МЕТОДОЛОГИЯ ЗА СЪЗДАВАНЕ НА ОБЕКТИВНИ И БАЛАНСИРАНИ ЛИНЕЙНИ ТЕСТОВЕ БАЗИРАНА НА ТЕОРИЯТА IRT

Иван Господинов, Борислав Славов, Стефан Филипов

Химикотехнологичен и Металургичен Университет – София,
Катедра „Програмиране и използване на компютърни системи”,
gospodinov.ivan@yahoo.com

***Резюме**: Оценяването на когнитивните способности на индивида е важна част от образователния процес. Главната цел на всяка тестова теория е да създаде тестове, които да са обективни по целия диапазон на способности на тестваната аудитория. Или казано с други думи - крайната оценка на всеки индивид да отговаря на знанията му, независимо колко големи или малки са те. В днешно време масово се използват две теории - Класическата Теория на Тестването (КТТ) и IRT (Item Response Theory). Настоящата статия разглежда някои от ограниченията на КТТ за създаване на обективни и балансирани линейни тестове и демонстрира как тези ограничения могат да бъдат отстранени като се приложи по-съвременната IRT.*

***Ключови думи**: обективен тест, балансиран тест, линеен тест, IRT*

1. Въведение

Класическата Теория на Тестването се прилага успешно от 100 години. В основата на тази теория заляга постулата, че оценката на способността на индивида, придобита от администриране на един тест, е равна на истинските му способности плюс някаква несистемна грешка. Тази грешка би била нула само ако тестът бъде даден на индивида безброй пъти и резултатите бъдат осреднени. Това е невъзможно, но е възможно например един тест да бъде даден на много индивиди или да бъдат конструирани множество паралелни, аналогични теста, които да се дадат на един и същи индивид [2],[7]. КТТ има редица недостатъци [1],[3], които затрудняват използването й. Два от тези недостатъка са особено важни за настоящия анализ.

Първият недостатък е, че КТТ постулира, че стандартната грешка на оценка на способностите при тестване е еднаква за всички тествани индивиди, което не е истина в общия случай.

Вторият недостатък е, че при КТТ характеристиките на тествания индивид (напр. нивото на знанията му) и характеристиките на теста (напр. трудността на въпросите) са свързани по цикличен начин – за да получим информация за нивото на знанията на индивида се нуждаем от трудностите на въпросите и обратно. Поради това е много трудно да се формулира КТТ модел, който да

изрази едните характеристики като функция на другите. Ако имахме такъв модел при зададен фиксиран набор от тестови въпроси бихме могли да определим информацията, които те носят, и оттам бихме могли да определим грешката при измерването за всеки един от тестваните индивиди. Тази грешка определя обективността на теста, както ще бъде дискутирано по-надолу. КТТ няма достъп до информацията, която носят отделните въпроси в теста, а само до информацията, която носи целия тест [10], което е голяма загуба на детайлност.

IRT е съвременна теория [4-6],[8-9], която за разлика от КТТ се фокусира върху отделните тестови въпроси. При тази теория, при използване на 3-параметричен модел, вероятността P за правилен отговор на един въпрос е функция на нивото на способност (ability) θ на тествания индивид и на три параметъра:

$$P(\theta) = c + \frac{1-c}{1+\exp[-a(\theta-b)]}, \qquad (1)$$

където *a* е дискриминативността на въпроса, *b* е неговата трудност, а *c* е вероятността за случайно отгатване на правилния отговор (виж фиг. 1а). Ако възможните отговори на въпросите са в така наречени multiple-choice формат параметърът *c* се намира директно докато параметрите *a* и *b* се получават при статистическа обработка на отговорите на целева група или на базата на експертни знания. Способността θ се измерва в единици z-score и обикновено се изменя в интервал [-3,+3]. Стойност за θ =-3 означава много ниска способност на индивида, докато стойност за θ =3 означава много висока способност. Ако превърнем z-score в оценка от 2 до 6 то -3 означава оценка 2, а +3 означава оценка 6. Едно от предимствата на IRT е, че теорията е способна да отчете вероятността *c* за отгатване отговора на въпроса. Забележете, че при IRT характеристиката на индивида θ и характеристиката на въпроса *b* „лягат” на една и съща ос и се изменят в едни и същи граници, което прави възможно получаването на отделна оценка за θ от отговора на всеки отделен тестови въпрос. Това е и основното предимство на IRT пред КТТ.

