

<https://doi.org/10.53656/math2023-5-5-cho>

Educational Technologies
Образователни технологии

ИЗБОР НА ПОДХОДЯЩ МОДЕЛ ЗА ИЗСЛЕДВАНЕ НА ЗАДАЧИТЕ ОТ НВО ПО МАТЕМАТИКА СЛЕД VII КЛАС СПОРЕД МЕТОДИТЕ НА IRT

Ас. Павлин Цонев

Висше военновъздушно училище „Г. Бенковски“ – Д. Митрополия

Резюме. В статията, с методите на Item Response Theory, са анализирани измерителните качества на тестовите задачи с избираем отговор от НВО за VII клас за областите София-град и Плевен, проведено през 2021 година. След съгласуване с емпиричните данни е обоснован изборът на двупараметричния модел по тази методика. Направено е сравнение на получените данни и тези от Класическата теория на тестовете.

Ключови думи: Item Response Theory; вероятностно моделиране; тестови задачи с избираем отговор; трудност; разграничителна сила

1. Увод

Националното външно оценяване (НВО) на учениците от VII клас е важен елемент от тяхното образование. Анализирането на резултатите, получени вследствие от това оценяване, е важно за образователната система. Всяка година се правят класирания по двата изпита – български език и литература и математика. Сравняват се резултатите от тези класирания. Правят се изводи за нивото на успеваемост в различните населени места. Сравнява се трудността на изпитите в различни години. Всичко това се прави с методите на Класическата теория на тестовете (КТТ).

Публикацията в сп. „Математика и информатика“ (Tsonev 2022) с методите на КТТ изследва измерителните качества на задачите с избираем отговор от проведеното през 2021 година НВО по математика за VII клас, като са използвани данните от областите София-град и Плевен. Настоящата статия прилага по-съвременни методи за такова изследване, а именно методологията на IRT (Item Response Theory), описана в много източници, като например (McDonald 1999), (Hambleton 1991) и др. В този смисъл, статията е продължение на цитираната публикация.

По-долу, с методите на IRT, са анализирани измерителните качества на гестовите задачи с избираем отговор на проведеното през 2021 година НВО по математика за VII клас, като са използвани данните от областите София-град и Плевен. Целият тест е достъпен на сайта на МОН¹. В резултат от анализа е направен извод за избор на подходящ IRT модел за изследване на задачите с избираем отговор в НВО по математика за VII клас.

Изследването на измерителните качества с помощта на КТТ е свързано с някои проблеми, като например:

- оценката на измерителните качества на задачите и на теста, като цяло, зависи от конкретната група лица и конкретната група задачи, които участват в изследването;
- трудността и разграничителната сила на задачите зависят от изследваните лица;
- измерването на надеждността чрез разделяне на теста на две половини или чрез повторно провеждане на теста са, по същество, изкуствени.

Някои от ограниченията на класическата теория могат да бъдат преодоляни, ако се приложи единна скала за представяне на трудността на гестовите задачи и постиженията на учениците в една и съща мерна единица. Такава възможност предлага IRT методологията, която на български език е известна като „Вероятностно моделиране“ (Bankov 2002) или „Теорията за отговор на тестов въпрос“ (Jalev 2014).

Тук ще разглеждаме и използваме IRT модели, които са за задачи от вида „решена – нерешена“, т.е. не се дават бонуси за частичен напредък по задачата. В основата на IRT стои предположението, че вероятността един ученик да реши правилно дадена тестова задача, може да бъде изразена чрез една абстрактна величина, наречена *способност*. Функцията, която пресмята тази вероятност, се нарича *характеристична функция*, а съответната ѝ графика – *характеристична крива*. На всяка задача се съпоставя една такава функция. Най-често разпространените методи за избор на характеристична функция са с един, два или три параметъра и се наричат съответно едно-, дву- и трипараметрични модели.

Еднопараметричният модел на IRT, който за краткост ще означаваме с **1PL**, се определя от характеристичната функция

$$P_i(\theta) = \frac{e^{\theta - b_i}}{1 + e^{\theta - b_i}},$$

където i е номерът на задачата, $P_i(\theta)$ е вероятността произволно избран ученик със *способност* θ да реши i -тата задача, а b_i е *трудността* на задачата. Така вероятността ученик със *способност* θ да реши задача с

трудност b_i , е равна на 0,5. Това дава възможност за способност на учениците θ и трудност на задачите b_i да се използва една и съща скала, наречена *скала на способностите* (ability scale). Интервалът, в който обичайно се изменя b_i , е $(-2; 2)$, като колкото е по-близка тази стойност до 2, толкова е по-висока трудността на i -тата задача. Графично това означава, че характеристичната крива на по-трудните задачи се разполага по-вдясно на чертежа.

