13,9%	25,2%	0,0%	4,2%
Средна група	27,0%	12,7%	20,5%	32,5%	1,1%	6,2%
Слаба група	21,5%	25,4%	34,2%	10,1%	0,7%	8,1%
Общо	29,0%	18,7%	22,9%	22,6%	0,6%	6,1%

Анализът на дистракторите по-горе подсказва, че задачата има малко или нищо общо с предмета на оценяването на теста. Това може да се получи в резултат на няколко причини, дори задачата да е валидна от съдържателна гледна точка.

– Изискванията към способността на учениците да четат са завишени, особено ако това не е тест, измерващ уменията за четене. Тази грешка може да се поправи, като се намали обемът на текста, така че изискванията да са ясни за всички ученици.

– Въпросът е двусмислен и не става ясно какво се иска в задачата. Тази грешка може да се поправи, ако при пилотното тестиране учениците бъдат помолени да мислят на глас, докато решават задачата. Недоразумения, предизвикани от информацията в условието (която може да звучи добре на учителите и авторите на задачи, но да не е разбираема за учениците), могат да бъдат поправени.

– Задачата може да бъде пристрастна към определена група ученици. Например в една математическа задача, в която се използва истинска футболна статистика, се проявява пристрастие към момчетата, които могат да са запознати с тази статистика и да отговорят правилно, без действително да са в състояние да решат подобна задача. Пристрастието в задачите може да се понижи чрез процедурата за решаване и мислене на глас, спомената по-горе.

Грешка 4. Ниска дискриминативна сила – твърде лесна или твърде трудна задача

Задача Q9	**1***	**2**	**3**	**4**	**8**	**9**
Силна група	10,2%	32,9%	23,2%	29,5%	0,0%	4,2%
Средна група	5,1%	34,6%	20,5%	32,5%	1,1%	6,2%
Слаба група	2,1%	24,3%	34,2%	30,6%	0,7%	8,1%
Общо	5,8%	30,6%	26,0%	30,9%	0,6%	6,1%

Макар вероятността да бъде избран верният отговор да се променя съобразно нивото на уменията, разграничителната способност на тази задача е твърде ниска, за да даде полезна информация. Всеки от подвеждащите отговори има по-голям шанс да бъде избран от верния отговор при всяка от категориите ученици. Прекомерно трудните задачи трябва да бъдат избягвани в широкомащабни (национални) оценявания. Подобен проблем възниква и при твърде лесните задачи. Ако една задача е решена вярно от почти всички ученици, тя не може да направи разграничение между различните нива на способностите. Чрез адекватна корекция на трудността на задачата ще се повиши и стойността на нейната дискриминативна сила.

5.3. Статистическа валидност

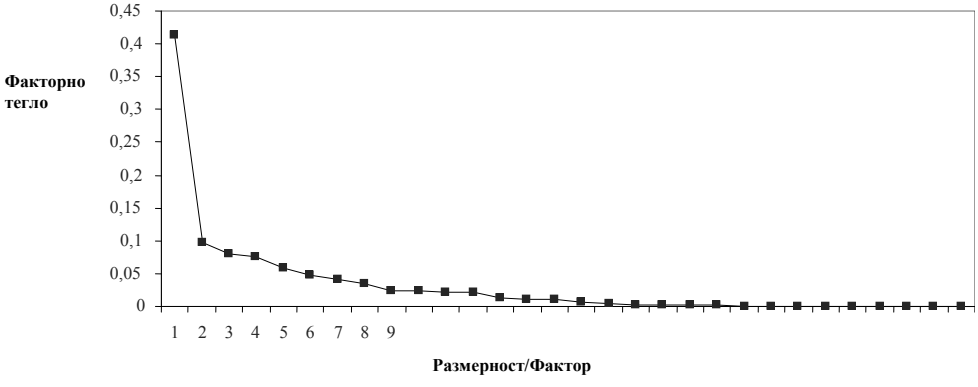
Статистическата валидност е по-скоро показател за качеството на изводите и преценките, направени въз основа на тестовите резултати, отколкото характеристика на самия тест. За да може коректно да подкрепи един извод обаче, представянето на учениците трябва да бъде разглеждано само в контекста на учебния материал, който се измерва, независимо от други характеристики. Съществуват два основни статистически метода за установяване на валидността на теста:

– Анализ на многомерността (факторен анализ) – всички задачи в един тест трябва да измерват един и същ конструкт (съдържателна област). Ако това не е изпълнено, различните резултати от теста ще описват различни конструкти, но ще се интерпретират като такива, отнасящи се само до един конструкт.

