Az Élő-pontrendszer és az item-válasz-elmélet ötvözése adaptív teszteléshez

Combining Elo rating system with Item Response Theory for Adaptive Item Sequencing

ANTAL Margit

Sapientia EMTE, Műszaki és Humántudományok kar, Marosvásárhely
adjunktus, manyi@ms.sapientia.ro

Abstract

In this paper, we present Elo rating used for adaptive testing. Since Elo rating was created only for calculating the skill levels of players, therefore Item information known from Item Response Theory (IRT) was used for selecting the next item in the process of adaptive item sequencing. The effectiveness of the method was tested on both real and simulated data. Item difficulty estimation was performed on a real dataset using three estimation methods. We found that Elo rating based method provides reliable estimates. The effect of different test lengths were examined using simulated data. Results indicate that Elo rating based method is slower than the IRT-based one, but offers estimations for item difficulties and provides reliable estimates providing that the test has at least 30 items.

Összefoglaló

A dolgozat célja bemutatni az Élő pontrendszer használatát adaptív tesztelésre. Mivel az Élő pontrendszer csak a képességszint újrabecslésével foglalkozik, ezért a következő item kiválasztásához az item-válasz elméletből ismert item-információt használjuk. A módszer hatékonyságát egy valós és egy szimulált adathalmazon ellenőriztük. A valós adathalmazon az item nehézségi paraméterének becslését végeztük el háromféle módszerrel, a szimulált adathalmazon pedig a teszt hosszát próbáltuk hangolni úgy, hogy az új módszerrel kapott képességszintek összemérhetőek legyenek az item-válasz-elmélet alapúakkal. Méréseink alapján az új módszer helyesen becsli a képességszintet, amennyiben kellő hosszúságú a teszt, ezen felül pedig lehetőséget teremt az itemek nehézségi paraméterének becslésére is.

Kulcsszavak: Élő-pontrendszer, Item-válasz-elmélet, item nehézség, adaptív tesztelés.

1. Bevezetés

Egyre több adaptív oktatási rendszer jelenik meg, amelyek megpróbálják minél jobban kielégíteni a felhasználók igényeit. Az adaptivitás alapján megkülönböztetünk a felhasználó tudásszintjéhez, az érdeklődéséhez, illetve szokásaihoz igazodó rendszereket [3]. Ebben a dolgozatban a tudásszinthez igazodó tesztrendszerekkel foglalkozunk, amelyeket adaptív tesztrendszereknek nevezünk. Nagyon sok nemzetközi intézmény adaptív tesztrendszereket használ a nemzetközileg elismert vizsgáihoz (TOEFL, GMAT, GRE). Ezen rendszerek a tesztitemeket előzetes kipróbálások alapján kalibrálják. Ahhoz, hogy megbízható paramétereket kapjunk, nagy mennyiségű, jól megválasztott tesztalanyra van szükség, ami egy nagyon költséges folyamat. Mi több, rosszul megválasztott tesztalanyok esetén torzulhatnak az item paraméterek [8]. Éppen ezért a dolgozatban megpróbálunk alternatív módszereket keresni az itemek paramétereinek kalibrálására.

Nagyon kevés olyan tanulmányt találtunk, amely ezzel a témával foglalkozik. Klinkenberg és társai [6] egy matematikai gyakorló rendszerbe (Math Garden) vezették be az Élő-pontrendszerre épülő képességszint, illetve item nehézség becslést. Wauters és társai [10] [11] összehasonlítottak hat módszert, amelyek az itemek nehézségeit hivatottak becsülni. Dolgozatunkban megismételjük Wautersék által elvégzett kísérletet saját adathalmazra, majd bevezetjük és összehasonlítjuk az item-válasz alapú adaptív tesztelést az általunk javasolt Élő-pontrendszerre épülővel.

2. A képességszint mérése

2.1. Élő-pontrendszer

Az Élő-pontrendszert Élő Árpád vezette be 1978-ban sakkozók rangsorolására [4]. Ma már több sportágban használják ezt a rendszert, sőt sok számítógépes játékban is ezzel rangsorolják a játékosokat. A módszer lényege, hogy minden játszma után mindkét játékos pontszámát módosítja, a mérkőzés eredménye és az ellenfél játszma előtti pontszáma függvényében. Erősebb játékos legyőzése több ponttal növeli a nyertes játékos pontszámát, mint gyengébb ellenfél legyőzése. Az Élő-pontrendszer alapján a játékosok pontszáma időben lassan változik.