Двупараметричният модел (**2PL**) зависи от още един параметър a_i , който се нарича *дискриминация* на задачата и се определя с функцията:

$$P_i(\theta) = \frac{e^{1,7a_i(\theta-b_i)}}{1 + e^{1,7a_i(\theta-b_i)}}.$$

По-високите стойности на дискриминацията задават по-стръмни характеристични криви и по-добре разграничават учениците със способности, близки до b_i . Най-често интервалът, в който приема стойности a_i , е $(0; 2)$. При трипараметричния модел (**3PL**) се въвежда и параметър c_i , който се счита за *коефициент на налучване*. Характеристичната функция на този модел е:

$$P_i(\theta) = c_i + (1 - c_i) \frac{e^{\theta-b_i}}{1 + e^{\theta-b_i}}.$$

Графично стойността на параметъра c_i изобразява хоризонтална асимптотата $P = c_i$ на характеристичната крива при $\theta \rightarrow -\infty$.

Някои от основните предимства на IRT моделите, които ги правят за предпочитане пред КТТ, са:

- оценката на параметрите a_i , b_i и c_i не зависи от контролната група лица, с която е правено изследването;
- статистическите свойства на задачите се изследват по-прецизно;
- тези статистически свойства се онагледяват с графиките на съответната функция $P_i(\theta)$ (по-долу има примери на такива графики за конкретни задачи), което улеснява интерпретацията и разбирането им. На графиките хоризонталната ос е оста на способностите θ (споменатата ability scale), а по вертикалната ос се нанася вероятността $P_i(\theta)$ за правилно решаване на задачата, пресметната с някоя от функциите по-горе.

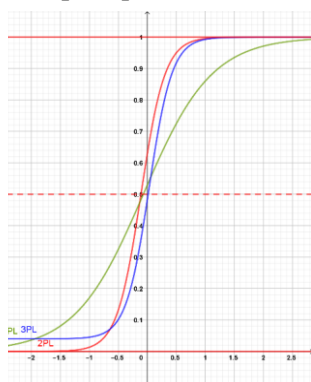
Данните в изследването са обработени с приложението jMetrik², от което са получени трите параметъра a_i , b_i и c_i , а графиките са начертани с приложението GeoGebra³.

2. Сравняване графиките на трите модела за параметризация

За изследването са използвани резултатите от проведеното НВО в края на VII клас по математика с 10 995 ученици от област София-град и 1845 ученици

от област Плевен, които са се явили на изпит през 2021 година. Разглеждани са само първите 18 задачи с избираем отговор. Обработени са данните за всички 18 задачи за учениците от двете области, взети заедно. Използваният софтуер jMetrik оразмерява скалата на способностите така, че разпределението на способностите на всички ученици е със средна стойност 0 и стандартно отклонение 1.

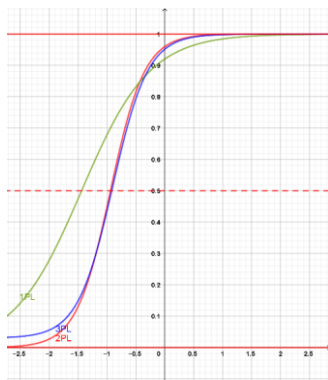
След като се начертаят графиките на характеристичните криви за всяка задача според трите IRT модела, се забелязват няколко групи, в които могат да се класифицират задачите.



