Többszakaszos adaptív tesztek felépítése, működése

Magyar Andrea (2013) Többszakaszos adaptív tesztek felépítése, működése. Oktatás-Informatika, 1-2. sz. http://www.oktatas-informatika.hu/2013/11/magyar-andrea-tobbszakaszos-adaptiv-tesztek-felepitese-mukodese

A technológia fejlődésével a mérés-értékelés hagyományos, papír alapú formáit fokozatosan felváltják a számítógép-alapú tesztek, melynek egyik legfejlettebb formája az adaptív tesztelés. Adaptív tesztelés során a tanulók válaszuk függvényében kapnak a következő lépésben könnyebb, vagy nehezebb feladatot, ezáltal lehetővé válik a képességszinthez igazodó mérés-értékelés. Az itemek kiközvetítésének módjától függően az adaptív teszteknek számos fajtája létezik. Az egyik leggyakrabban alkalmazott teszttípus a többszakaszos adaptív teszt, melynél különálló itemek helyett különböző nehézségű rövid, fix tesztek kerülnek kiosztásra. A tanulmány célja rövid áttekintést adni a többszakaszos adaptív tesztek felépítéséről, működéséről.

With the development of technology the traditional paper-based measurement-evaluation forms are gradually being replaced by computer-based tests. Among the many forms of the computer based assessment, the adaptive test is one of the most advanced forms. During the adaptive testing procedure the students are administered test questions selected on the basis of the correctness of their previous responses. If the answer is right, they get a more difficult item, if no then an easier item is selected and administered to the students. This way the tests became tailored to the students' ability levels. Depending on the administration process, there are several forms of adaptive tests. One of the most commonly used forms is the multi-stage test, in which short, fixed tests (called modules) are administered instead of separate items.

The aim of this study is to give a brief overview of the structure and the operation of multi-stage tests. For adaptive tests a calibrated itempool is essential condition. In the first part we discuss the main points of the development of the itempool with the role of the item response theory in it. In the adaptive testing process the use of item-response theory is essential as during the testing procedure the items are administered according to their certain parameters, such as their difficulty, discrimination or guessing. Many itempools contains other parameters, like content areas, as well. It the next part the different structures of multi-

stage tests are introduced and the steps of the development of the following stages. Here we introduce some research issues dealing with the sufficient number of items, stages and modules in a certain test. In the last part there are some issues of the branching rules and scoring methods. Comparing the multi-stage tests and the item-based adaptive tests, the multi-stage tests have a lot of advantages, such as easier development, greater administrative control and simpler testing procedure. The students have opportunities to review their previous steps, this way they are helped to reach better results. As the modules are tailored to the students' ability level, there is a greater challenge for each student.

With this paper we would like to contribute to the spread of adaptive testing methods.

Az utóbbi két évtizedben a számítógépek rohamos térhódításával a számítógépes tesztelés (l. *Molnár*, 2010; *Csapó, Tóth, Molnár, Pap-Szigeti és R. Tóth*, 2009; *R. Tóth, Molnár, Latour és Csapó*, 2011) legfejlettebb formájává az adaptív tesztelés (CAT – Computerized Adaptive Testing) vált (*Amstrong*, 2002; *Csapó, Molnár és R. Tóth*, 2008). Adaptív tesztelés során a tesztkérdések a tanulók előző válaszai alapján kerülnek kiközvetítésre. Amennyiben jó választ adnak, nehezebb item következik, amennyiben nem, akkor könnyebb. Ezáltal a tesztelés a tanulók képességszintjéhez igazodik (*Csapó, Molnár* és *R. Tóth*, 2008; *Keng*, 2008).