Параметърът *b* отговаря на $P(\theta)$ със стойност 0.5+c/2. От което следва, че ако на един индивид се зададат достатъчно много въпроси с трудност *b* и той отговори на около *половината* от тях (плюс c/2) тогава неговата способност θ е равна на *b*. Или с други думи, този индивид знае за оценка *b*. Това се разминава с конвенционалната представа за трудност на един въпрос, при която например студент за 6 трябва да може да отговори на (почти) *всички* въпроси за 6. От уравнение (1) можем да получим израз за информацията $i(\theta,b)$ която носи отговорът на един въпрос с трудност *b* зададен на индивид със способност θ [12],[11]:

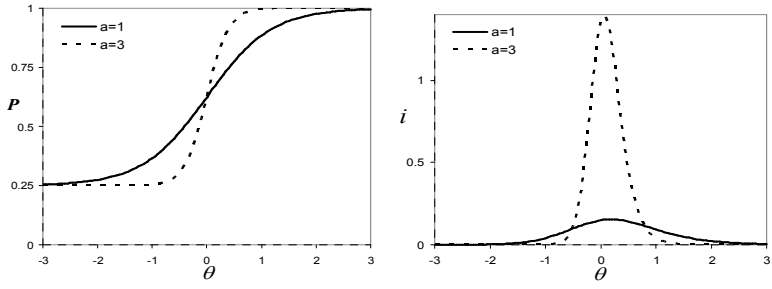
$$i(\theta,b) = \frac{[P'(\theta)]^2}{P(\theta)[1-P(\theta)]} = a^2 \frac{[1-P(\theta)][P(\theta)-c]^2}{P(\theta)[1-c]^2}, \quad (2)$$

където $P'(\theta)$ е производна по θ, $[1\text{-}P(\theta)]$ е вероятността за грешен отговор на въпроса, а $P(\theta)$ се дава от уравнение (1). Информацията $i(\theta,b)$ има максимум около θ =*b* тъй като $P(\theta)$ има инфлексия там, т.е. един въпрос е толкова по-информативен, колкото трудността му е по-близо до истинската способност на индивида. От $i(\theta,b)$ директно можем да получим стандартната грешка на оценяване [9] на θ за един въпрос:

$$\sigma_i(\theta,b) = \frac{1}{\sqrt{i(\theta,b)}}, \quad (3)$$

Уравнение (3) ни дава строг математически критерий за точност на оценяване на индивида, който може да бъде използван за конструиране на балансиран и обективен тест. Грешката $\sigma_i(\theta,b)$ показва каква е точността в определянето на оценката на способността θ на *един* индивид отговорил на *един* въпрос с параметри *a*, *b* и *c*. Забележете елегантния начин по който IRT прави връзка между грешката на оценяване, информацията и вероятността за правилен отговор на един тестови въпрос. При ССТ такава връзка не може да бъде изведена.

Фигура 1а показва графики на $P(\theta)$ а фиг. 2б показва графики на $i(\theta,b)$ за два въпроса с еднакви *b* и *c* (*b*=0, *c*=0.25) но с две различни дискриминативности *a*=1 и *a*=3 съответно. При въпросът с по-малко *a* графиките са „по-разлети”. От тях се вижда, че при високо *a* с малко въпроси можем много точно да определим способностите само на индивиди, чиито θ са много близо до *b*. При ниско *a* можем да определим способностите на индивиди в по-широк диапазон около *b* но за целта са ни нужни повече въпроси за да наваксаме падането на нивата на $i(\theta,b)$. С фиксирано количество въпроси при високо *a* се обхващат по-малко индивиди, но с по-висока точност на оценяване, докато при ниско *a* се обхващат повече индивиди, но с по-ниска точност на оценяване.



Фигура 1. P и i като функции на θ за b=0 и c=0.25 за две различни стойности на a.