1. Стойността на израза $2021 - 2020 \cdot (-0,1)$ е:

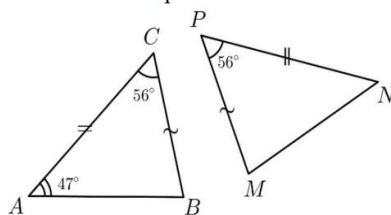
- А) $-0,1$
- Б) 1
- В) 2223
- Г) 2219

Фигура 1. IRT характеристични криви на задача 1



10. Дадени са $\triangle ABC$ и $\triangle MNP$, за които $AC = NP$, $BC = MP$, $\angle CAB = 47^\circ$ и $\angle ACB = \angle MPN = 56^\circ$. Мярката на $\angle NMP$ е:

- А) 97°
- Б) 77°
- В) 56°
- Г) 47°

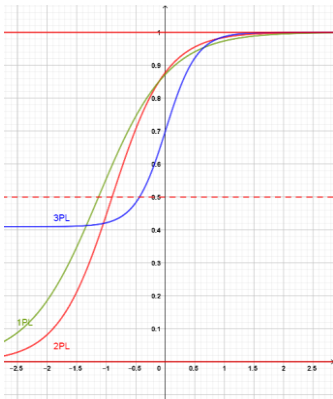


Фигура 2. IRT характеристични криви на задача 10

- По отношение на 3PL модела повечето задачи са с висока стойност на параметъра на налучкване, но има и такива като задачи 1 (фиг. 1) и 10 (фиг. 2), при които този параметър е нисък. За някои задачи той е над 0,3, като

например задачи 2, 4, 10 и 12. Например графиката 3PL на задача 2 (фиг. 3) показва, че този модел не може да разграничи учениците със способности, по-ниски от $\theta < -0,5$, т.е. учениците с такива способности могат с вероятност около 0,4 да посочат правилния отговор на задачата поради ефекта на налучкването.

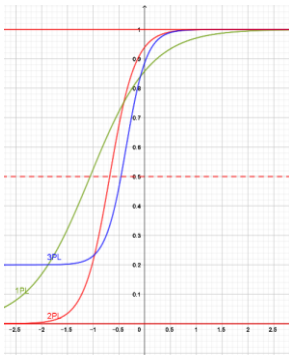
• Задачи, в които 1PL моделът се различава съществено от другите два модела – задачи 1 (фиг. 1), 4, 10 (фиг. 2) и 12. Графиките на тези задачи по 2PL и 3PL моделите са по-стръмни, защото при тях параметърът $a > 1$, т.е. задачите имат по-висока разграничителна сила.



2. Стойността на израза $a^2 - b^2$, при $a = 10,5$ и $b = 9,5$ е:

- А) 18,5
- Б) 19
- В) 19,5
- Г) 20

Фигура 3. IRT характеристични криви на задача 2



9. В кутия има 3 червени молива и два пъти повече зелени. Вероятността, произволно изваден молив да е червен, е:

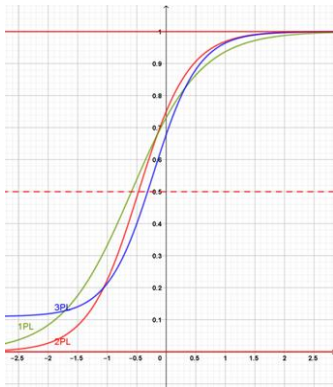
- А) $\frac{1}{3}$
- Б) $\frac{2}{5}$
- В) $\frac{3}{5}$
- Г) $\frac{2}{3}$

Фигура 4. IRT характеристични криви на задача 9

• Задачи, в които трите графики имат големи разлики – задачи 9, 18 и 14.

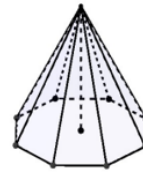
Например в задача 9 (фиг. 4) моделът 1PL дава слаба разграничителна сила, а моделът 3PL може да разграничи само ученици със способности $-1 < \theta < 0,5$.

• Трите модела имат сходни графики – например задача 11 (фиг. 5), която има много добра разграничителна сила и нисък коефициент на налучкване.



11. На чертежа е дадена правилна осмоъгълна пирамида
Лицето на една от околните ѝ стени е 18 cm^2 .
Лицето на околната повърхнина на пирамидата е:

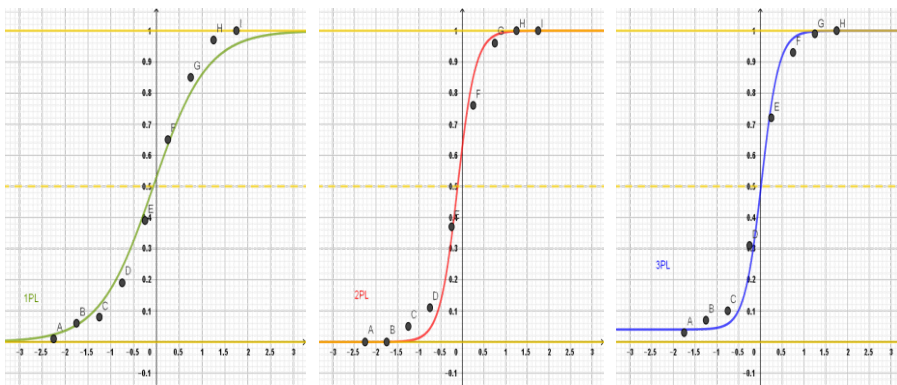
- A) 108 cm^2
B) 126 cm^2
B) 144 cm^2
Г) 162 cm^2