A fix tesztekhez képest az adaptív tesztek számos előnnyel rendelkeznek: A tanulók személyre szabott feladatokat kapnak, így a pontosabb képességmérés lehetőségét teremtik meg (*Linacre*, 2000). A vizsga során minden tanuló különböző kérdést kap, ezért nincs lehetőség a válaszok előzetes betanulására, biztonságosabbá válik a tesztelés (*Wainer*, 2000). A tesztelési idő átlagosan felére csökken, ezáltal kevésbé fárasztó a tanulók számára (*Frey és Seitz*, 2009). A CAT számos előnye ellenére hátrányokkal és korlátokkal is rendelkezik. Az adaptív tesztek előállítási költsége jóval magasabb, mint a hagyományos teszteké *(Linacre*, 2000; *Wainer*, 2000; *Meijer* és *Nering*, 2000). Az itemek a kiközvetítés során esetlegesen befolyásolhatják egymást, valamint ugyanaz az item lehet könnyebb, vagy nehezebb attól függően, hogy milyen itemek után következik. Gyakran tartalmi átfedések lehetnek az itemek között (*Wainer*, 2000). Az adaptív tesztelésre vonatkozóan további előnyöket és hátrányokat részletez *Molnár* (2013) és *Magyar* (2012).

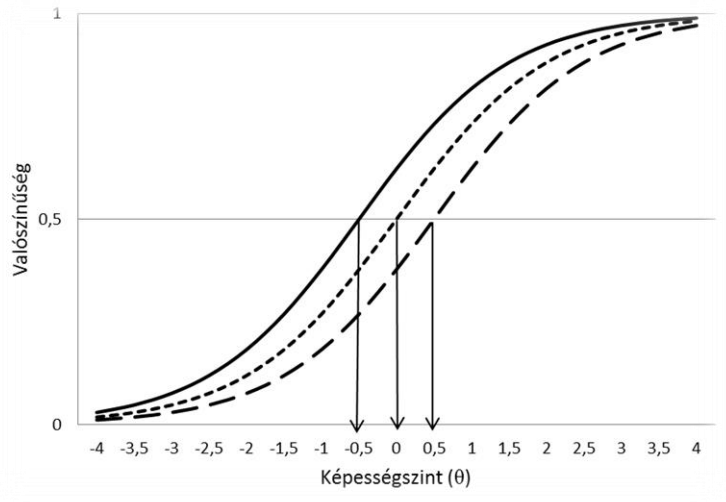
Ezek közül több probléma kiküszöbölését oldja meg a többszakaszos adaptív teszt (MST – Multi Stage Test), mely egyesíti magában a hagyományos fix és az adaptív tesztek tulajdonságait, egyrészt a kérdéseket a tanuló képességszintjéhez igazítja, másrészt lehetőséget ad az itemek sorrendjének előzetes meghatározására (*Amstrong*, 2004; *Molnár*,

2013). Szerkezetüket tekintve a többszakaszos tesztek a lineáris fix és az item alapú adaptív tesztek között félúton helyezkednek el (*Jodoin*, 2006; *Patsula*, 1999). A tesztelés során több szakaszban itemek helyett modulok kerülnek kiosztásra, melyek tulajdonképpen különböző nehézségi szintű rövid fix tesztek. Egy teszt minimum két szakaszból áll. Egy szakaszon belül két vagy több modul lehet, melyek nehézségi szintjükben különböznek. Miután a tanuló végez egy-egy modullal, képességszintje becslésre kerül, és ez alapján kap a következő szakaszban újabb nehézségi szintűt (*Zenisky, Hambleton és Luecht*, 2010).

A tanulmány célja áttekintést adni a többszakaszos adaptív tesztek alkalmazási lehetőségeiről. Bemutatjuk a többszakaszos tesztek felépítését, a kalibrált feladatbank előállításának lépéseit, tesztelési algoritmusát, különböző típusait, a modulok pontozási lehetőségeit, valamint az MST előnyeit és hátrányait.

Kalibrált feladatbank előállítása

Az item-alapú adaptív tesztekhez hasonlóan a többszakaszos tesztek előállításának hátterében is feladatbank áll, mely az itemeket és azok jellemzőit is tartalmazza. Ezek egyrészt tartalmi és adminisztratív információk, másrészt pedig az itemek empirikus paraméterei (*Eggen*, 2007). Az itemek kalibrálása a valószínűségi tesztelméletet (IRT – Item Response Theory) felhasználva egy-, két-, vagy három-paraméteres logisztikus modellek felhasználásával történhet. Az egy-paraméteres logisztikus modell (más néven Rasch modell) a személyparaméter mellett egy paramétert tartalmaz, az item nehézségi mutatót. Az item nehézségi mutató azt adja meg, hogy milyen képességszint szükséges ahhoz, hogy az adott itemet 50% valószínűséggel oldja meg a tanuló (*Molnár*, 2013). Az *1. ábra* különböző nehézségű itemek karakterisztikus görbéit ábrázolja:



1. ábra: Három különböző nehézségű item karakterisztikus görbéje

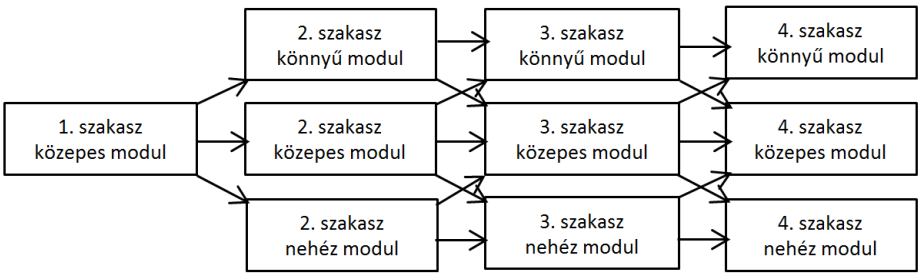
A baloldali görbe a legkönnyebb, a középső görbe az átlagos, a jobboldali a legnehezebb item helyes megoldásának valószínűségi görbéjét ábrázolja. A példán szereplő itemek nehézségi mutatói: -0,5, 0 és 0,5. Az itemek karakterisztikus görbéjének pontonkénti összegzésével előállítható a teszt karakterisztikus görbe, mely a teljes tesztre vonatkozóan közvetít információt (*Baker*, 2001; 3. ábra).

A kalibrált itemek a feladatbankba kerülnek, majd megfelelő nehézségi szintű modulba kerülnek besorolásra. Ez kétféleképpen történhet: alulról felfelé, illetve felülről lefelé. Alulról felfelé való teszt konstrukciónál a tartalmi és statisztikai jellemzők a modulok szintjén történnek megadásra. Ezáltal egymáshoz nagyban hasonló, felcserélhető modulok szerkeszthetők. Ezzel szemben, a felülről, a teszt szintjén való szerkesztésnél a teszt egészére történik a jellemzők megadása, így a modulok nem feltétlenül lesznek felcserélhetőek (*Luecht*, 1998).

A tesztek szerkesztése

A többszakaszos teszteknek többféle változata létezik, ennek megfelelően szerkesztésük is különböző. A legegyszerűbb szerkezetűek a kétszakaszos (’two-stage’) tesztek, melyeknél a tanulók először egy azonos kezdő modult (routing test) kapnak, mely különböző nehézségű itemeket tartalmaz. Ezen teszten elért eredményük függvényében kerül kiosztásra a tanulók saját képességszintjéhez illeszkedő nehézségű felmérő teszt a második részben *(Adema*, 1990; *Puhan*, 2003). A felmérő tesztek különböző nehézségi szintű fix tesztek, melyek nagyjából azonos nehézségi szintű kalibrált itemeket tartalmaznak. A felmérő teszt megoldásával és értékelésével be is fejeződik a tesztelés. Többszakaszos tesztek esetében több felmérő teszt követi egymást, nehézségi szintjük szerint modulokba rendezve (*Hendrickson*, 2007).

Az MST nagy változatosságot enged a szakaszok, a szakaszokon belül a modulok, és a modulokon belül az itemek számát illetően (*Davis* és *Dodd*, 2003). A *2. ábra* egy négy szakaszból álló tesztet ábrázol, szintenként három különböző nehézségű modullal. A szintekről való továbblépés a teljesítmény függvényében egy-egy szinttel változhat.