Az Élő-pontrendszer egyszerű számításokat igényel. Jelölje Θ_A és Θ_B az A és B játékosok képességszintjeit (pontszámát). Minden játszma után a játékosok képességszintjeit az (1) és (2) képletekkel újraszámítjuk.

$$\hat{\theta}_A = \theta_A + K(S_A - E(S_A)) \tag{1}$$

$$\hat{\theta}_B = \theta_B + K(S_B - E(S_B)) \tag{2}$$

$$E(S_A) = \frac{1}{1+10^{(\theta_B-\theta_A)/400}} \tag{3}$$

A fenti képletekben S_A a játszma eredménye az *A* játékos szemszögéből nézve (0 – veszített, 0.5 - döntetlen és 1 –nyert) és $E(S_A)$ a játszma az *A* játékos általi megnyerésének esélye. Az (1) és (2) képletekben használt K a képességszint változásának súlyozását hivatott végezni, és ezt általában állandóként szokás megválasztani.

Adaptív tesztrendszerek esetében az egyik játékos a vizsgázó, a másik pedig az item (tesztkérdés, feladat). Legyen Θ_A a vizsgázó képességszintje és *b* a kiválasztott item nehézsége, mindkét értéket ugyanazon a skálán mérjük, ez általában [-3, 3]. A játszma ezúttal a vizsgázó és az item között van, az (1) és (2) képletek pedig a következőképpen alakulnak:

$$\hat{\theta} = \theta + K(S - E(S)) \tag{4}$$

$$\hat{b} = b - K(S - E(S)) \tag{5}$$

S értéke 1, ha a vizsgázó helyesen válaszolt, ellenkező esetben 0. Így minden egyes helyes válasz esetén a vizsgázó képességszintje növekszik és az item nehézsége csökken. Ebben a dolgozatban K értékét 0.4-nek választottuk [12]. Habár K-t állandónak szokás választani, lehetne függvényt is használni. A Wauters és társai [12] által javasolt függvény viszont nem bizonyult megfelelőnek. Az E(S) számításához a (6) képlettel megadott logisztikai függvényt használtuk.

$$E(S) = \frac{1}{1+e^{(b-\theta)}} = \frac{1}{1+e^{-(\theta-b)}} \tag{6}$$

2.2. Item-válasz-elmélet

Az *Item-válasz-elmélet* egy valószínűségi tesztelmélet, amelynek fő célja a tesztitemek igazítása a vizsgázó képességi szintjéhez. A vizsgázó képességi szintjének becslése folyamatosan, a tesztelés során történik. Az adaptív tesztelés a következő lépésekből áll: (i) egy kezdeti képességszint beállítása, (ii) az adott képességszinthez a legmegfelelőbb kérdés kiválasztása, (iii) a képességszint újrabecslése a kérdésre adott válasz alapján. A második és a harmadik lépést addig ismételjük amíg a befejezési feltételek nem teljesülnek [1]. A következőkben a háromparaméteres logisztikai modellt mutatjuk be. Ebben a modellben minden egyes itemhez egy item-karakterisztikus görbét rendelünk, amely megmutatja, hogy adott Θ képességszintű diák milyen valószínűséggel válaszol helyesen az adott itemre [7]. Az item-karakterisztikus görbe egyenlete a következő:

$$P(\theta) = c + \frac{(1-c)}{1+e^{-D a(\theta - b)}} \quad (7)$$

ahol *a* az item diszkriminációja, *b* a nehézsége és *c* pedig a válasz kitalálásának valószínűsége. Az item nehézsége azonos skálán mozog a vizsgázó képességszintjével. Elméletileg ez az érték -∞ és +∞ között mozog, a gyakorlatban azonban elegendő -3 és +3 közötti intervallum. A diszkrimináció azt mutatja meg, hogy az item mennyire jól választja szét a vizsgázókat az adott nehézségi szinten. Ez a paraméter a görbe meredekségét határozza meg annak középső szakaszában. Minél meredekebb a görbe ezen szakasza, annál nagyobb az item diszkriminációja az item nehézségi szintjén. A kitalálási faktor egy valószínűségi érték, például egy igen/nem választ váró kérés esetében értéke 0.5, a *D* pedig egy skálázási faktor, amelynek 1.7 értéket szokás használni [7].