От изложеното дотук се вижда, че за разлика от КТТ, IRT ни дава достъп до информацията, която ни носи всеки един въпрос от теста за индивид с всякаква способност. Това ни позволява да администрираме тест, при който да направим оценка за θ на индивида още след първия въпрос и да адаптираме трудността на следващия въпрос според тази оценка – ако индивидът отговори правилно му се дава по-труден въпрос, ако отговори грешно му се дава по-лесен въпрос. Този принцип заляга в алгоритмите за Компютърно Адаптирано Тестване (CAT – Computer Adaptive Testing), които използват IRT. Тези алгоритми са много сигурни, ефективни и точни поради което са бъдещето на тестването. При адаптивните тестове дължината и съдържанието на теста не са фиксирани и обективността е гарантирана по много елегантен математически начин. Един недостатък на тези алгоритми е, че CAT тестове могат да се администрират само на компютър, което не винаги е удобно. Често пъти това е невъзможно при изпитване на големи потоци от студенти от по 50 до 250 човека. Тогава се налага да се администрира така наречения хартиен (paper-and-pencil) тест. Класическият хартиен тест е така наречения линеен тест. Този вид тест е с фиксирани въпроси, които са еднакви за всички тествани индивиди. Линейният тест не позволява адаптивност, независимо дали използваме КТТ или IRT. При конструирането на такъв тест е от изключителна важност да се осигури обективност. Информацията, показана в (2), и стандартната грешка при определянето на крайната оценка, показана в (3), са естествените критерии за обективност. При всеки линеен тест е много важно да се гарантира обективно оценяване на индивидите по целия диапазон на способности. За целта е необходимо внимателно балансиране на съдържанието на теста, например балансиране на въпросите по трудност, дискриминативност, теми и т.н.

При съвременната генерация на тестове въпросите се „изтеглят” от банка с въпроси, параметрите (*a*,*b*) на които са предварително оценени. При балансирането на един линеен тест се питаме следните важни неща: Въпроси с какви параметри (*a*,*b*) да изтеглим така, че всеки индивид да е еднакво точно оценен? Какъв е минималният брой въпроси, който би ни гарантирал, че грешката при оценяването не е по-голяма от някаква максимално допустима грешка? Авторите на тази статия предлагат методология, която да даде отговори на тези въпроси. В последващия анализ е дефинирано понятието обективен тест и са изведени математически изрази помагащи ни да определим дали един тест е обективен.

2. Методология

Отговорът на един индивид със способност θ на един тестови въпрос с трудност *b* носи на администратора на теста информацията $i(\theta,b)$, която се дава от уравнение (2). Информацията е адитивна величина. Тоталната

информация за θ на един индивид, получена от отговорите му на всички въпроси в теста е сума от информациите на отделните въпроси или:

$$I(\theta) = \int_{-3}^{3} i(\theta, b) g(b) db, \quad (4)$$

където $g(b)$ е разпределението на въпросите в теста по трудност, което трябва да спазва:

$$\int_{-3}^{3} g(b) db = N_q, \quad (5)$$

където N_q е броя въпроси в теста. Стандартната грешка при оценяването на един индивид базирана на отговорите му на всички въпроси е:

$$\sigma_I(\theta) = \frac{1}{\sqrt{I(\theta)}} \quad (6)$$

Забележете, че $\sigma_I(\theta)$ е базирана на информацията получена от *всички* въпроси (откъдето идва и индекса I) докато $\sigma_i(\theta, b)$ в уравнение (3) е базирана на информацията получена от *един* въпрос (откъдето идва и индекса i). Двете са свързани посредством уравнение (4) в което участва разпределението $g(b)$. Информацията получена от отговорите на всички студенти на всички въпроси е:

$$I_{total} = \int_{-3}^{3} I(\theta) f(\theta) d\theta, \quad (7)$$

където $f(\theta)$ е разпределението на тестваните индивиди по способности θ. Средната информация от целия тест за един тестван индивид е:

$$I_{ave} = \frac{I_{total}}{N_s}, \quad (8)$$

където е N_s е броя тествани индивиди. За разпределението $f(\theta)$ е в сила:

$$\int_{-3}^{3} f(\theta) d\theta = N_s \quad (9)$$

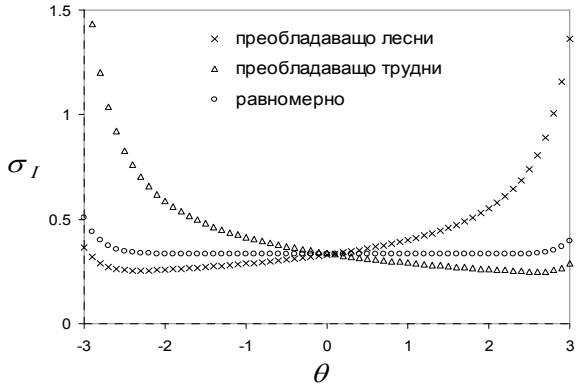
Средната стандартна грешка на оценяване за един индивид е:

$$\sigma_{ave} = \frac{1}{\sqrt{I_{ave}}} \quad (10)$$

При анализа на обективността на един линеен тест важната характеристика е $\sigma_I(\theta)$, показана в (6), защото тя ни дава грешката на оценяване за един индивид със способност θ на базата на отговорите му на всички въпроси в теста. Следователно, *един напълно обективен тест*

