СЪПОСТАВИМОСТ НА ОЦЕНЕНИТЕ СПОСОБНОСТИ СПРЯМО МОДЕЛИТЕ НА ITEM RESPONS THEORY*

Галин Г. Дръжков, Боряна Хр. Узунова-Димитрова

COMPARABILITY EVALUATED ABILITIES TO ITEM RESPONS THEORY MODELS

Galin Г. Drazhkov, Boryana Hr. Uzunova-Dimitrova

ABSTRACT: *Comparability between Characteristic Curve obtained by application One-Parameter model of Georg Rasch and Two-Parameter model of Alan Birnbaum of Item Response Theory, out on the same group of tested persons evaluated out on the same test built up out on the same studied material.*

KEYWORDS: *Classical Test Theory (CTT), Item Respons Theory (IRT), the difficulty parameter, the discrimination parameter.*

През последните години все повече навлизат тестовете, чрез които се оценяват придобитите и усвоени знания от обучаемите. За така използвания и предпочитан метод на оценяване се прилагат от една страна Класическата тестова теория, а от друга Теорията за отговор на тестови въпрос с алтернативни отговори за оценяване чрез латентни черти. Въпреки съществените различия между двете теории, те споделят някои общи теоретични конструктиви, каквито са характеристиките на въпросите – трудност, дискриминативна сила и налучкване на верния отговор.

Новата психометрична теория има редица теоретични предимства пред Класическата тестова теория. Спрямо параметрите на тестовите въпроси те се изразяват в прецизността на техните оценки, независимостта им от извадката, независимостта им една от друга. Тези характеристики, но в противоположен смисъл, се разглеждат като съществени недостатъци на Класическата теория.

Появата и развитието на IRT теорията и нейното налагане като основа на психологическите измервания, се разглежда от мнозина изследователи в оценяването [7]. Въпреки това емпиричните процедури на измерването в рамките на тази теория не се различават съществено от тези при Класическата теория. Най-често се разработва или се използва готов специализиран инструмент за измерване, съставен от множество въпроси, всеки от които е ориентиран към отделен елемент от съответната оценявана способност, която е пряко свързана с изследователския интерес. Отговорите на тестираното лице се оценяват дихотомично – като за правилен отговор се присъжда 1 точка, а за неправилен – 0 точки, т.е. използва се бинарното представяне на данните. Необходимостта от този начин на представяне на отговорите предопределя във висока степен обстоятелството, че предпочитаният вид въпроси в тестовото изпитване са въпросите от тип множествен избор със структуриран отговор, чиято форма съответства на това изискване.

Изчисленията се основават на база двумерна матрица от типа A=CxQ с емпирични данни, т.е. дихотомичните оценки на отговорите на тестираните лица. IRT се фокусира

* Научноизследователски проект №РД-08-113/2016 Модели и алгоритми за извличане на знания от големи данни, симулиране на невронни мрежи и оптимални учебни процедури - ръководител доц. д-р Найден Ненков, Шуменски университет "Епископ Константин Преславски", ФМИ, катедра "Компютърни системи и технологии".

върху отговорите на тестираните лица на всеки отделен въпрос и въз основа на тази информация, чрез съответния математически апарат, дава възможност за оценка на ненаблюдаемите индивидуални способности.

IRT съществува под формата на различни модели и може да се разглежда по-скоро като обща теоретична концепция за обяснение на латентните променливи. Теорията включва разнообразни модели на връзката между изпълнението на тестовите въпроси от тестираните лица и техните способности чрез прилагане на вероятностни подходи. Това прави измерването в рамките на IRT теорията базирано на модела (model-based measurement) [14].

В IRT теорията понятията като знания, умения и компетенции в различните предметни области са заменени с използваното от теорията понятие „способност“ (ability), което се явява основната латентна черта. [1].

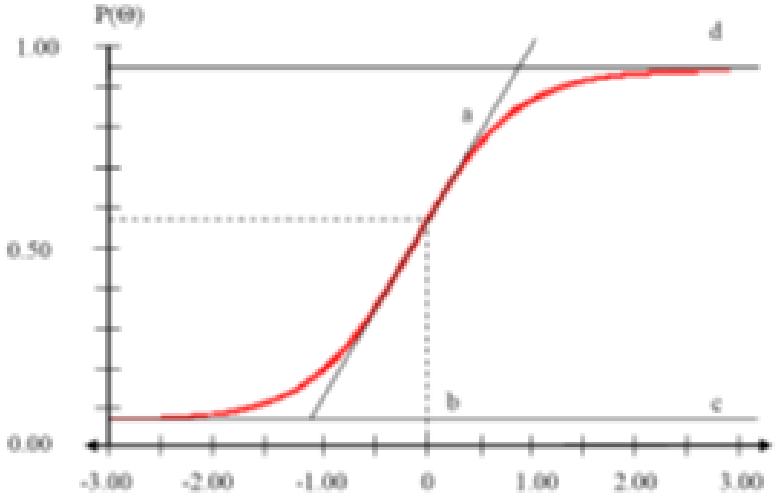
Друг важен елемент от използваните модели на IRT теорията са характеристичните криви. [5, 6, 9, 13, 14].