Ebben a dolgozatban az egyparaméteres vagy más néven Rasch modellt használtuk, amelyben az itemre adott helyes válasz valószínűségét úgy kapjuk meg, hogy a (7) képletben a b paraméter értékének 1-et, a c paraméternek pedig 0-át helyettesítünk.

A logisztikai modell másik fontos eleme az item-információ függvény, amely méri, hogy az item segítségével mennyire lehet pontosan becsülni a képességszintet. Ha az item-információ nagy, nagyobb pontossággal lehet meghatározni egy item után a képességszintet, így a nagyobb információs mutatóval rendelkező kérdés kerül a vizsgázó elé, nem pedig az, amelyre nagyobb valószínűséggel tud válaszolni. Az item-információ számítására a következő képletet használtuk [7]:

$$I_i(\theta) = \frac{P_i'(\Theta)^2}{P_i(\Theta)(1 - P_i(\Theta))} \quad (8)$$

A $P_i'(\theta)$, a $P(\theta)$ elsőrendű deriváltja.

Az adaptív tesztelés harmadik lépése az új képességszint becslése az előzetesen megválaszolt itemek alapján. A szakirodalom erre több képletet is ajánl, amelyeket rendre kipróbálva, a Rudner [7] által ajánlott bizonyult a legmegfelelőbbnek:

$$\hat{\Theta}_{s+1} = \hat{\Theta}_s + \frac{\sum_{i=1}^{N} S_i(\hat{\Theta}_s)}{\sum_{i=1}^{N} I_i(\hat{\Theta}_s)} \quad (9)$$

ahol

$$S_i(\Theta) = (u_i - P_i(\Theta)) \frac{P_i'}{P_i(1 - P_i)} \quad (10)$$

valamint u_i az i. kérdésre adott válasz alapján 1, amennyiben a válasz helyes, illetve 0 ellenkező esetben.

Az *Item-válasz-elmélet* két esetben mond csődöt, amikor minden kérdésre helyes, vagy minden kérdésre helytelen választ ad a vizsgázó. Ezeket a szélsőséges eseteket figyelni kell, és egy adott kérdésszám után le kell állítani a tesztelést. Bármilyen más esetben a megállási feltételt a standard hibához kötjük. A standard hiba a képességszint becslésének pontosságát jellemzi, ezért ha ez az érték egy küszöbérték alá csökken, leállíthatjuk a tesztelést.

A standard hiba kiszámításához felhasználjuk a teszt-információ függvényt, amelyet a következő képlettel számítunk [7]:

$$T(\Theta) = \sum_{i=1}^{N} I_i(\Theta) \quad (11)$$

Ezután a standard hibát pedig így számíthatjuk [7]:

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \tag{12}$$

3. Itemek nehézségének becslése – mérési eredmények

Az itemek nehézségi paramétereit a tesztelés elkezdése előtt be kell állítani. Ezt általában úgy végzik, hogy egy nagy teszthalmazon előzetes méréseket végeznek, majd egy IRT paraméterek becslésére alkalmas programmal (BILOG-MG, MULTILOG, PARAM-3PL stb.) meghatározzák a paramétereket. Sajnos az előzetes mérések végzése költséges. Mi több, amennyiben a tesztalanyok nem megfelelően vannak kiválasztva, torzulhat az item paraméterek becslése. Éppen ezért fontos lenne alternatív megoldásokat találni item paraméterek becslésére. Egy pár tanulmány [2][5][6][10][11] megpróbált alternatív megoldást nyújtani erre a problémára. Wauters és társai [10][11] nemcsak megoldásokat nyújtottak, hanem ezeket össze is hasonlították. Ebben a dolgozatban egy hasonló mérést mutatunk be, amelyet egy saját, valós adathalmazon mértünk.

3.1. Résztvevők

A mérésekhez az adatokat 2011 és 2012-ben gyűjtöttük az Objektumorientált programozás vizsgán. Két adathalmazt alkottunk OOP1 és OOP2, ahol az OOP1 a 2011-ben, illetve az OOP2 a 2012-ben gyűjtött adatokat tartalmazza. Az OOP1 adathalmaz 74 vizsgázó, illetve az OOP2 63 vizsgázó adatait tartalmazza. A vizsgázók Automatizálás, Informatika és Számítástechnika szakos hallgatók voltak.

