MAGYAR PEDAGÓGIA
113. évf. 3. szám 181–193. (2013)

SZÁMÍTÓGÉP ALAPÚ ADAPTÍV ÉS RÖGZÍTETT FORMÁTUMÚ TESZTELÉS ÖSSZEHASONLÍTÓ HATÉKONYSÁGVIZSGÁLATA

Magyar Andrea* és Molnár Gyöngyvér**
** SZTE Neveléstudományi Doktori Iskola*
*** SZTE Neveléstudományi Intézet, MTA-SZTE Képességfejlődés Kutatócsoport*

Az utóbbi két évtizedben a számítógépek rohamos elterjedésével a papír-ceruza alapú tesztek helyét – mind hazai, mind nemzetközi szinten – fokozatosan felváltották, illetve felváltják a számítógép alapú mérések (*Molnár*, 2011). A tradicionális papír-ceruza teszteléssel szemben számos új lehetőség adódik alkalmazásukkal, például az innovatív, multimédiás elemeket is tartalmazó dinamikusan változó itemek megjelenítése (*Wang*, 2011; *Greiff*, *Wüstenberg* és *Funke*, 2012; *Greiff*, 2012) vagy a személyre szabott, adaptív tesztelés megvalósítása (*Eggen* és *Straetmans*, 2009).

A hagyományos, többnyire rögzített formátumú tesztek használata esetén minden tesztelt személy ugyanazon sorrendben ugyanazon feladatokat kapja a tesztelés során, függetlenül képességszintjétől és teljesítményétől. Azonban ezzel a technikával csak egy viszonylag szűk képességtartomány mérhető a szükséges pontossággal. Ha a tesztet szélesebb képességtartomány mérésére tesszük alkalmassá, azaz a feladatok nehézségi indexei széles skálán mozognak, akkor minden adatfelvételben részt vevő személy számára csak a teszt néhány feladata jelent kihívást, melyek nehézségi szintjei közel állnak a tesztet megoldó személy képességszintjéhez. A teszt többi feladata, esetlegesen a teszt nagyobb része jóval kisebb mértékben járul hozzá a személy képesség- vagy tudásszintjének pontos meghatározásához. Azok túl könnyűek vagy túl nehezek, azaz vagy nem jelentenek kihívást, vagy frusztrálóak a tesztelt személy számára, ami jelentős mértékben csökkentheti a teszteléssel kapcsolatos érdeklődését, attitűdjét, motivációját (*Csapó*, *Molnár* és *R. Tóth*, 2008).

Az adaptív tesztelési technika alkalmazása során a teszt feladatai nem előre meghatározott fix sorrendben követik egymást, hanem azokat egy feladatbankból választják ki a tesztmegoldó korábbi feladatokon nyújtott teljesítménye alapján. Például feladatszintű adaptivitás esetén (l. később) amennyiben a tanuló meg tudja oldani a teszt egy feladatát, következőleg egy nehezebbet kap, ha nem, akkor könnyebbet. Ezen algoritmus alkalmazásával a tesztelés végén minden tanulóhoz hozzárendelhető egy képességszint, melynél könnyebb feladatokat nagyobb valószínűség mellett old meg helyesen, mint helytelenül. Erre az alapfeltevésre épülnek a tudástér-elméletek is (*knowledge theory*; l. *Tóth*, 2005). Miután adaptív tesztelés során a tanulók összességében különböző feladatokból, itemekből összeállított teszteket oldanak meg, eredményük a hagyományos, klasszikus tesztel-

mélet eljárásaival nem összehasonlítható. A feladatok közös sajátossága, hogy azok azonos feladatbankból származnak, ami megteremti egyrészt a valószínűségi tesztmodellek alkalmazásának lehetőségét, másrészt a diákok teljesítményének közös képességskálán történő kifejezhetőségét (erről bővebben l. *Molnár*, 2013).