2. ábra: 1-3-3-3 szerkezetű négyszakaszos teszt (*Molnár*, 2013 alapján)

A szakaszok száma nagyban befolyásolja a teszt összetettségét. Minél több szakaszból áll a teszt, annál több a lehetséges útvonal száma és annál többféle nehézségi szintű teszt állítható elő. (A 2. ábrán összesen 17 különböző útvonal lehet). A szakaszok számának növelésével azonban egyre összetettebbé válik a teszt adminisztráció a mérési precizitás arányos növekedése nélkül (*Amstrong,* 2004; *Hendrickson*, 2007). Ezen meggondolások alapján a leggyakrabban 2-4 szakaszos tesztek alkalmazása terjedt el (*Zenisky,* 2010).

A szakaszokon belül a modulok is változó számúak lehetnek. Általában a teszt egy modullal kezdődik, és azt követően háromra, esetenként ötre nő az egy szinten lévő modulszám (*Patsula*, 1999). A megfelelő pontosság eléréséhez azonban általában három, maximum négy modul elegendő szintenként (*Amstrong*, 2004).

Az itemek számára vonatkozóan oszlanak meg leginkább a vélemények. *Hendrickson* (2007) tanulmánya szerint a különböző kutatásokban az itemek száma egy-egy modulon belül 1-90 közötti, azonban átlagosan öt itemből álló modulok fordulnak elő a leggyakrabban. További kutatási kérdés, hogy a kezdő, vagy a következő modulok legyenek-e hosszabbak. *Davis* és *Dodd* (2003), valamint *Kim* és *Plake* (1993) a bevezető mérés pontosságának kiemelt jelentőségével indokolja a hosszabb kezdő teszt alkalmazását. (Részletesebben l. *Keng*, 2008).

Elágazási szabály és pontozási lehetőségek

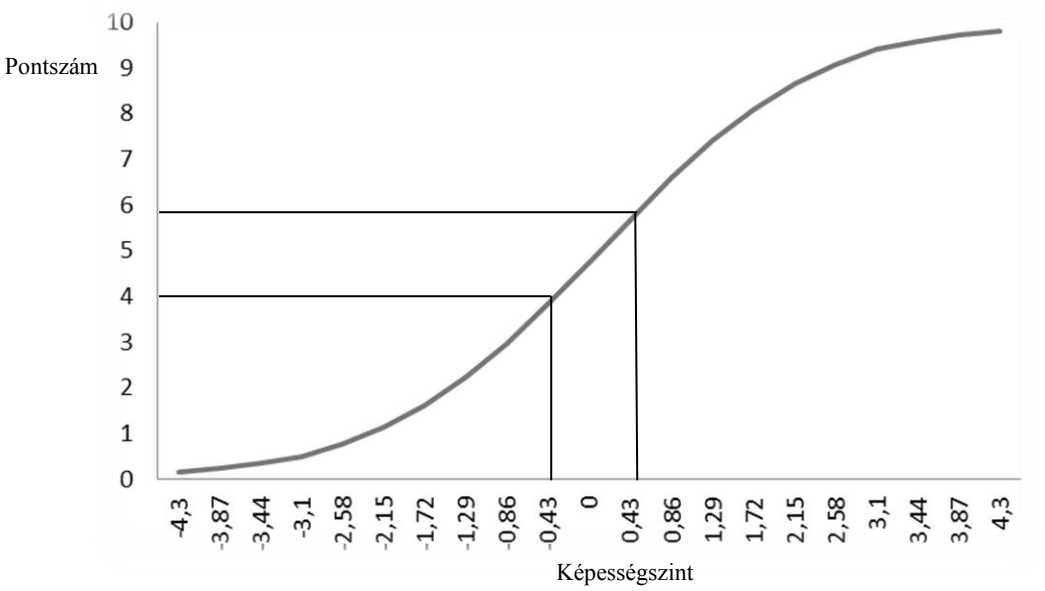
A tesztelés során a tanulók az előző modulon elért pontszámtól függően kapnak könnyebb vagy nehezebb modult a következő lépésben. Az elágazásoknál alkalmazott algoritmus alapvetően meghatározó a tesztelés során, mivel ennél a pontnál történik a tanulók hozzárendelése a különböző nehézségű modulokhoz (*Zenisky* és *Hambleton*, 2004). *Zenisky* és *Hambleton* (2004) négy leggyakrabban használatos módszert említenek: a DPI, az AMI, a NC és a random módszer.