3.2. Adatok és módszerek

Az adatokat két, egyenként 30 itemet tartalmazó teszt szolgáltatta, amelyet a Moodle oktatási környezeten keresztül bonyolítottunk le, ellenőrzött körülmények között. A hallgatók jártasak voltak a környezetben, hiszen egész félév során használták a környezetet, sőt hasonló ismeretfelmérő tesztet is végeztek korábban.
A begyűjtött adathalmazon háromféle módszerrel becsültük az itemek nehézségét:

- Item-válasz elmélet alapú becslés az *R* statisztikai program *ltm* csomagja segítségével (IRT). A becslést az *ltm* csomag *rasch* függvénye segítségével végeztük.

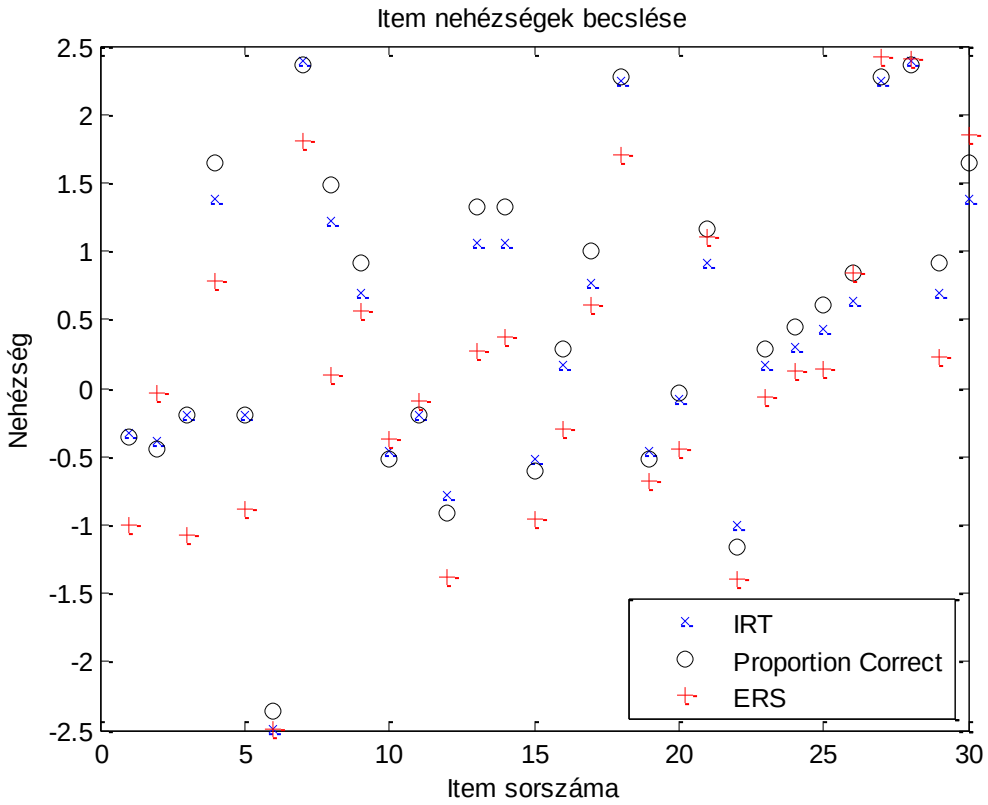
- Élő-pontrendszer alapú becslés, a (4), (5), és (6) képletek segítségével (ERS). Ennél a becslési módszernél az item nehézségeket kezdetben 0-ra állítottuk, majd rendre minden egyes vizsgázó adataira a (4), (5), és (6) képletek segítségével újrabecsültük az item nehzéségét és a képességszintet.

- A helyes válaszok aránya alapú becslés a (13) képlet segítségével történt (ProportionCorrect).

$$\hat{b}_i = 1 - \frac{n_i}{N_i} \tag{13}$$

ahol N_i az összes válasz száma, n_i a helyes válaszok száma az *i*-ik itemre. Azért, hogy az így kapott nehézségek összemérhetők legyenek, az értékeket a [-3, 3] intervallumra skáláztuk.

3.3. Eredmények



1. ábra *Item nehézségek becslése három különböző módszerrel*

Az 1. ábra a három különböző becslési módszerrel elért eredményeket szemlélteti az itemek nehézségi paraméterére vonatkozóan. A három módszerrel mért item nehézségek korrelációját is kiszámítottuk mindkét adathalmazra. Az 1. és 2. táblázatok ezeket a korrelációkat szemléltetik. A mérések alapján kijelenthető, hogy a helyes válaszok aránya alapú- és az IRT –becslés közötti korreláció a legnagyobb. A fenti ábrán viszont az is látszik, hogy az ERS becsléssel kapott item nehézségek a legkisebbek.