– Анализ на различното функциониране на задачите за различни групи ученици (DIF). Чрез този анализ проверяваме дали учениците с еднакво ниво на способностите избират с еднаква вероятност верния отговор на всяка задача – независимо от личните си характеристики.

5.3.1. Анализ на многомерността **(факторен анализ)**

Всяка задача от теста теоретично може да измерва нещо напълно различно от останалите. Ако спецификацията на теста е добре разработена и задачите са написани в съответствие с нея, тогава всички задачи имат общи характеристики. Проблем има само когато общите характеристики не произтичат от връзката на задачите с основната размерност на теста – измервания конструкт. За да потвърдим, че задачите се отнасят към една и съща размерност, се използва графично представяне „scree plot", което идентифицира евентуалното наличие на многомерност.



Фигура 5.1. „Scree plot" графика за тест с 30 задачи

Графиката на теста трябва да притежава следните характеристики:
– Първата стойност трябва да е много по-голяма от останалите.
– Всички останали стойности, без първата, трябва да са със сходна големина.

Големината на факторното тегло на задача онагледява връзката между тази задача и всички размерности на теста. В идеалния случай факторното тегло на всички задачи трябва да бъде сходно с това на теста като цяло.

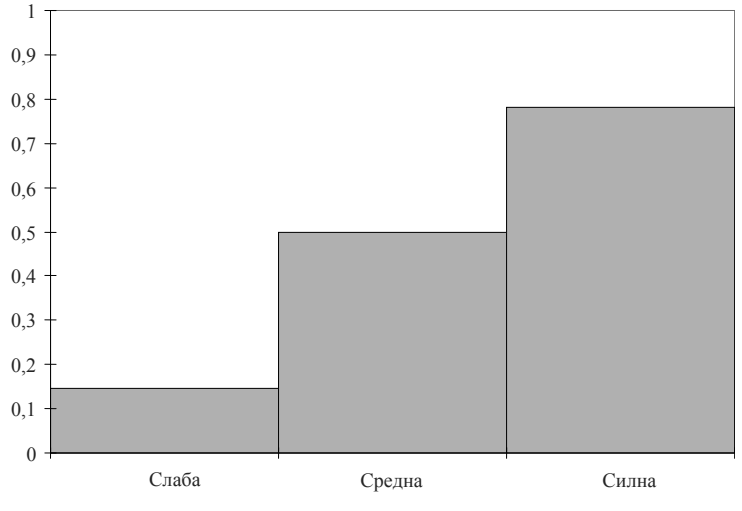
Ако „scree plot" графиката на един тест не изглежда като тази по-горе, задачите трябва да започнат да се отстраняват една по една, докато се постигне желаната едномерност. Факторните тегла на задачите се подреждат в графиката в низходящ ред на стойностите си. Това означава, че задачите, намиращи се на второ и трето място, се различават най-много от измервания от теста конструкт (фактор). Критерият за премахване на задачи е те да имат силно факторно тегло във втора или трета размерност. Най-проблемните задачи трябва да бъдат отстранени първи.

5.3.2. Анализ на различното функциониране на задачите за различни групи ученици (DIF)

Целта на този анализ е да предотврати някои подгрупи ученици да реагират на задачите по различен начин. Макар различните групи ученици да се справят по различен начин с теста, има еднаква вероятност учениците с еднакви умения от една и съща група да отговорят вярно на една и съща задача. Задачи с големи разминавания във вероятността за постигане на верен отговор между групите трябва да бъдат разгледани от експертната група, която да установи дали дадена задача не е несправедлива по отношение на определена група ученици. Ако експертната група прецени, че задачата е несправедлива, екипът по разработването на теста може да реши да я премахне от теста.