Фигура 5. IRT характеристични криви на задача 11

3. Сравнение на трите модела спрямо параметъра θ на способностите на учениците.

За да се избере подходящ IRT модел, се търси този, който има най-добро съгласуване с емпиричните данни. Ще приложим метод, който е описан на стр. 66 – 67 в книгата (Hambleton 1991). За всеки от трите модела всеки от учениците получава стойност на параметъра θ , който е число в интервала $(-3; 3)$. Този интервал се разделя на 12 равни части. След това за всяка задача

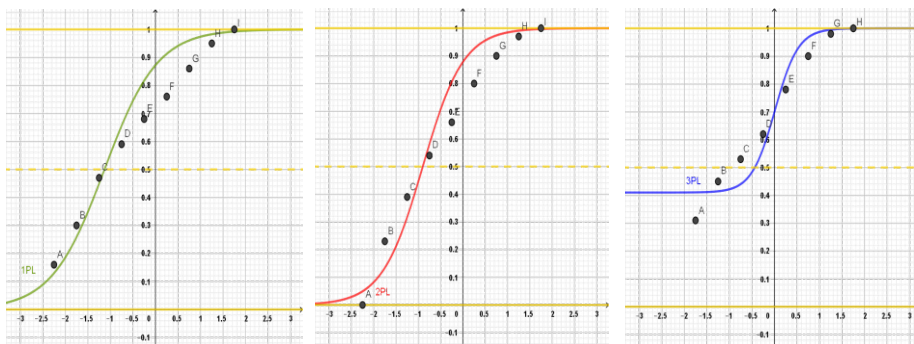


Фигура 6. Графики с остатъците на задача 1 за трите IRT модела

и за всеки подинтервал се намира отношението на броя на учениците, които са отговорили вярно на въпроса, и броя на всички ученици в този подинтервал. Така на фигура 6 са изобразени графиките за задача 1 (наречени графики с остатъците).

Всяка точка $A, B, C, \dots$ има абсциса средата на интервал с дължина 0,5 и ордината частта от учениците, решили вярно задачата, които имат способност θ в този интервал. За тази задача може да се забележи, че:

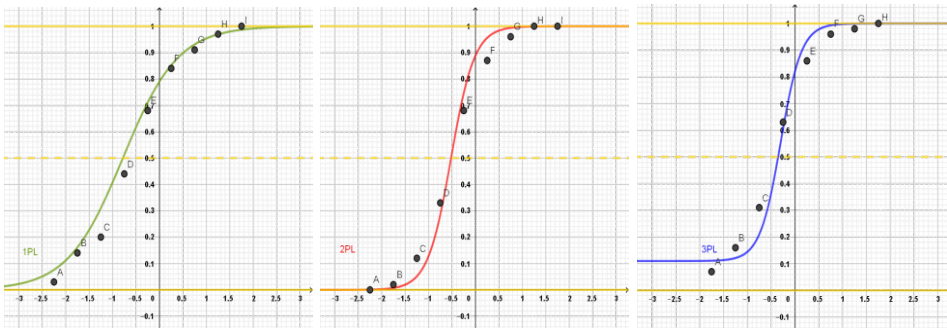
- има оптимална трудност;
- трите криви са близки по вид;
- липсват точки, съответстващи на ученици със способности по-малки от -2,5 и по-големи от 2;
- и при трите модела точките са близко разположени до графиките; това показва добра съгласуваност и за трите модела;
- 1PL моделът има по-слаба разграничителна сила;
- графиките на 2PL и 3PL моделите малко по-добре се доближават до емпиричните данни, но при 3PL модела няма ученици със способности, по-малки от -2.