	IRT	ERS	Prop. Correct
IRT	1	0.942	0.994
ERS		1	0.937
Prop. Correct			1

OOP1 adathalmaz

	IRT	ERS	Prop. Correct
IRT	1	0.930	0.991
ERS		1	0.931
Prop. Correct			1

OOP2 adathalmaz

1. táblázat *Pearson korreláció a különböző módszerekkel becsült item nehézségek között.*

4. A képességszint becslése – mérési eredmények

4.1. Adatok és módszerek

Ebben a részben bevezetjük az Élő-pontrendszer alapú adaptív tesztelést és összehasonlítjuk az item-válasz elmélet alapúval. Az adaptív tesztelés a következő egyszerű algoritmuson alapszik:

1. lépés Adjunk a vizsgázónak egy pár közepes nehézségű tesztitemet és állapítsuk meg a válaszokból a kezdeti képességszintjét.

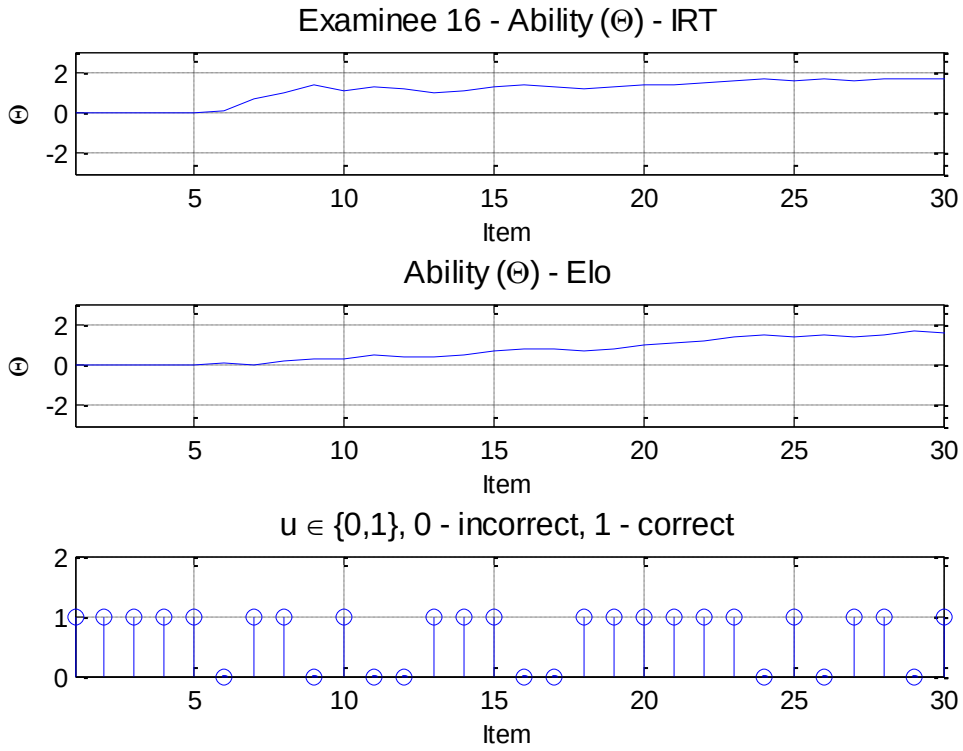
2. lépés Válasszuk ki azt a tesztitemet, amelynek információtartalma a legmagasabb a vizsgázó aktuális képességszintjén.

3. lépés A vizsgázó válasza alapján becsüljük újra a képességszintjét.
4. lépés Ha teljesülnek a megállási feltételek, leállítjuk le a tesztelést, különben folytatjuk a *2. lépéssel*.