A DPI módszer (DPI – Defined Population Interval) a tanulókat egyenlően osztja szét a következő szakaszban következő modulokba mindegyik tanulóhoz hozzárendelve az adott modul alapján becsült képességszintjét. Ez alapján egy meghatározott képességszint alatti tanulókat a könnyű modulba, az e felettieket a közepes modulba sorolja. Ha több modulra ágazik a következő szakasz, akkor több osztópont alapján dönt az algoritmus. Az osztópont standard normál eloszlású minta esetén (N(0,1)) a tanulók három modulba történő sorolásánál -0,43 és +0,43-nál lesz; két modul esetén, ha a mintát felezni szeretnénk, akkor 0,0-nál.

Az AMI módszernél (AMI – Approximate Maximum Information) az osztópont a soron következő teszt teszt információs függvénye (*Baker*, 2001; *Molnár,* 2013) alapján kerül kiszámításra, és amelyik modul a legtöbb információt szolgáltatja, azaz értéke a legközelebb

áll a tanuló eredményéhez, az választódik ki az elágazásnál. Mivel ez a technika áll a legközelebb az item alapú adaptív teszteknél használatos módszerekhez (*Keng*, 2008), ettől várható a legnagyobb mérési pontosság.

Az NC módszer (NC – Number Correct) az első és második modul közötti elágazásnál a kezdő modul teszt-karakterisztikus görbéje alapján osztja a tanulókat három hozzávetőlegesen egyenlő részre. Így, ha két modulból áll a második szakasz, akkor 0,0 képességponthoz tartozó pontszámnál lesz az osztópont, három modul esetén pedig -0,43 és +0,43 által meghatározott értékeknél kerülnek a tanulók besorolásra a modulon elért pontszámuk alapján (*3. ábra*) (A 0,0 képességpont az átlagos képességet jelenti, ettől balra helyezkednek el az ettől gyengébb képességsávok, és a negatív számok jellemzik, jobbra pedig az átlag felettiek, melyeket a pozitív számok szemléltetnek). A következő elágazásoknál hasonló módon kerülnek meghatározásra az osztópontok, az addig megtett útvonal alapján. Az NC módszer előnye, hogy alkalmazásával elkerülhető a tanulók minden egyes modul utáni képességbecslése, és hozzávetőlegesen egyenlő számú tanuló sorolható minden modulhoz *(Zenisky* és *Hambleton*, 2004).



3. ábra: Osztópontok meghatározása az NC módszer szerint: 10 itemből álló modulnál 4 és 6 pontnál helyezkednek el az osztópontok

A random módszernél véletlenszerűen történik a tanulók a modulokhoz való hozzárendelésre, az egyetlen szempont, hogy a minta egyenlő arányban kerüljön elosztásra (*Zenisky* és *Hambleton*, 2004). A felsorolt módszereknek számos kombinációja létezik, részletesen lásd *Zenisky* (2010).

Az elágazási szabályokhoz szorosan kapcsolódik a modulok és a teljes teszt pontozása. Amint az elágazási szabályoknál látható volt, a tanulók képességszintje egy-egy szakasz végeztével becslésre kerül. Gyakran a becsült képességpontokat NC (number-correct) pontszámokká konvertálja az algoritmus (*Keng*, 2008). Azonban, míg a modulok pontozására elegendő az NC pontozás, a teljes teszt pontozására nem megfelelő, mivel a tanulók statisztikailag különböző itemeket kapnak (*Zenisky*, 2010). Ezért a teljes teszt pontozására az item-alapú adaptív teszteknél használatos módszerek alkalmazhatók a többszakaszos tesztek esetén is, a megfelelő IRT modellt alkalmazva (*Keng*, 2008).