Ez a típusú feladatadás és tesztösszeállítás a hagyományos, mindenki számára azonos itemeket azonos sorrendben tartalmazó, rögzített formátumú tesztekkel szemben a teljesítmények sokkal finomabb mérését teszi lehetővé (*Linacre*, 2000). Jelentős mértékben megnő a tesztelés során kinyerhető itemekre és személyekre vonatkozó információ nagysága (l. *Molnár*, 2013). Elhanyagolhatóvá válik annak valószínűsége, hogy a tesztelt személyek ugyanazon feladatokat ugyanabban a sorrendben kapják, azaz növekszik a tesztelés biztonsága (*Wainer*, 2000). Mindez új lehetőségeket terem a mérés-értékelés területén. Ha nem törekszünk több információ kinyerésére, azaz megelégszünk a hagyományos tesztelés során elérhető pontossággal, akkor a kiközvetített feladatok száma, vagyis a teszt hossza (*Thompson*, 2007), ezzel párhuzamosan a teszt megoldásához szükséges idő is jelentős mértékben rövidül, utóbbi átlagosan felére csökken (*Frey* és *Seitz*, 2009, 2011).

Az adaptivitás mértékétől függően különböző típusú adaptív teszteket különböztetünk meg (*Al-A'ali*, 2007; *Magyar*, 2012). Feladatszintű adaptivitás esetén teljes mértékben biztosított, hogy ha a tesztelt személy helytelenül/helyesen oldja meg a teszt egyik feladatát, akkor a teszt következő feladata egy könnyebb/nehezebb feladat lesz. Azonban a feladatszintű adaptív tesztelés egyik fő problémája, hogy a feladatok paraméterei annak függvényében változnak, milyen feladatok veszik körül az adott feladatot, illetve, az a teszt melyik (elején, közepén, végén) részén helyezkedik el (*Molnár*, 2013). Ezen probléma megoldását kínálja a részteszt szintű, többszakaszos adaptív tesztelés.

A többszakaszos adaptív tesztelés egyesíti a rögzített és az item alapú adaptív tesztek jó tulajdonságait (*Jodoin*, 2006). Mindamellett, hogy a teszt nehézségi szintjét a tanuló képességszintjéhez igazítja, a résztesztek szintjén lehetőséget ad a kérdések sorrendjének előzetes meghatározására (*Amstrong*, 2004; *Molnár*, 2013). A feladatonkénti adaptív tesztelés során megismert eljárás előre meghatározott, különböző nehézségű résztesztekkel valósul meg. A többszakaszos adaptív tesztek legalább két szakaszból állnak, egy szakaszon belül legalább két vagy három különböző nehézségi szintű rögzített formátumú résztesztet tartalmaznak. Az alkalmazott algoritmus szerint a tesztelt személy képességszintje ebben az esetben nem feladatonként, hanem résztesztenként becsülhető meg. Ha a tesztelt személy egy előre meghatározott teljesítmény alatt/felett teljesít az adott részteszten, akkor a teszt következő résztesztjének feladata egy könnyebb/nehezebb részteszt (*Zenisky*, *Hambleton* és *Luecht*, 2010), egy számára a korábbinál nagyobb diagnosztikus erővel bíró részteszt lesz.

A számos előnyös tulajdonság a hagyományos rögzített formátumú tesztekkel szemben vonzóvá teszi az adaptív tesztelés alkalmazását. Ennek hatására a legjelentősebb nemzetközi oktatási vonatkozású projektekben, kutatásokban is fokozatosan előtérbe kerül. Mind az európai OECD PISA-, mint az amerikai *No Child Left Behind*- (NCLB) kutatások kapcsán is felmerült az igény alkalmazására.

Frey, *Seitz* és *Kröhne* (2011) szimulációs kísérletben vizsgálták az adaptív tesztelés bevezetésének hatását, lehetőségeit a PISA-felmérések vonatkozásában. A kutatás adat-

bázisát 14 624 15 éves tanuló korábbi (2000-es, 2003-as és 2006-os) PISA-méréseken elért válasza adta. Az elemzést összesen 348 PISA–itemen végezték el. A szimulációban az itemszintű elemzések helyett részteszt szintű elemzéseket végeztek. Eredményeik szerint a mérés hatékonysága (mérési precizitás/prezentált itemek száma) 74%-kal nőtt. Ugyanazon pontosság biztosítása esetén a szükséges itemszám a korábbi rögzített formátumú tesztek alkalmazása során szükséges 55-ről 26-ra csökkent, és a tesztelés időtartama 120 percről 57 percre csökkent. Ezen eredmények alapján a PISA szakértői csoportja a többszakaszos adaptív tesztek részleges bevezetését javasolta a 2015-ös méréstől kezdődően (*OECD*, 2012).