A 2. lépésben mindig a legnagyobb információjú itemet választjuk, az iteminformációt pedig a (8) képlettel számítjuk. A 3. lépésben az Élő pontrendszer estén a (4) képletet, az item-válasz elmélet esetében pedig a (9) képletet használjuk. Az Élő pontrendszer esetében újrabecsülhetnénk az item nehézségét is az (5) képlettel, de ezt nem használtuk ebben a kísérletben.
Összesen 1000 vizsgázót szimuláltunk. Annak érdekében, hogy a két módszer összehasonlítható legyen, a válaszokat előre generáltuk egy egyenletes eloszlású véletlenszám generátort használva. Az itembankot is mesterségesen állítottuk elő, 200 itemet, amelyek nehézségei egyenletes eloszlásúak a [-3, 3] intervallumban. Az első öt itemet véletlenszerűen választottuk ki közepes nehézségűek közül. A képességszint becslése minden vizsgázó esetében az 5. item után kezdődik. Minden vizsgázó ugyanannyi itemet kapott, tehát nem használtunk megállási feltételeket. A szimulációt különböző rögzített itemszámra mindkét módszerre elvégeztük és a kapott képességszinteket összehasonlítottuk.

4. 2. Eredmények

A 2. ábra a 16. vizsgázó válaszait, illetve az ezeknek megfelelő képességszint becsléseket szemlélteti egy 30 itemet tartalmazó teszt esetében.



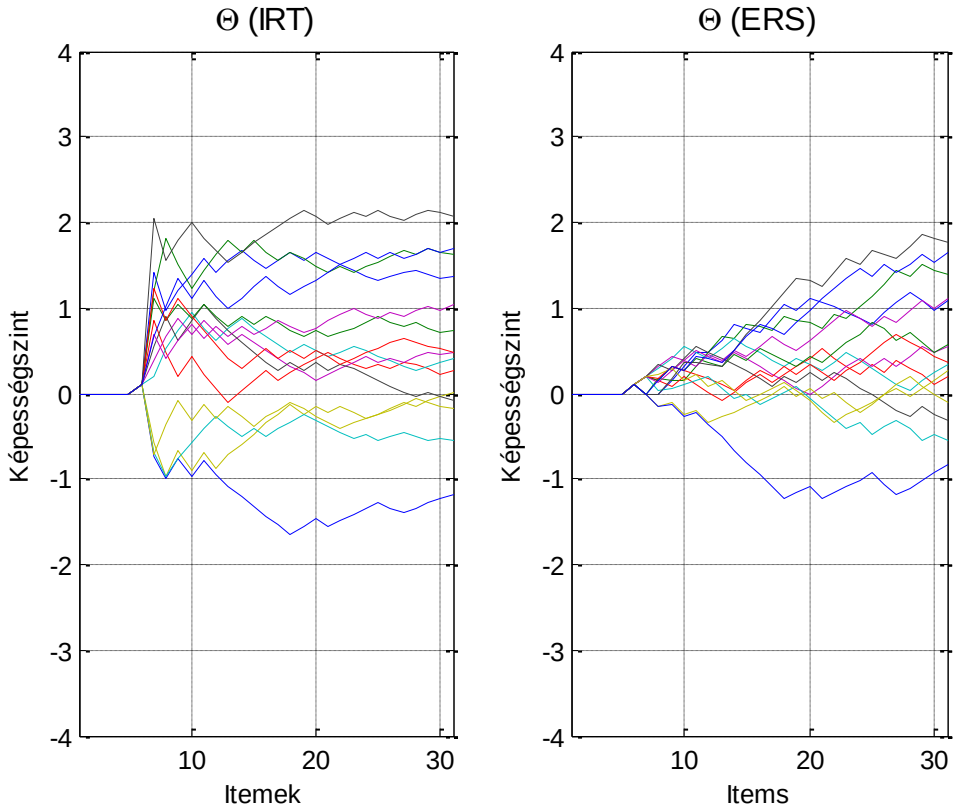
2. ábra *Képességszint becslése IRT és ERS módszerekkel*

A szimulációt ötször végeztük el 10, 15, 20, 25 és 30 itemet tartalmazó tesztekre. A szimulációk végén képeztük a két módszerrel kapott képességszintek különbségét, amelynek kiszámítottuk az átlagát és a szórását is. A 2. táblázat ezeket az eredményeket foglalja össze. Az első oszlop a teszt hosszát tartalmazza, a második a két módszerrel kapott képességszintek különbségének átlagát, a harmadik pedig ugyanennek a szórását. A negyedik és az ötödik oszlop az IRT esetében mért hibák átlagát, illetve ezek szórását tartalmazza.