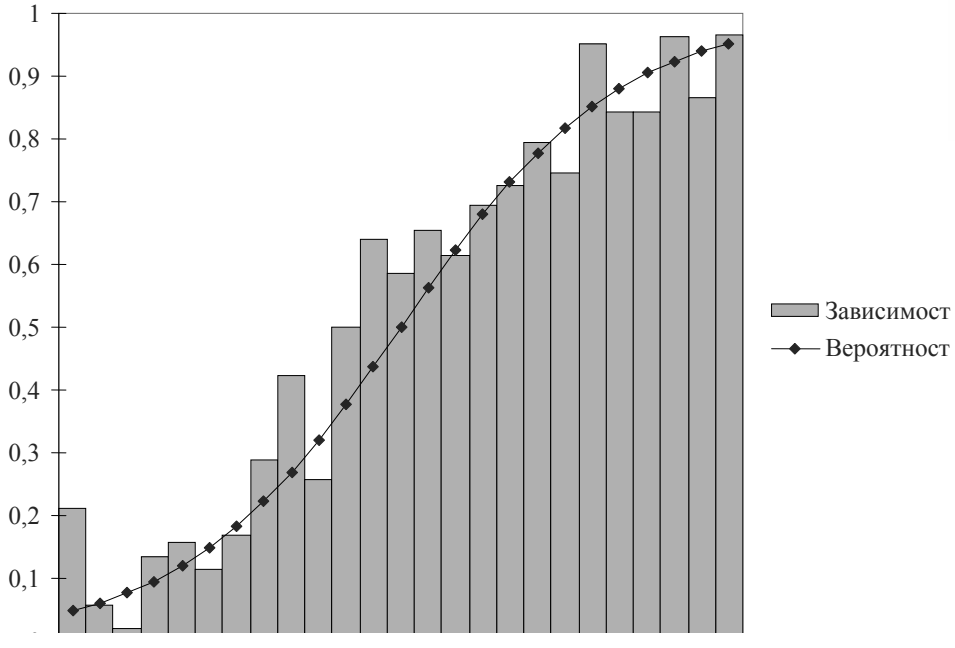
5.4. IRT Анализ

Теорията за вероятностното моделиране (IRT) приема, че вероятността учениците с определено ниво на способностите да отговорят вярно на дадена задача е независима от популацията, към която принадлежат, и от наличието на други задачи в теста. Този принцип надгражда класическата дискриминативна сила. Класическата дискриминативна сила се изчислява, като от частта на верните отговори на учениците от силната група се извадят тези на учениците от слабата група. Ако искаме да изобразим тези отношения, включващи и частта на верните отговори и на учениците от средната група, то ще получим следната графика:



Фигура 5.2. Зависимост на верните отговори спрямо силната, средната и слабата група ученици

Учениците могат да се разпределят в много повече от три групи в зависимост от способностите им в така наречените нива на способностите. **Фигура 5.3.** показва едно такова представяне.



Фигура 5.3. Зависимост на верните отговори спрямо нивата на способностите на учениците

Тъй като въз основа на данните се правят допълнителни изчисления, нараства вероятността за възникване на грешка на извадката при всяко от изчисленията. Статистически грешката на извадката може да бъде намалена, ако се приеме, че зависимостта е монотонно нарастваща функция.

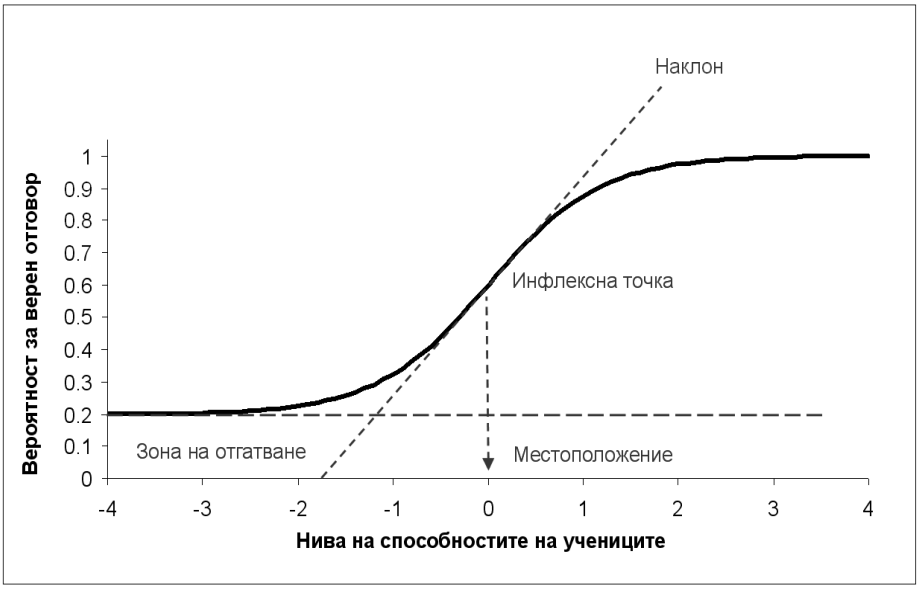
Кривата, наречена характеристична крива на задачата (item response function – IRF), представя вероятността за избор на верния отговор за всяко ниво на способностите. Тъй като IRF е монотонно нарастваща крива, наклонът ù се променя при преминаване от ниско към високо ниво на успеваемост. Точката, където кривата е най-стръмна, се нарича инфлексна точка.