Kingsbury (2004) az NCLB-törvény előírásainak megfelelő teszt kidolgozásával kapcsolatban végzett vizsgálatot adaptív tesztekre vonatkozóan. A szimulációs kísérlet 4. és 8. évfolyamos tanulók számára készült matematikai és szövegértési rögzített formátumú teszteket hasonlított össze adaptív verziójukkal. A kutatás eredményei szerint az adaptív teszt mindkét szinten több információt, vagyis pontosabb mérési eredményt szolgáltatott, mint a hagyományos verzió.

Azonban egy jól működő feladat- vagy részteszt szintű adaptív rendszer kidolgozása bonyolult és összetett feladat, miközben a rögzített formátumú tesztelésről az adaptív tesztelésre való átállás számos kérdést is felvet. A tanulmányban bemutatott kutatásnak négy célja volt: (1) egy 5–8. évfolyamos tanulók induktív gondolkodás-fejlettségi szintjének meghatározására – többszakaszos adaptív tesztelés használatával – alkalmas itembank összeállítása, (2) a diákok hagyományos, rögzített formátumú teszten és részteszt szintű, adaptív teszten elért teljesítményeinek összehasonlítása, (3) az adaptív tesztelés során kiosztott itemek, illetve résztesztek nehézségi szintjének, ennek változásmintázatának jellemzése, valamint (4) a rögzített és az adaptív tesztelés során kinyert információ és a mérési hiba nagyságának összehasonlítása képességszint szerinti bontásban.

Módszerek

Minta

Az adatfelvétel 2012 őszén 158 5–8. évfolyamos diák részvételével zajlott. A diákok 45%-a volt fiú. A minta évfolyam és nem szerinti eloszlását az 1. táblázat tartalmazza.

1. táblázat. A minta évfolyam és nem szerinti eloszlása

Évfolyam	*N (fő)*	*Nemek aránya*
5.	22	1,64
6.	44	1,52
7.	51	1,59
8.	41	1,49

Megjegyzés: fiú: 1, lány: 2.

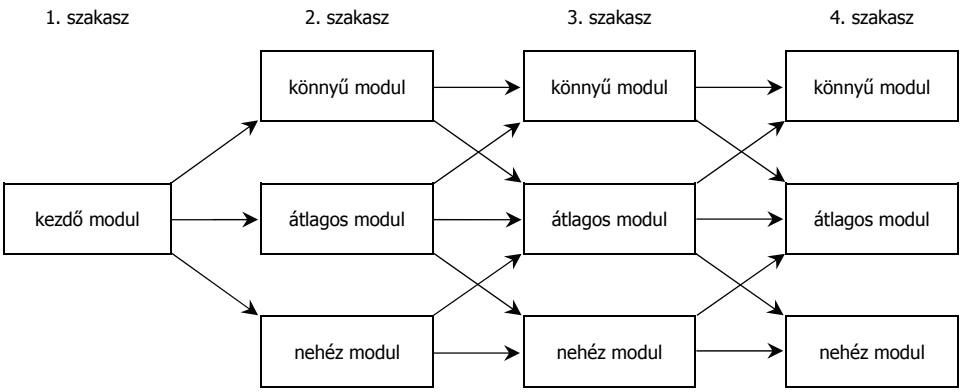
Mérőeszköz

Az induktív gondolkodás fejlettségét korábbi kutatásokban már gyakran alkalmazott itemek (l. pl. *Molnár* és *Csapó*, 2012) felhasználásával mértük. A több teszt feladatait tartalmazó, a korábbi kutatási eredmények alapján kalibrált itembank 96 verbális és nonverbális itemet tartalmaz.