Teszt hossza	Átlag (Θ_{IRT}-Θ_{ERS})	Szórás (Θ_{IRT}-Θ_{ERS})	Átlag (hiba)	Szórás (hiba)
10	0.68	0.52	0.46	0.034
15	0.49	0.41	0.35	0.033
20	0.31	0.27	0.30	0.011
25	0.23	0.20	0.26	0.009
30	0.18	0.14	0.23	0.007

2. táblázat *A teszt hossza és a képességszint becslése közötti összefüggés*

A két módszer közötti különbséget jól szemlélteti a 3. ábra, ahol 10 vizsgázó képességszintjének becslését szemléltettük két különböző módszert használva. Az IRT sokkal gyorsabban képes megbecsülni a képességszintet, mint az Élő pontrendszer alapú változat.



3. ábra *Képességszintek becslése 10 vizsgázó esetén egy 30 itemet tartalmazó tesztet használva*

5. Következtetések

A dolgozatban bemutattuk az Élő-pontrendszert, illetve ennek használatát adaptív tesztelésre. Habár ez a módszer lassúbb a képességszint becslésében, mint az item-válasz elmélet alapú, mert legalább 30 itemet tartalmazó teszt szükséges, azzal az előnnyel rendelkezik, hogy a képességszint mellett az itemek nehézségét is képes folyamatosan becsülni. Így lehetővé válik, hogy egy erre a módszerre épülő adaptív tesztrendszer folyamatosan finomítsa az itemek nehézségi paramétereit. Mivel az Élő-pontrendszer csak a nehézség paraméter becslését teszi lehetővé, ezért az összehasonlításokban az 1-paraméteres, Rasch modellel dolgoztunk. A teszt hosszára vonatkozó mérési eredmények összhangban vannak van der Maas, Wagenmarkers és társai eredményeivel [9], akik arra a következtetésre jutottak, hogy sakkozók esetében az Élő-pontrendszer 25 játszma lejátszása után ad

megbízható eredményt.

A mérések igazolják, hogy az Élő-pontrendszer jól használható adaptív tesztelésre. A módszer lassúsága miatt viszont ajánlott inkább adaptív gyakorló rendszerekbe beépíteni.

Köszönetnyilvánítás

A dolgozatban bemutatott kutatást a *Sapientia Alapítvány - Kutatási Programok Intézete* támogatta.

Hivatkozások

[1] Antal, M., Erős, L. (2010). Item-válasz-elmélet alapú adaptív tesztelés (Item Response Theory Based Adaptive Testing), SZAMOKT XX., October 7-10, 2010, Satu Mare, Romania, pp. 101-106.

[2] Brinkhuis, M. J. S., Maris, G. (2009). Dynamic Parameter Estimation in Student Monitoring Systems. CITO-report.

[3] Brusilovsky, P. , Millan, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In: P. Brusilovsky, A. Kobsa and W. Neidl (eds.): The Adaptive Web: Methods and Strategies of Web Personalization. Lecture Notes in Computer Science, Vol. 4321, Berlin Heidelberg New York: Springer-Verlag, 3-53.

[4] Elo, A. E. (1978). The rating of chess players, past and present. London: B.T. Batsford, Ltd.

[5] Kingsbury GG. (2009). Adaptive Item Calibration: A Process for Estimating Item Parameters Within a Computerized Adaptive Test. In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.

[6] Klinkenberg S, Straatemeier M, van der Maas HLJ. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. Computers & Education. 57 (2), 1813-1824.

[7] Rudner, L. M. (1998). *An online, interactive, computer adaptive testing tutorial.* http://echo.edres.org:8080/scripts/cat/catdemo.htm

[8] Stocking, M.L. (1990). Specifying optimum examinees for item parameter estimation in Item Response Theory. *Psychometrika 55*(3), 461-475.

[9] Van der Maas, H. L., Wagenmakers, E.-J. (2005). A psychometric analysis of chess expertise, American Journal of Psychology,, Vol. 118, No. 1, pp. 29-60.

[10] Wauters, K., Desmet, P., Van Den Noortgate, W. (2012). Item difficulty estimation: An auspicious collaboration between data and judgement. Computers & Education, 1183-1193.

[11] Wauters, K., Desmet, P., Van Den Noortgate, W. (2011). Acquiring Item Difficulty Estimations: a Collaborative Effort of Data and Judgment. Educational Data Mining EDM 2011, 121-128.

[12] Wauters, K., Desmet, P., Van Den Noortgate, W. (2011). Monitoring Learners' Proficiency: Weight Adaptation in the Elo Rating System. Educational Data Mining EDM 2011, 247-252.