трябва да има еднакви стойности на $\sigma_I(\theta)$ *за всеки тестван индивид, независимо от способността му* θ. Един тест е толкова по-обективен колкото по-равномерно разпределена е грешката на оценяване по целия интервал от способности. Естествено, предварително трябва да сме се убедили, че всеки въпрос в теста има коректни параметри *a*, *b* и *c*, което не е тема на настоящия анализ.

За да определим обективността на един тест е нужно да зададем разпределението $g(b)$ и да видим дали това разпределение ще ни даде константно $\sigma_I(\theta)$ в целия диапазон на θ. За целта използваме уравнения (1), (2), (4), и (6) - тоест заместваме $P(\theta)$ от (1) в (2) и полученото $i(\theta,b)$ заедно със зададеното $g(b)$ заместваме в (4), откъдето получаваме $I(\theta)$, което на свой ред заместваме в (6). Полученият в (4) интеграл не може да бъде решен аналитично, поради което задачата е решена числено по следния начин: Способностите θ и трудностите *b* са изменяни в интервала [-3,+3] със стъпка $\Delta\theta$ =0.05 и Δb=0.05, съответно. За всяка стойност на θ произведението $i(\theta,b)g(b)\Delta b$ е сумирано по целия диапазон на *b* и полученото $I(\theta)$ е използвано за намирането на $\sigma_I(\theta)$.



Фигура 2. Грешката при оценяване като функция на способностите на индивидите за три разпределения на въпросите по трудност.

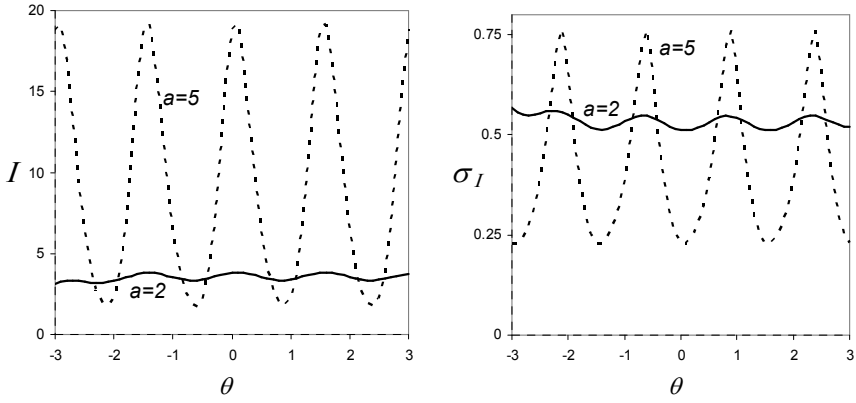
Фигура 2 показва числено получени графики за σ_I като функция на θ за три различни непрекъснати разпределения $g(b)$ – линейно намаляващо с преобладаващи лесни въпроси, линейно увеличаващо с преобладаващи трудни въпроси и равномерно разпределение по трудност. Резултатите са получени за *a*=5 и *c*=0.25. Равномерното разпределение е най-обективно, защото дава хоризонтална графика (освен близо до краищата си). Този резултат не е изненадващ. Интуицията подсказва, че ако искаме обективно да

тестваме група от индивиди трябва да дадем еднакво количество въпроси за всички способности, независимо какво е разпределението $f(\theta)$ - то не участва в уравнения (1) до (6).

Според IRT, информацията при тестване на индивид със способност θ, се носи предимно от въпроси с трудност b равна на θ, но също така и от въпроси с трудност близка до θ. При фиг. 2в по-ниската точност на оценяване (по-високата грешка) в краищата на интервала за θ произтича от това, че за тези гранични стойности информация „идва” само от едната страна. Това лесно може да бъде поправено като към теста се включат въпроси с допълнителни трудности -3,5 и 3,5 съответно. Това до голяма степен ще изравни стойностите на σ_I в целия интервал за θ от -3 до +3. При фиг. 3в по-голямата неточност в левия отколкото в десния край се дължи на ненулевата стойност на c, която прави формата на графиката асиметрична.