A többszakaszos tesztek előnyei és hátrányai

Az item alapú adaptív tesztekkel összehasonlítva a többszakaszos tesztek számos előnnyel rendelkeznek. A modulok előre tervezhetőek és szerkeszthetőek, így nagyobb kontrolt biztosítanak a teszt adminisztráció számára. Ezáltal kiküszöbölhetővé válik, hogy az itemek egymásnak információt szolgáltassanak (*Hendrickson,* 2007). Különösen előnyös alkalmazásuk a tartalmi korlátozások esetében *(Hendrickson*, 2007). További fontos előnyük, hogy a modulokon belül a tanulóknak lehetőségük van a visszalépésre és javításra. Mivel adaptivitás csak a modulok között valósul meg, így ez nem veszélyezteti a teszt algoritmusát és segíti a tanulókat a minél magasabb pontszám elérésében (*Vispoel,* 2000). Az item alapú adaptív tesztekhez képest jóval kevesebb adminisztrációt és számítógépes számításokat igényelnek (*Hendrickson*, 2007).

Előnyei mellett *Hendrickson* (2007) hangsúlyozza, hogy bizonyos hátrányokkal is rendelkeznek a többszakaszos tesztek. Általában több itemre van szükség azonos precizitás eléréséhez. A tesztszerkesztőknek több munkába kerül előállításuk, mivel az itemeken túl azok egymásra hatását is ellenőrizniük kell. A kétszakaszos teszteknél könnyen előfordulhat, hogy a kezdő teszt nagyobb hibával méri be a tanulók képességszintjét.

Összefoglalás

Adaptív tesztelés során a kérdések a tanulók képességszintjéhez illeszkednek, a kérdésekre adott helyes vagy helytelen válaszuk alapján kapnak a következő lépésben nehezebb vagy könnyebb kérdést. Ez a módszer számos előnnyel jár a fix tesztekhez viszonyítva: Pontosabbá válik a tesztelés, nő a tesztbiztonság, a tesztelési idő is átlagosan felére csökken.

Attól függően, hogy különálló itemeket, vagy több itemből álló rövid fix teszteket közvetít a rendszer, az adaptív teszteknek különböző típusa létezik. Az egyik legelterjedtebb

típus a többszakaszos adaptív teszt (MST – Multi Stage Test), mely a hagyományos fix és az item alapú tesztek között félúton helyezkedik el, és egyesíti magában mindkét teszt tulajdonságait. A többszakaszos tesztekben itemek helyett modulokat oldanak meg a tanulók, melyek tulajdonképpen különböző nehézségi szintű rövid fix tesztek, és az adott modulon elért eredmény függvényében kerül kiosztásra a következő nehézségi szintű modul. Ezáltal nagyobb kontroll biztosítható, és könnyebben megvalósíthatóak kisebb méréseknél is.

A többszakaszos tesztek nagymintás mérésekben egyre népszerűbbek és gyakorlati alkalmazásuk nemzetközi viszonylatban egyre terjed, melyek közül a legismertebbek a Graduate Record Examination (GRE) és a Graduate Management Admission Test (GMAT), valamint a PISA mérésekben is tervezik részleges bevezetésüket a 2015-ös méréstől kezdődően (*Frey* és *Seitz*, 2009).

Köszönetnyilvánítás

A tanulmány megírását a TÁMOP 3.1.9-11 kutatási program támogatta.

Irodalom

Adema, J. J. (1990): The construction of customized two-stage tests. *Journal of Educational Measurement*, 27. 3. sz. 241-253.

Amstrong, R. D. (2002): *Routing rules for Multiple-Form Structures*. (Computerized Testing Report 02-08). Law School Admission Council.

Armstrong, R. D., Jones, D. H., Koppel, N. B., és Pashley, P. J. (2004): Computerized adaptive testing with multiple-form structures. *Applied Psychological Measurement*, 28. sz. 147-164.

Baker, F. B. (2001): *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation. College Park, MD: University of Maryland.

Csapó Benő, Molnár Gyöngyvér és R. Tóth Krisztina (2008): A papír alapú tesztektől a számítógépes adaptív tesztelésig: a pedagógiai mérés-értékelés technikájának fejlődési tendenciái. *Iskolakultúra*, 3-4. sz. 3-16.