Подобно на психологическите черти, способностите на оценяваното лице в определена област могат да бъдат разгледани като латентна черта. Тази латентна черта може да бъде представена като едномерно пространство (континуум), а всеки тестиран, който притежава такива способности – като точка на този континуум. Позицията на всеки оценяван на континуума се определя от равнището (количеството) на неговите способности, което може да бъде изразено чрез определена числова стойност (*ability score*), обозначавана с Θ (тита). Тъй като в общия случай различните индивиди се различават по своите способности, техните точки биха били позиционирани на различни места на континуума, т.е. биха били асоциирани с различни числови стойности на Θ.

Съгласно втората теорема на П. Супес и Дж. Зинес [3], съвкупността от тестираните лица, заедно с приписаните им числови стойности, отразяващи равнището на техните индивидуални способности, образуват скала на компетентността (*ability scale*), която най-често се обозначава също с Θ. От теоретична гледна точка скалата на латентните способности е неограничена отляво и отдясно, поради което скаловите стойности на индивидите (индивидуалните Θ*i*) могат да варират от отрицателна до положителна безкрайност.

В общия случай, отделните тестирани лица се различават по своите способности и поради това биха отговорили правилно на даден въпрос с различна вероятност $P(\Theta)$, която може да варира в границите $0.00 \leq P(\Theta) \leq 1.00$. Вероятността се изменя плавно, а не скокообразно. Тя би била сравнително по-висока при тези лица, които имат по-големи способности и съответно по-ниска при онези, които имат по-малки способности. IRT разглежда вероятността $P(\Theta)$ на тестираните лица да определи верния отговор на даден въпрос като функция от техните способности. Графичното изображение на тази функция е представено на фиг. 1, характеристичната крива на въпроса, която има формата на плавна, монотонно нарастваща S-образна крива, неограничена отляво и отдясно.

Ординатите на характеристичната крива в левия край на континуума на способностите имат ниски стойности (лицата със слаби способности, биха отговорили вярно на въпроса с вероятност, близка до 0.00), издига се плавно нагоре (лицата със средни способности биха дали верен отговор с вероятност около 0.50), за да премине към десния край на скалата с ординати, близки до 1.00 (лицата с високи способности, които почти сигурно биха дали верен отговор).



Фиг. 1. Общ вид на характеристична крива на тестови въпрос

Тази форма на характеристичната крива съответства на допускането, че с нарастването на способностите нараства и вероятността за верен отговор.

S-образната крива от фиг. 1 се характеризира с 5 свойства: (1) наклон на кривата в средната ѝ част, (2) позиция спрямо хоризонталната ос (скалата на способностите), (3) долна хоризонтална асимптота, (4) горна хоризонтална асимптота и (5) симетричност. Тези няколко свойства са достатъчни за описанието на всяка характеристична крива, следователно и на всякакъв тип връзки между способностите на лицата и изпълнението им на тестовите въпроси.

Подобно на Класическата теория и IRT борави с понятията "суров тестов бал" (*raw test score*) и "действителен бал" (*true score*). В рамките на модерната теория те имат аналогични значения – съответно на сума от дихотомично кодираните отговори на тестираното лице на въпросите от теста и на средна стойност на множеството от наблюдаваните тестови балове на това лице, получени при независими тестирания.

Различен е обаче начинът за определяне на действителния бал, който се използва в IRT. Съгласно подхода, предложен от Д. Лоули, формулата за определяне на действителен бал е следната [12, цит. в 5]:

$$(1) \quad TS_i = \sum_{j=1}^{k} P_j(\Theta_i),$$

където:

TSi - действителен бал на тестираните лица с равнище на компетентност Θj;

j - пореден номер на тестови въпрос;

k - брой на въпросите в теста;

Pj(Θ*i*) - вероятност от правилен отговор на *j*-тия тестови въпрос от тестирано лице с равнище на способност Θ*i*.