Тъй като $g(b)$ се конструира посредством изтегляне на въпроси от банка с въпроси с дискретни стойности на a и b то не е възможно да се получим непрекъснато и равномерно $g(b)$. Възниква въпросът: Кое е най-подходящото дискретизиране на $g(b)$? Най-удобно е да се изберат три до пет категории въпроси с постепенно нарастващи трудности, например въпроси с трудност 2; 3; 4; 5 и 6, които превърнати в z-score може да са -3; -1,5; 0; 1,5; 3. Често пъти разполагаме само с въпроси в три категории – леки, със средна трудност и трудни. При дискретизиране на $g(b)$ възниква друг въпрос: Могат ли тези дискретни категории на трудност да оценят студенти със способности намиращи се помежду тях – например 2,5; 3,5; и т.н.? Ключът към отговора на този въпрос е избирането на подходяща дискриминативност a, както показва фиг. 1. Въпроси с ниско a са подходящи за „попълване” на празнините между отделните категории трудности, но носят по-малко тотална информация. Това трябва да се компенсира с по-голям брой въпроси. Както се вижда от (4), (5), и (7) по-големия общ брой въпроси N_q ще доведе до увеличаване на тоталната информация и съответно понижаване на грешката σ_{ave}. Въпроси с високо a няма да успеят добре да „попълнят” празнините, въпреки че ще носят повече информация в другите z-score области.

Фигура 3 показва числени решения за две различни стойности на a за информацията $I(\theta)$ (фиг. 3а) и за грешката $\sigma_I(\theta)$ (фиг. 3б) за тест с въпроси дискретизирани в пет категории по трудност -3; -1,5; 0; 1,5; 3 с по 4 въпроса във всяка категория. Стойността a=5 довежда до недопустимо ниски информационни нива и съответно големи грешки в областите между категориите. По-малко дискриминативна стойност a=2 довежда до много по-малки флуктуации на информацията/грешката, тоест до много по-равномерно разпределение на информацията/грешката по целия интервал от способности на индивида θ.

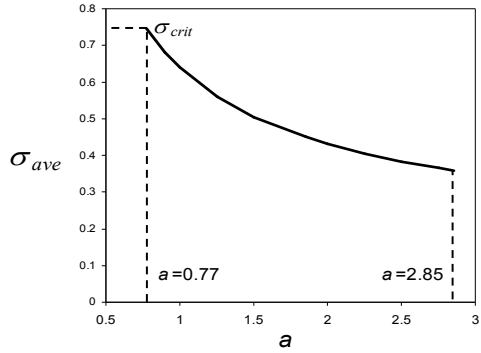


(а) (б)

Фигура 3. Числени резултати за $I(\theta)$ и $\sigma_I(\theta)$ при две различни стойности за a при тест с дискретизация на въпросите по категории на трудност.

От изложеното дотук се вижда, че за създаването на обективен тест не е достатъчно да конструираме равномерно разпределение $g(b)$ с дискретизирани категории въпроси по трудност. При всеки тип дискретизация е нужен правилен избор за параметъра *a*. Задачата за намиране на $\sigma_I(\theta)$ може да бъде решена числено за произволно разпределение $g(b)$, такова, при което параметрите *a*, *b* и *c* са различни за всеки въпрос.

Изискванията при създаване на линеен тест не са само към неговата обективност. Нужно е не само тестът еднакво точно да определя стандартната грешка за всеки индивид независимо от способността му, но и тази грешка да бъде по-малка от някаква предопределена стойност σ_{crit}. Тоест необходимо е $\sigma_{ave}<\sigma_{crit}$. При оценяване по шестобалната система половинките се закръгляват на по-високата оценка поради което грешката по шестобалната скала може да бъде най-много 0,5. Превърнато в z-score скала това прави σ_{crit}=0.75. При фиксирано $g(b)$ грешката σ_{ave} зависи от *a* и от броя въпроси N_q. За намиране на σ_{ave} е нужно да се укаже разпределението $f(\theta)$. Често пъти не разполагаме с такова. Тогава го приемаме за равномерно, тоест приемаме, че способностите на индивидите са равномерно разпределени в сегмента [-3,3].