Фигура 7. Графики с остатъците на задача 2 за трите IRT модела

На фигура 7 са изобразени графиките с остатъците на задача 2. При нея се вижда, че:

- задачата е сравнително лесна;
- моделът 3PL дава висок коефициент на налучкване, което не се съгласува добре при учениците с ниски способности (точка A е далеч от кривата);
- моделите 1PL и 2PL се съгласуват добре с данните за учениците с по-ниски способности, като 2PL има малко по-добра разграничителна сила.



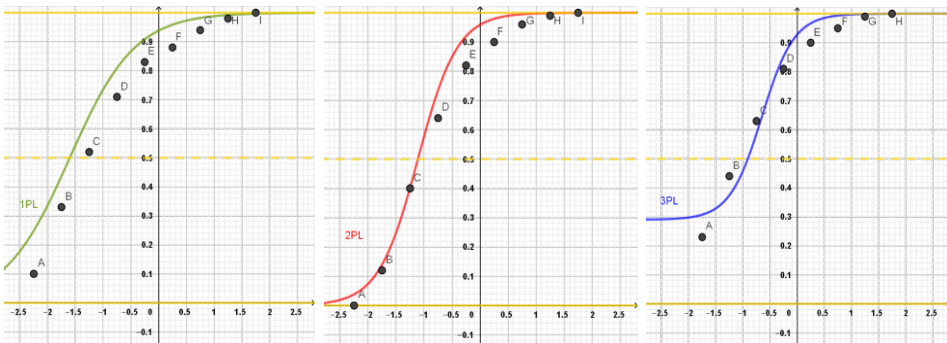
Фигура 8. Графики с остатъците на задача 4 за трите IRT модела

На фигура 8 са изобразени графиките с остатъците на задача 4. При нея се вижда, че:

- задачата е сравнително лесна и с добра разграничителна сила;
- добро съгласуване на данните, особено за 2PL модела.

На фигура 9 са изобразени графиките с остатъците на задача 18. При нея се вижда, че:

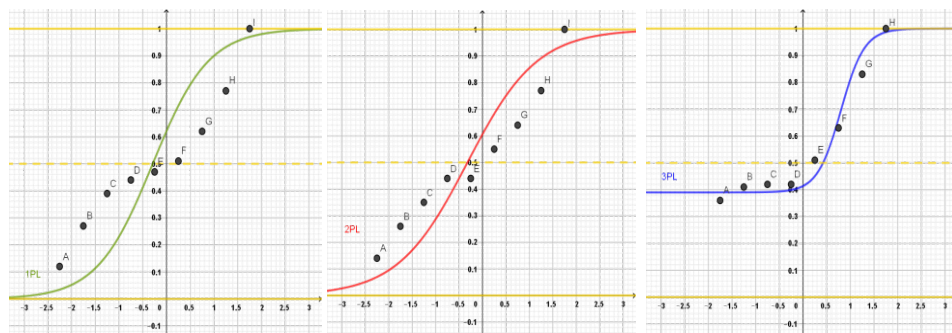
- задачата е лесна и с добра разграничителна сила;
- добро съгласуване на данните особено за 2PL модела;
- моделът 1PL дава теоретично по-силно представяне на учениците, отколкото са емпиричните данни;
- моделът 3PL дава високо ниво на налучкване по-сериозни разлики при учениците с по-ниски способности.



Фигура 9. Графики с остатъците на задача 18 за трите IRT модела

Подобни изводи могат да се направят за почти всички задачи с изключение на задача 17. При нея се наблюдават по-сериозни разлики в кривите и емпиричните резултати (фиг. 10). Тя е пример за това, че задача с недобри

характеристики според Класическата теория на тестовете (в случая тя има слаба разграничителна сила), не се съгласува добре с моделите от IRT. Вижда се голямо разминаване, особено в 1PL и 2PL моделите, както и трудността при тях и тази при 3PL. За първите два модела може да се каже, че слабите ученици се представят по-добре, а силните ученици – по-зле, отколкото дава теоретичната крива. За тази задача 3PL има най-добра съгласуваност с емпиричните данни. Този модел обаче има добра разграничителна способност само за учениците с високи стойности на θ .



Фигура 10. Графики с остатъците на задача 17 за трите IRT модела

В заключение, изглежда, че трите модела дават добра съгласуваност с емпиричните данни. Сериозен недостатък на модела 3PL е, че липсват ученици със способност под -2 , като в много от задачите той дава високи стойности на коефициента на налущване и с това се различава съществено от другите два модела. При това, за ниски стойности на θ този модел няма добра разграничителна сила и не се съгласува много добре с емпиричните данни. Моделът 1PL не отчита добре разграничителната сила на задачите. Следователно може да се каже, че 2PL моделът е за предпочитане за моделиране на данните с IRT.