Видът на всяка характеристична крива на задача се определя от не повече от 3 характеристики:

– местоположението на инфлексната точка на „S-образната“ крива спрямо хоризонталната ос определя трудността на задачата;

– наклонът на характеристичната крива, зададен от големината на ъгъла в инфлексната точка на „S-образната“ крива и хоризонталната ос, определя дискриминативната сила на задачата;

– точката, в която левият край на „S-образната“ крива би пресякъл вертикалната ос, определя вероятността учениците с ниски способности да налучкат верния отговор.



Фигура 5.4. Характеристична крива на задача (IRF)

При определена „зона на отгатване“ наклонът на характеристичната крива за отделните нива на способности илюстрира способността на една задача да разграничи учениците, чиито способности са под и над необходимото ниво за правилното ù решаване. Характеристичните криви на всички задачи могат да бъдат обединени, за да се получи характеристичната крива на теста (test characteristic curve – TCC).

Характеристичната крива на задачата може да бъде трансформирана по математически път във функция, наречена информационна функция на задача (item information function – IIF), която илюстрира колко полезна е една задача при оценяване на учениците от различните нива на способностите. Една добре

написана задача е действително прецизна само за определено ниво на способностите. Информационната функция на задачата позволява на авторите на теста да подберат задачи, покриващи специфични нива на способностите, така че да увеличат прецизността на теста съобразно конкретната му цел.

Макар при някои групи ученици да се наблюдават променливи нива на способностите, IRT теорията приема, че характеристичната крива на една задача е една и съща за всички ученици до ниво линейна трансформация. Това свойство, наречено инвариантност, позволява лесно да се приравняват оценките от различни тестове, съдържащи група от едни и същи тестови задачи.

5.4.1. Изчисляване на резултатите чрез IRT

За разлика от резултатите, получени чрез класическата теория на тестовете, които представляват общия брой правилно решени задачи, разделен на общия брой решени задачи, IRT резултатите оползотворяват цялата информация от отговорите на учениците. IRT резултатът на теста включва информацията от характеристичните криви на всички задачи. Тъй като IRT резултатите използват повече информация от класическите резултати, те са по-точни и по-добре разграничават учениците, особено при по-кратките тестове.

5.4.2. Селекция на задачите чрез IRT

Поради различните си цели различните тестове изискват различни нива на точност спрямо способностите на учениците. Ако резултатите се използват за преценка дали учениците покриват определено ниво на знания и умения, тестът трябва да постига максимална точност по отношение на конкретното ниво на способностите. **Таблица 5.5.** показва идеалните нива на точност при различните нива на способности за тестове с различна цел.