Pj(Θ*i*) се определя в зависимост от конкретния модел на характеристичната крива на дадения въпрос. Както се вижда, съгласно подхода на Д. Лоули действителният бал за дадено равнище на компетентност се определя като сума от вероятностите за верен отговор на всички въпроси в теста. Тъй като вероятността за даден въпрос се изменя в границите $0.00 \leq Pj(\Theta) \leq 1.00$, то действителният бал може да приеме стойности в границите $0.00 \leq TS \leq k$.

Кривата, която отразява функционалната връзка между скалата на способностите Θ и действителния бал *TS*, е характеристичната крива на теста. Тя има форма на монотонно нарастваща функция, която не се задава чрез конкретен математически израз, не се описва чрез параметри и поради това нейната форма е конкретна за всеки тест. Обикновено тя е S-образна, подобна на характеристичната крива на въпросите съставящи теста. По подобен начин формата на характеристичната крива не зависи от разпределението на тестираните лица на скалата на способностите. При прилагането на модел от IRT теорията, левият ѝ край клони неограничено към 0.00, когато равнището на компетентност клони към -∞, а десният ѝ край – към *k*, когато равнището на компетентност клони към +∞.

Обект на настоящото изследване е съпоставимостта между получените характеристични криви, получени при прилагането на еднопараметричния модел на Георг Раш и двупараметричния модел на Бирнбаум от IRT теорията (Item Response Theory) върху една и съща група от тестирани лица, върху един и същи изучаван материал, които способности са оценени посредством един и същи тест съдържащ еднакви тестови въпроси от тип множествен избор с предоставени отговори, от които само един е верен.

Използваните логистични функции са функции от „четвърто поколение". Чрез тях се моделира връзката между скалата на латентните способности и представените променливи. Тя е предпочетена заради по-високата си адекватност спрямо емпиричните данни и поради по-лесната си изчислимост. Заради тези нейни качества днес тя е сред най-често използваните вероятностни модели за представяне на тази връзка.

Общият вид на логистичната функция е следният:

(2) $$P(t) = \frac{1}{1 + e^{-t}}$$

където:
P(*t*) - зависима променлива;
t - независима променлива;
e - константа, основа на натуралния логаритъм, с приблизителна стойност 2.718.

Еднопараметричният логистичен модел известен още като модел на Раш, включва само един параметър и това е трудността на въпросите *b*. Уравнението, изразяващо функционалната връзка между латентната променлива и вероятността от верен отговор, има следния вид:

(3) $$P(\theta) = \frac{1}{1+e^{-(\Theta - b_j)}};\ -\infty \le b \le +\infty$$

където:
Pj(Θ) - вероятност за правилен отговор на въпрос *j*
b – трудността на задачата, която е различна за всяка различна задача;
θ - способността на тестираното лице да даде верен отговор на задачата.
e = 2.718 - основа на натуралния логаритъм (неперово число)

Включването в модела само на параметъра на трудността *b* има своите основания, тъй като това е единственият параметър, който е разположен на скалата на способностите, с която се образува смесен континуум. Оттук теоретичните граници на изменение на параметъра *b* са същите, както и тези на Θ (При въпроси с нулева дискриминативна сила стойността на *b* е неопределена.): -∞ ≤ *b* ≤ +∞, но практически рядко надхвърлят -3.00 ≤ *b* ≤ +3.00.

Двупараметричният логистичен модел известен още като модел на Бирнбаум добавя още един параметър – дискриминативната сила на въпроса *a*, с което горното уравнение добива следния вид:

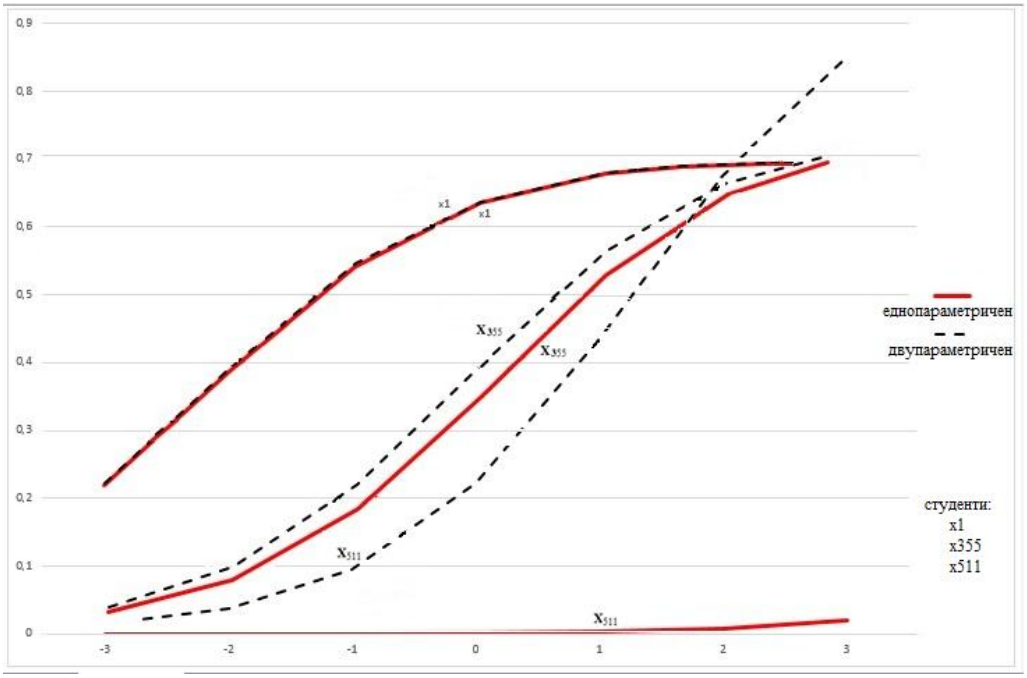
(4) $$P(\theta) = \frac{1}{1+e^{-a(\Theta-b_j)}};\ -\infty \le a \le +\infty,\ -\infty \le b \le +\infty$$

където:
aj - дискриминативна сила на въпрос *j*.

Поради S-образната форма на характеристичната крива нейният наклон се променя с изменение на равнищата на компетентност. Той има максимална стойност в инфлексната ѝ точка, в която равнището на компетентност е равно на трудността на въпроса. Поради това дискриминативната сила на въпроса е свързана не толкова с общата форма на кривата, колкото с нейния наклон в точка $\Theta = b$. Действителната стойност на наклона в тази точка е *a*/4, но възприемането на *a* като наклон на характеристичната крива е приемлива апроксимация, която прави интерпретацията на този параметър по-лесна (Baker, 2001).

Добавянето на параметъра *a* води до това, че в общия случай всеки въпрос е представен от характеристична крива с различен наклон. Теоретичните граници на изменение на дискриминативния параметър са $-\infty \le a \le +\infty$, но практически той е ограничен между $-2.80 \le a \le +2.80$ [8, 14].

На фиг. 2. са показани получените характеристични криви на три тестирани лица, с различен тестови бал – минимален, среден и максимален, получени чрез използването на еднопараметричния и двупараметричния модел. Наблюдава се, че кривите се различават само по местоположението си, т.е. кривите са успоредни и не пресичащи се както е и характерно за характеристичните криви на тези модели от IRT теорията. За построяването на характеристичните криви е необходима подготвеността (θ) на тестирания, вероятността му (P(θ)) да даде правилен отговор на съответния тестови въпрос и дискриминативната сила на използвания тест.



Фиг. 2. Характеристични криви на трима студента x1, x355 и x511 спрямо модела на Раш и модела на Бирнбаум, дискреминативна сила на теста a=0.75

Анализирайки получените резултати, използвайки модела на Раш и модела на Бирнбаум, представени на фиг. 2, можем да направим следните изводи:

- **Наблюдава се, че характеристичните криви отразяващи слабите резултати на тестираното лице се припокриват с незначителни разлики, което от своя страна говори, че дори и включването на втория параметър, а именно дискриминативната сила на теста, не променя крайния резултат за тестираното лице.**
- **Характеристичните криви отразяващи средните резултати на тестираното лице, се наблюдава че те леко се отдалечават една от друга, като тази отговаряща на еднопараметричния модел се намира под кривата съответстваща на двупараметричния, което говори от своя страна, че измерванията направени чрез модела на Бирнбаум отчитат по-точни резултати, а също така и не променят характерната форма на характеристичната крива.**
- **Кривите отразяващи резултатите на отлично представилото се тестирано лице, се наблюдават съществени различия, при еднопараметричния тя е почти успоредна на абсцисата и е загубила характерната си форма, докато при двупараметричния модел кривата е запазила своето характерно представяне въпреки наподобяването ѝ на овал.**
- **При многократни реални експерименти, статистическите данни могат да се използват за «напасване» ("Within population item-fit") на входните параметри (критериите) към изискванията на моделите. Лесните (които се удовлетворяват от всички) и трудните (които никоя работа не удовлетворява) критерии могат да бъдат изключени.**