Фигура 4. σ_{ave} като функция на a в интервал на стойности, които водят до обективен и точен тест

Фигура 4 показва зависимостта на σ_{ave} от *a* при фиксиран N_q=30. Резултатът е получен за $g(b)$ дискретизирано в пет категории по трудност -3; -1,5; 0; 1,5 и 3 и равен брой въпроси във всяка категория и равномерно $f(\theta)$. *Критерият за обективност* е флуктуациите в $\sigma_I(\theta)$ да са в рамките на ±15% от σ_{ave}. *Критерият за точност* е $\sigma_{ave}<\sigma_{crit}$, където σ_{crit}=0,75. От фигурата се вижда, че σ_{ave} намалява с нарастването на *a*. Намерен е интервал от стойности за *a* [0,77; 2,85], които водят до създаване на обективен и едновременно с това точен тест. Лявата граница на този интервал е стойността на *a*, в ляво от която е нарушено изискването за точност. Дясната граница на интервала е стойността на *a*, в дясно от която е нарушено изискването за обективност. Оптималната стойност на *a* е стойността в дясната граница защото тя отговаря на най-ниско σ_{ave}, тоест най-висока точност на теста (за сметка на максимално допустимия компромис с обективността).

Последно е нужно да разгледаме зависимостта на σ_{ave} от N_q при фиксирано *a*. При анализ на уравнения (4) и (5) се вижда, че $I(\theta)$ зависи линейно от N_q. От (6) тогава следва, че σ_I се изменя както 1/корен квадратен от N_q.

3. Заключение

Изведени бяха уравнения за информацията и грешката при оценяване на линеен тест използвайки теорията IRT. Дефиниран бе критерий за обективност на линеен тест и бяха показани резултати за обективността на няколко вида разпределения на тестове по трудност на въпросите. Дефиниран бе интервал на дискриминативност на въпросите в теста при който един линеен тест е обективен и точен в рамките на предварително зададени допуски.

Литература

1. Hambleton, R., Jones, R. (1993): *Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development.* Educational Measurement: Issues and Practice, 12(3), 3847
2. Bechger, T., Gunter, M., Huub, H., & Béguin, A. (2003): *Using classical test theory in combination with item response theory.* Applied psychological measurement, 27(5), 319-334.
3. Crocker, L., Algina, J. (1986): *Introduction to classical and modern test theory.* New York: Holt, Rinehart and Winston, Inc.
4. Hambleton, R. K., Robin, F., & Xing, D. (2000): *Item response models for the analysis of educational and psychological test data.* In H. Tinsley & S. Brown (Eds.), Handbook of applied multivariate statistics and modelling. San Diego, CA: Academic Press.
5. Hambleton, R., Swaminathan, H., Rogers, H.: *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Press.
6. de Ayala, R.:. *The Theory and Practice of Item Response Theory*, New York, NY: The Guilford Press. (6.12), p.144
7. Novick, M. (1966): *The axioms and principal results of classical test theory*, Journal of Mathematical Psychology Volume 3, Issue 1, February, Pages 1-18
8. Hambleton, R., Yue Zhao: *Item Response Theory (IRT) Models for Dichotomous Data,* Encyclopedia of Statistics in Behavioral Science, University of Massachusetts, Amherst, MA, USA.
9. Keller, L.A. (2000): *Ability Estimation Procedures in Computerized Adaptive Testing* (AICPA technical report). Ewing: The American Institute of Certified Public Accountants. AICPA.
10. Mansoor, Al-A'ali (2007): *Implementation of an Improved Adaptive Testing Theory*. Educational Technology & Society, 10 (4), 80-94
11. Linden**,** Wim J.**,** Glas**,** Cees A.W. (2010): *Elements of Adaptive Testing*, Springer, ISBN 978-0-387-85461-8
12. Lord, F. M. (1977): *A broad-range tailored test of verbal ability.* Applied Psychological Measurement, 1, *95-100.*

METHODOLOGY FOR CREATING OF OBJECTIVE AND BALANCED LINEAR TESTS BASED ON THE ITEM RESPONSE THEORY

Ivan Gospodinov, Borislav Slavov, Stefan Filipov

University of Chemical Technology and Metallurgy, Sofia,
Department of Programming and Computer Systems Application,
gospodinov.ivan@yahoo.com

***Abstract**: The estimation of the cognitive ability of an individual is an important part of the educational process. The main goal of every test theory is to create tests that are objective along the entire range of abilities of the tested audience. Or in other words – the final ability estimate of every individual needs to reflect his/her knowledge, regardless of the amount of knowledge he/she possesses. Two theories are widespread nowadays – the Classical Test Theory (CTT) and the Item Response Theory (IRT). This paper reviews some of the disadvantages of CTT for creation of objective and balanced tests and demonstrates how these disadvantages can be overcome when the more modern IRT is applied.*