4. Съпоставка на резултатите от КТТ и IRT

В таблици 1 и 2 са представени обобщените резултати на осемнадесетте задачи според КТТ и според двупараметричния 2PL модел на IRT. Задачите са подредени низходящо според съответния коефициент, получен от КТТ.

Коефициентът на корелация в таблица 1 е отрицателен, защото висока стойност на коефициента за трудност от КТТ всъщност означава, че задачата е решена от повече ученици, т.е. тя е лесна, докато това съответства на ниските стойности на параметъра b от IRT. Абсолютната стойност на този

коэффициент е практически равен на 1. Следователно и двете теории класифицират по еднакъв начин трудността на задачите.

Таблица 1. Наредба според трудността				Таблица 2. Наредба според rbpis			
№	Трудност	b-param		№	Rpbis	a-param	
12	0,82	-1,11	лесна	1	0,68	2,84	много добра
18	0,78	-1,11	лесна	4	0,59	2,35	много добра
10	0,75	-0,94	лесна	9	0,57	2,38	много добра
2	0,70	-0,90	лесна	14	0,56	1,87	много добра
9	0,69	-0,68	оптимална	7	0,54	1,70	много добра
16	0,67	-0,71	оптимална	5	0,50	1,53	много добра
15	0,65	-0,68	оптимална	13	0,49	1,48	много добра
4	0,64	-0,52	оптимална	10	0,49	1,99	много добра
11	0,61	-0,47	оптимална	12	0,48	2,58	много добра
5	0,60	-0,41	оптимална	8	0,48	1,42	много добра
13	0,59	-0,39	оптимална	16	0,47	1,50	много добра
14	0,58	-0,34	оптимална	11	0,47	1,38	много добра
8	0,55	-0,25	оптимална	6	0,47	1,34	много добра
17	0,55	-0,32	оптимална	15	0,44	1,30	много добра
7	0,53	-0,15	оптимална	18	0,44	1,67	много добра
1	0,51	-0,11	оптимална	3	0,43	1,20	много добра
3	0,45	0,19	оптимална	2	0,41	1,29	много добра
6	0,39	0,40	оптимална	17	0,31	0,79	добра
Коефициент на корелация -0,99079				Коефициент на корелация 0,82955			

Според таблица 2 корелацията между разграничителната сила на задачите, която се дава от коефициентите rbpis и параметъра a , също има висока стойност. Всички задачи, с изключение на задача 17 имат параметър $a > 1$, което означава, че имат много добра разграничителна сила.

Две задачи правят впечатление:

- задача 17 има ниска разграничителна сила и по двете теории. Едно обяснение е дадено в (Tsonev 2022);

- задачите са съсредоточени в средната третина на интервала; това означава, че липсват задачи за много-слабите и много-силните ученици;
- и задачите, и учениците имат сравнително добро „нормално“ разпределение;
- най-много задачи имат коефициент на трудност $b = -0.2$, докато най-големият брой ученици са с малко по-ниски способности;
- липсват ученици със способност 1,6, която е малко под най-високата измерена способност 1,8;
- както показаха изследванията с КТТ (Tsonev 2022) задачите с номера 1, 3, 6 и 7 са най-трудни, а задачите с номера 2, 10, 12 и 18 са най-лесни, което много добре се съгласува с резултатите от КТТ.

Разлики в двете диаграми:

- пиковите стойности за област София-град са две – първата е в средата на разглеждания интервал, а втората е в десния край, където са учениците с най-високи способности;
- графиките над и под реда θ/b се съгласуват по-добре за област Плевен, отколкото за област София;
- голяма част от учениците от област София имат способности, които надвишават максималното ниво на трудност на задачите $b = 0,4$.

Изводите, които могат да се направят, са, че в област София-град голяма част от учениците са отлично подготвени за решаването на задачите с избираем отговор, тези задачи не са ги затруднили в голяма степен и не биха били полезни, сами по себе си, за класиране след VII клас. За тази цел задачите с отворен отговор ще дадат повече възможности. За учениците в област Плевен може да се каже, че и само с тези 18 задачи те биха се диференцирали в много добра степен в гимназиите след VII клас.