От графиката ясно се наблюдава, че характеристичните криви на студент x1 спрямо двата модела са чувствително изпъкнали, което говори за слабо подготвен студент, който не е усвоил учебния материал и не демонстрира необходимите проверявани знания. Неговата оценка би била незадоволителна, т.е. Слаб 2. Характеристичните криви на студент x355 имат стандартна форма традиционна за двата модела. Тълкуването им би било следното, че студента е усвоил материала в общи линии, но не изцяло, ясно се забелязва, че половината от кривата е изпъкнала, а другата – вдлъбната, т.е. знанията му са сравнително задоволителни, т.е. оценката би трябвало да е Добър 4. Характеристичните криви на студент x511 за еднопараметричния модел е вдлъбната и приблизително успоредна на оста на абсцисата, а за двупараметричния модел характерната S – образна форма наподобява повече овал, което говори за един добре подготвен студен, който е усвоил цялостно учебния материал и е дал верни отговори на тестовите въпроси, т.е. неговата оценка би трябвало да е Отличен 6 [4].

В рамките на даден модел, характеристичната крива има една и съща форма за всички айтеми, независимо от броя на параметрите, които описват теста. Прилагането на тези методи почти винаги води до отхвърлянето на по-малко или повече айтеми, които не съответстват на избрания модел, независимо от това дали той е 1- или 2-параметричен. Това означава, че по правило айтемите в един и същи тест могат да бъдат описани с характеристични криви в рамките на различни модели.

IRT въвежда характеристична крива на теста, която отразява функционалната връзка между скалата на способностите Θ и действителния бал на тестираните лица. Тя играе важна роля в етапа на интерпретиране и представяне на резултатите от теста, като служи за преобразуване на скалата на способностите в скала, която е свързана с тестовия бал в Класическата теория и поради това е лесно разбираема за потребителите на тестовите резултати.

Измерванията чрез моделите от IRT теорията, има своето логично обяснения – това е теоретична рамка, която осигурява по-добре от останалите теории постигането на основната цел на измерването, а именно получаване на такива оценки на личностовия

параметър (позицията на изпитваното лице и на континуума Θ), които да бъдат неизместени, състоятелни и надеждни [2, 9, 10, 11].

ЛИТЕРАТУРА

1. **Анастази**, А., Урбина, С. (2001). Психологическое тестирование. Санкт-Петербург: Питер, стр: 516-517.
2. **Гласс**, Дж, Стэнли, Дж. (1976). Статистические методы в педагогике и психологии. Москва: Прогресс.
3. **Суппес**, П., Зинес, Дж. (1967) Основы теории измерений. В: Психологические измере-ния. Под ред. Л. Д. Мешалкина. Москва: Мир, сс. 9-110.
4. **Узунов-Димитрова,** Б., Методи и алгоритми на изкуствения интелект за обективизация при оценяването във висшите училища. Шумен 2016.
5. **Baker**, F. B. (2001). The basics of Item response theory. ERIC Clearinghouse on Assessment and Evaluation, 2-nd ed.
6. **DeVellis**, R. F. (2003). Scale development: theory and applications. Applied Social Research Methods Series, Vol. 26. Thousand Oaks, Ca.: Sage Publications, Inc., 2-nd ed.
7. **Embretson**, S. E., Reise, S. P. (2000). Item response theory for psychologists, pp. 13.
8. **Embretson**, S. E., Hershberger, S. L. (eds.) (1999). The new rules of measurement: What every educator and psychologist should know. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
9. **Hambleton**, R. K., Swaminathan, & H., Rogers, H. J. (1991). Fundamentals of Item response theory. Newbury Park, Ca.: Sage Publications, Inc.
10. **Harvey**, R. J. (2003, april). Applicability of binary IRT models to job analysis data. In Meade, A. (Chair), Applications of IRT for measurement in organizations. Symposium presented at the Annual conference of the Society for industrial and organizational psychology, Orlando.
11. **Hulin**, C. L., Drasgow, F. & Parsons, C. K. (1983). Item response theory: Applications to psy-chological measurement. Homewood, IL.: Dow Jones-Irwin.
12. **Lawley**, D. N. (1943, January). On problems connected with item selection and test construc-tion. In: Proceedings of the royal society of Edinburg. Section A. Mathematical and Physical Sciences, Vol. 61, Issue 03, pp. 273-287
13. **Lord**, F. M. (1980). Applications of Item response theory to practical testing problems. Hills-dale, NJ: Lawrence Erlbaum Associates, Inc.
14. **Weiner**, I. B., Freedheim, D. K., Schinka, J. A., & Velicer, W. F. (2003). Handbook of Psychology: Research methods in psychology. NJ: John Wiley & Sons, Inc.