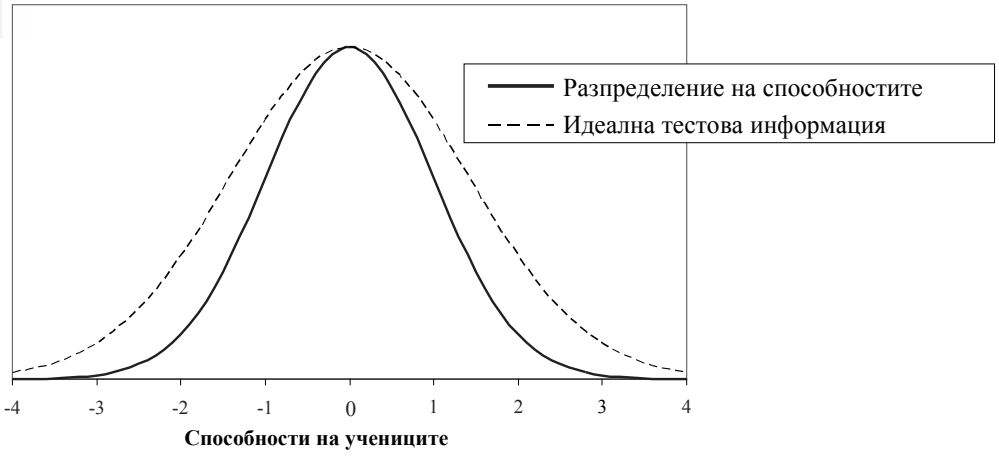
Таблица 5.5. Нива на точност

Цел	Ниво на способностите		
	Ниско	Средно	Високо
Обща грамотност	Средна точност	Висока точност	Средна точност
Завършване на образователен етап (образователен минимум)	Висока точност	Средна точност	Ниска точност
Селекция	Ниска точност	Средна точност	Висока точност

Информационната функция на задачата описва колко точна е всяка задача за всяко ниво на способностите. Информационните функции на всички задачи в един тест се сумират, за да се получи информационната функция на теста.

Тя описва точността на теста за всяко ниво на способностите. Задачите трябва да бъдат подбрани така, че информационната функция на теста да отговаря на целта на оценяването.

Идеалната тестова информация е малко по-широка от успеваемостта на учениците, която се изследва. Идеалната информационна функция за едно национално оценяване обикновено изглежда така:



5.5. Постигане на сравнимост между тестове чрез IRT

Ако един и същ набор от задачи се даде на две групи ученици с еднакви способности, характеристичните криви на задачите ще бъдат идентични. Ако групите се различават по способности обаче, функциите на задачите трябва да се приравнят, за да се получат сравними резултати. Само след това може да се постигне сравнимост между тестовете, дори те да съдържат и други задачи, които не са дадени и на двете групи.

С помощта на IRT може да се постигне сравнимост между резултатите от два теста с общи задачи, като се следва следният алгоритъм:

1. Изчисляване на параметрите и резултатите на единия тест.
2. Изчисляване на параметрите и резултатите на другия тест.
3. Пресмятане на характеристичната крива на теста за общите задачи в двата теста.
4. Намиране на константите на линейното уравнение, използвано за минимизиране на разликата между характеристичните криви на двата теста.
5. Прилагане на линейното уравнение към параметрите на задачите на единия тест за постигане на сравнимост между характеристичните криви на задачите.

6. Проверка дали приравняването на задачите е извършено правилно за всяка обща задача. Ако приравняването не е било успешно по отношение на някоя задача, тя се премахва от списъка с общите задачи и се повтарят стъпки от 3 до 5.

7. Прилагане на линейното уравнение към резултатите на учениците за постигане на сравнимост на тестовите резултати.

6. Интерпретиране на тестовите резултати

Оценяването е създадено, за да служи на определена цел – да предоставя информация за извличане на определени изводи или вземане на решения. Интерпретирането на резултатите трябва също да бъде съобразено с нуждата от информация за постигане на целта на оценяването. Има два метода, които обикновено се използват за трансформиране на тестовите резултати в подходящи за докладване данни:

– Преобразуване на резултатите (Rescaling).

– Стандартизиране (стандартни скàли).

6.1. Преобразуване на резултатите

Разпределението на IRT резултатите е близко до стандартното нормално разпределение, което е със средна стойност, равна на 0, и стандартно отклонение, равно на 1. Поради тази причина половината от резултатите ще бъдат отрицателни и няма да има ограничение за определяне на максимален или минимален резултат. Широката общественост не е запозната с нормалното разпределение и трудно би разтълкувала резултатите, представени чрез такава скала. Резултатите трябва да бъдат представени по възможно най-разбираем начин на широката публика. Използват се два метода за преобразуване (rescaling) на IRT резултатите:

– Чрез линейна трансформация – резултатите трябва да бъдат умножени по определено число, за да се получи желаното стандартно отклонение. Към получения резултат се добавя определено число, за да се получи желаната средна стойност. Най-важната цел на линейната трансформация е да преобразува резултатите на всички ученици в неотрицателни оценки. Някои популярни (стандартизирани) скàли, използващи линейна трансформация, са със средна стойност 50 и стандартно отклонение 10, или средна стойност 500 и стандартно отклонение 100.