Csapó Benő, Molnár Gyöngyvér, Pap-szigeti Róbert és R. Tóth Krisztina (2009): A mérés-értékelés új tendenciái: a papír és számítógép alapú tesztelés összehasonlító vizsgálatai általános iskolás, illetve főiskolás diákok körében. In: Perjés István és Kozma Tamás (szerk.): *Új kutatások a neveléstudományokban.* Hatékony tudomány, pedagógiai kultúra, sikeres iskola. Magyar Tudományos Akadémia, Budapest. 99-108.

Davis, L. L., és Dodd, B. G. (2003): Item Exposure Constraints for Testlets in the Verbal Reasoning Section of the MCAT. *Applied Psychological Measurement*, 27 5. sz. 335-356.

Eggen, T. J. H. M. (2007): *Choices in CAT models in the context of educational testing.* In: D. J. Weiss (Ed.): Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.

Frey, A. és Seitz, N. N. (2009): Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation*, **35.** 2-3. sz. 89-94.

Hendrickson, A. (2007): An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, **26**. 2. sz. 44-52.

Jodoin, M. Zenisky A. és Hambleton, R. K. (2006): Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education*. 19. 3. sz. 203-220.

Keng, L. (2008*): A Comparison of the performance of testlet-based computer adaptive tests and multistage tests*. The University of Texas, Austin.

Kim, H., és Plake, B. S. (1993): *Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta.

Linacre, J. M. Ph.D. (2000): *Computer-adaptive testing: A methodology whose time has come.* MESA Psychometric Laboratory, University of Chichago.

Lord, F. M. és Norvick, M. R. (1968): *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Luecht, R. M., és Nungester, R. J. (1998): Some practical examples of computer-adaptive sequential testing. Journal of Educational Measurement, **35**. 3. sz. 229–249.

Magyar Andrea (2012): Számítógépes adaptív tesztelés. *Iskolakultúra*, 6. sz. 52-60.

Molnár Gyöngyvér (2010): Technológia-alapú mérés-értékelés hazai és nemzetközi implementációi. *Iskolakultúra*, 7-8. sz. 22-34.

Molnár Gyöngyvér (2013): *A Rasch modell alkalmazási lehetőségei az empirikus kutatások gyakorlatában*. Gondolat Kiadó, Budapest.

Patsula, L. N. (1999): *A comparison of computerized adaptive testing and multistage testing.* Electronic Doctoral Dissertations for UMass Amherst. Paper AAI9950199.

Puhan, G. és Gierl, M. J. (2003): *Evaluating the comparability of English- and French-speaking examinees on a science achievement test administered using two-stage testing.* Paper Presented at the Annual Meeting of the National Council on Measurement in Education (NCME) at the Symposium entitled, "Test Adaptations and Translations: Developments and Evaluation Advances" Chicago, Illinois, U.S.A. April 22-24.

R. Tóth Krisztina, Molnár Gyöngyvér, Thibaud Latour és Csapó Benő (2011): Az online tesztelés lehetőségei és a TAO platform alkalmazása. *Új Pedagógiai Szemle*, 61. 1-2-3-4-5. sz. 8-22.

Vispoel, W. P., Hendrickson, A. B. és Bleiler, T. (2000): Limiting answer review and change on computerized adaptive vocabulary tests: psychometric and attitudinal results. *Journal of Educational Measurement*, **37** 1. sz. 21-38.

Wainer, H. (2000): *Computerized adaptive testing: A primer* (2nd Edition). Hillsdale, NJ: Erlbaum.

Zenisky, A. L. és Hambleton, R. K. (2004): *Effects of Selected Multi-Stage Test Design Alternatives on Credentialing Examination Outcomes*. Paper presented at the annual meeting of NCME, San Diego, CA.

Zenisky, A., Hambleton, R. K. és Luecht, R. M. (2010): Multistage testing: Issues, designs and research. In: der Linden, W. J. és Glas, C. A. W. (Eds.*): Elements of adaptive testing*, New York: Springer.

Elérhetőségek:

Magyar Andrea

SZTE Neveléstudományi Doktori Iskola

II. éves PhD hallgató

Lakcím: 6800 Hódmezővásárhely, Ady Endre út 15/a

Telefon: 0630 2831157

e-mail: mandrea@edu.u-szeged.hu

magyarandrea8@gmail.com