6. Заключение

Две са основните цели на провежданите НВО по БЕЛ и математика – да подредят учениците след VII клас според техните постижения и да оценят доколко добре те са изучили учебния материал по тези два предмета.

В България изследването на тестовете се прави най-често чрез Класическата теория на тестовете. С помощта на методите на по-съвременната IRT може да се открият допълнителни характеристики, да се потвърдят или отхвърлят вече намерени свойства както на тестовите задачи, така и на способностите на учениците.

От направеното изследване върху резултатите от областите София-град и Плевен може да се каже, че:

- трите модела на параметризация според IRT добре представят характеристиките на задачите;
- моделът с два параметъра най-добре се съгласува както с емпиричните данни, така и с изводите от КТТ;
- в по-малките области на България, където конкуренцията при класирането за гимназиите е по-слаба, дори само първите 18 тестови задачи с избираем отговор могат да подредят учениците според техните постижения;
- за по-големите области тези задачи не са достатъчни за такава наредба;
- въпреки тези разлики, трудността на задачите се съгласува добре със способностите на учениците

В следващите изследвания на автора ще бъдат разгледани тестовите задачи със свободен отговор.

БЕЛЕЖКИ

1. Национално външно оценяване за VII клас, МОН, Retrieved 14.06.2023 from: https://web.mon.bg/upload/26732/NVO-MATH_7kl_18062021.pdf
2. jMetrik, Retrieved 14.06.2023 from: <https://itemanalysis.com/jmetrik-download/>
3. GeoGebra, Retrieved 14.06.2023 from: <https://www.geogebra.org/>

ЛИТЕРАТУРА

- БАНКОВ, К., 2002. Вероятностно моделиране за измерване на ученическите постижения, *Математика и информатика*, год. 45, кн. 4.
- ДЖАЛЕВ, Л., 2014. Приложимост на Класическата тестова теория и Теорията за отговор на тестов въпрос: преглед на литературата по въпроса. *Българско списание по психология*, 1, брой 1 – 3.
- HAMBLETON, R.K., SWAMINATHAN, H. & ROGERS, H.J., 1991. *Fundamentals of Item Response Theory*. Sage.
- McDONALD, R.P., 1999. *Test Theory: A Unified Treatment* (1st ed.). New York: Psychology Press.
- ЦОНЕВ, П., 2022. Някои изводи върху резултатите от националното външно оценяване по математика за VII клас, *Математика и информатика*, год. 65, кн. 6, с. 587 – 601.

REFERENCES

- BANKOV, K., 2002. Probabilistic Modeling for Measuring Student Achievement. *Mathematics and Informatics*, vol. 45, no. 4 [in Bulgarian].
- JALEV, L., 2014. Applicability of Classical Test Theory and Test Question Answer Theory: A Literature Review. *Bulgarian Journal of Psychology*. no. 1, pp. 1 – 3. [in Bulgarian]
- HAMBLETON, R.K., SWAMINATHAN, H. & ROGERS, H.J., 1991. Fundamentals of Item Response Theory. Sage.
- McDONALD, R.P., 1999. *Test Theory: A Unified Treatment* (1st ed.). New York: Psychology Press.
- TSONEV, P., 2022. Some conclusions on the results of the national external assessment in mathematics for grade VII. *Mathematics and Informatics*, vol. 65, no. 6, pp. 587 – 601. [in Bulgarian]

CHOOSING AN APPROPRIATE MODEL FOR INVESTIGATING THE TASKS OF SCHOOL IN MATHEMATICS AFTER GRADE VII ACCORDING TO IRT METHODS

Abstract. In the article the methods of Item Response Theory were used for analyzing the measuring qualities of the test tasks with a selectable answer from the National External Evaluation for the Grade 7 conducted in 2021 in the Sofia-city and Pleven districts. After agreement with the empirical data, the choice of the two-parameter model was justified using this methodology. A comparison was made between the obtained data and those from the Classical Test Theory.

Keywords: Item Response Theory; probabilistic modeling; multiple-choice test items; difficulty; discriminating power

✉ **Dr. Pavlin Tsonev, Assist. Prof.**

ORCID iD: 0000-0003-1289-3750

Department of Natural Sciences and Humanities
“Georgi Benkovski” Bulgarian Air Force Academy
Dolna Mitropolia, Bulgaria
E-mail: ptsonev@yahoo.com