– Чрез нелинейна трансформация – резултатите се преобразуват чрез нелинейна функция, като характеристичната крива на теста, за да има в новата скàла абсолютна минимална и максимална стойност. Предимството на нелинейната трансформация е, че представя резултатите с ограничен абсолютен максимум. IRT резултатите, получени чрез линейна трансформация на характеристичната крива на теста, се наричат IRT суров бал, защото се интерпрети-

рат като класически тестови резултати, но са доста по-прецизни.

6.2. Установяване на стандарти за качествени оценки (прагови стойности)

Критериите за преобразуване на точките в оценка помагат на заинтересованите страни да си отговорят на въпроса „Как са се представили учениците?“ на прост език, а не на езика на статистиката. Те обикновено се наричат „нива на постижения на учениците“ или „оценъчни стандарти“.

В най-простия случай оценяването използва само един стандарт, който се нарича „прагова стойност“. Счита се, че учениците, които са покрили стандарта, са преминали теста успешно, докато тези, които не са го достигнали, са се провалили. Други примери за стандарти са „отговарят на очакванията“ и „надхвърлят очакванията“. Обикновено стойността на надеждността на теста (която определя и стандартната грешка на измерването) не позволява използването на повече от 4–5 стандарта.

Има няколко метода за установяване и използване на оценъчните стандарти, но те всички следват една и съща обща процедура:

1. Оценъчните стандарти се определят от заинтересованите страни.
2. Задачите се съставят, за да се измерят стандартите.
3. Събират се данни (тестът се провежда).
4. Статистическият анализ подсказва оптималния праг за разпределяне на учениците по нива на постижения.
5. Заинтересованите страни определят финалното разпределение.
6. Резултатите на учениците се превръщат от точки в оценки.

При повечето оценявания заинтересованите страни са представени от екип по изготвяне на теста или от специална комисия за определяне на стандартите.

Има два основни статистически метода за определяне на оптималните прагови стойности:

– основани на съдържанието;

– основани на статистически изчисления.

И при двата метода се налага определена част от стандартите за знанията и уменията на учениците да бъдат определени предварително.

6.2.1. Методи, основани на съдържанието

Методите, основани на съдържанието, изискват следването на определени стъпки:

1. Експертната група, която одобрява задачите, или екипът по изготвяне на теста определя кое ниво на постиженията измерва всяка задача.
2. Екипът или подкомисията по установяване на праговете преглежда задачите за всяка група и определя броя на задачите, които един ученик трябва да реши правилно, за да покрие прага за даденото ниво на постижения.

3. Броят се усреднява, за да се изчисли вероятността за избор на верния отговор, която да се използва за интерпретиране на резултатите от теста.

4. За всички задачи от дадено ниво на постиженията се изчислява характеристична крива на теста.

5. Праговете, определящи нивата на постиженията, се изчисляват, като стойностите на вероятностите за избор на верен отговор се прилагат към характеристичната крива на теста.

5. Комисията за определяне на стандартите преглежда получените прагове и ако е необходимо, ги преработва за по-добро докладване на резултатите. Обикновено това означава да се провери дали интервалът на IRT резултатите за всяко ниво на постиженията е еднакъв.

6. Резултатите на учениците се превръщат в оценки или учениците се причисляват към определено ниво на постижения, като се използват окончателно одобрените прагове.

6.2.2. Методи, основани на статистически изчисления

Методите, основани на статистически данни, изискват следните стъпки:

1. Параметрите a и b (наклон и инфлексна точка според **Фигура 5.4.**) се използват, за да се изчислят групите задачи с еднакви или близки резултати по тези параметри, използвайки процедурата за K-броя независими клъстера, при която всяка задача се оценява според максималната стойност на наклона на характеристичната ù крива, а броят на клъстерите (K) отговаря на броя на стандартите.

2. Центровете на всеки клъстер се намират, като се пресметне средната трудност на задачите, определени за същия клъстер. Всяка задача се претегля (умножава) със стойността на максималния наклон на характеристичната крива. Пресмята се праговата стойност като средна стойност на съседни клъстерни центрове.

3. Комисията по установяване на стандартите преглежда изчислените прагове и ако е необходимо, ги преработва за по-добро докладване на резултатите. Обикновено това се свежда до проверка дали интервалът от резултатите за всяко ниво на постиженията е еднакъв.

4. Резултатите на учениците се превръщат в оценки и учениците се причисляват към дадено ниво на постижения, като се използват окончателно одобрените прагове.

6.3. Вторичен анализ

Вторичният анализ се използва в случаите, когато се интересуваме от успеваемостта на учениците по групи или индивидуално. Има голям брой статистически методи, широко използвани за анализ на тестови резултати. Специфичните методи за анализ на резултатите от един тест трябва да са съ-

образени с целта на теста. Има две основни цели, които налагат вторичен анализ на резултатите:

– Мониторинг и оценка – резултатите от теста се използват за оценка на състоянието на съществуващата учебна програма.

– Вземане на стратегическо решение – резултатите от теста предоставят фактическата база, въз основа на която да се реши как да се промени съществуващата или да се създаде нова учебна програма.

Статистическите методи, подходящи за оценявания, резултатите от които ще се използват за мониторинг и оценка, включват стандартните описателни статистики – като честота, средна стойност и стандартно отклонение. Повечето резултати се докладват с помощта на графики със стълбове, онагледяващи броя или процента ученици от различните нива на постижения.

Напредъкът с течение на времето трябва да бъде онагледяван с графики с линии, сравняващи средните резултати в различни времеви отрязъци, или процента ученици над определено ниво на постиженията. Сравняването на групите трябва да се ограничи до важни демографски характеристики, като географски регион, пол и майчин език.

Резултатите трябва да бъдат съобщени на отделните училища, но училищните ръководства и заинтересованите страни не бива да използват резултатите, за да оценяват нивото на училището като цяло. Вместо това резултатите трябва да послужат за мониторинг на потенциални проблеми или незадоволени потребности на учениците в определено училище или регион. Във всички случаи подробните сравнения обикновено са съпътствани от стандартна грешка на измерването. Големите таблици със статистически коефициенти трябва да бъдат сведени до минимум или да бъдат публикувани само в статистическите приложения. Интерпретацията на резултатите трябва да бъде сведена до прости описателни твърдения.

Статистическите методи, подходящи за хората или институциите, вземащи стратегически решения, са съотношения и статистически модели с няколко променливи наведнъж. Анализът за информиране на лицата, вземащи политически решения, може да използва и друга информация освен данните от оценяването.

Резултатите трябва да бъдат докладвани с помощта на точкови диаграми, регресионни криви и разпределения на вероятностите. За вземането на политическо решение е необходимо далеч по-голямо усилие за набавяне на доказателства, които да потвърдят направените изводи въз основа на резултатите. Тук ограниченията за размера и естеството на представената информация са по-малко. Например може да се наложи определени изводи да бъдат подкрепени именно с големи статистически таблици.

БЕЛЕЖКИ

1. Ръководството е разработено от Фернандо Картрайт и Джери Мусио за Център за контрол и оценка на качеството на училищното образование по проект, финансиран от Световната банка.

DEVELOPMENT OF TESTS, A MANUAL

Abstract. Education has become an important determinant of the social and economic progress of individuals and of nations, and it typically ranks as one of the top priorities established by national governments.

National student assessment programs play a central role in measuring the quality of education and monitoring the impacts of education reform. As such, assessment agencies have a responsibility to ensure that the development of tests and examinations meet the highest professional standards.

The purpose of this manual is to assist a staff responsible in the design and development of student tests, and the analysis and reporting of results. Our focus here is on the practical application of test theory based on our experiences in carrying out a variety of assessment programs in Canada and internationally.

The main themes are as follows: planning the assessment, test design, item development, test development & pilot testing, item and test analysis, interpreting test results.

Fernando Cartwright
Jerry Mussio