Adatfelvétel és eljárások

Az adatfelvétel első fázisában minden osztályt véletlenszerűen kettéosztottunk. A diákok egyik része egy 28 itemből álló (n=79), rögzített formátumú induktívgondolkodásztesztet oldott meg, a másik fele egy szintén 28 itemes, négyszakaszos, 1-3-3-3 szerkezetű, többszintű adaptív tesztet. Két hét elteltével, az adatfelvétel második fázisában, a korábban rögzített formátumú tesztet megoldó tanulók adaptív tesztet, az adaptív tesztet megoldó tanulók rögzített formátumú tesztet kaptak. A diákok számára rendelkezésre álló idő mindkét esetben 45 perc volt. Az itemek nehézségi paraméterei -4,3 és +4,3 logitegységszint között mozogtak. A rögzített formátumú teszt feladatai széles képességtartomány vizsgálatát célozták meg, azaz a teszt különböző nehézségű feladatokból állt. A tanulók képességszintjének megállapítása a Rasch-modell segítségével történt, majd a logitegységben adott értékeket egy 500 pontos átlagú és 100 pontos szórású skálára transzformáltuk (az eljárásról részletesebben l. *Molnár*, 2013).

A többszakaszos adaptív teszt kezdő résztesztje (1. szakasz) 10, szélesebb itemnehézségi skálán mozgó itemet tartalmazott. A 2–4. szakasz résztesztjei 6-6 itemből álltak, és minden szakaszon belül három különböző nehézségi szintű résztesztet tartalmaztak. Összesen 10 különböző részteszt (modul) kialakítása történt, melyből 17 különböző tesztváltozat összeállítására volt lehetőség. A résztesztek (modulok) egymáshoz való viszonyát és az egész tesztelés során elfoglalt helyét az 1. ábra szemlélteti.



1. ábra
A többszakaszos adaptív teszt szerkezete

A modulok közötti elágazási szabály meghatározása az NC-módszer (*Number Correct*) segítségével történt (*Zenisky* és *Hambleton*, 2004). Minden feladatban a helyes válaszért 1, a helytelen vagy hiányzó válaszért 0 pontot adott a rendszer. Az első és a második modul közötti elágazásnál az osztópontot a kezdő modul tesztkarakterisztikus görbéje (erről részletesebben l. *Molnár*, 2013) alapján számítottuk ki a korábbi adatfelvételek eredményeire alapozva, így a 4-nél kevesebb pontot elért tanulók a könnyű modult kapták, a 7-nél több pontot elérők a nehezet (457 és 543 képességpontok által meghatározott értékek). A második és a harmadik elágazás esetén is hasonló módon határoztuk meg az osztópontokat: a könnyű modulnál 0–4 pont elérésekor a könnyű modul felé, 5–6 pont esetén a közepesen nehéz modul felé ágazott el a teszt. A legnehezebb részteszten 0–2 pontot elérő diákok a tesztelés következő szakaszában a közepes nehézségű modult kapták, míg a 3–6 pontot teljesítők maradtak a legnehezebb itemeket tartalmazó tartományban. A közepes (átlagos) nehézségű résztesztet megoldók három irányban léphettek tovább. A 0–2 pontot elérők a legkönnyebb, a 3–4 pontot teljesítők a közepes nehézségű, az 5–6 pontot kapott diákok a legnehezebb feladatokat tartalmazó részteszt felé léptek tovább. Ezzel a módszerrel, előzetes hipotézisünk szerint, a negyedik szakasz végére három egyenlő részre osztottuk a diákokat képességszintjük szerint.

A rögzített és az adaptív teszten elért teljesítmények összehasonlíthatóságát horgonyitemek biztosították. A bevezető tesztben 10, a többi modulban 2-2 horgonyitem szerepelt. Ezáltal bármelyik útvonalon ment végig a tanuló, 16 horgonyitem biztosította az eredmények közös képességskálán való megjeleníthetőségét. A tesztekből kinyerhető információ nagyságát a tesztinformációs görbékkel jellemeztük (erről részletesebben l. *Molnár*, 2013), ami a tesztből kinyert információ nagyságát a tesztet megoldó tanulók átlagos képességszintje és az itemek nehézségi szintje közötti különbségek segítségével jellemzi. A kinyert információ nagyságát akkor tekintettük maximálisnak, ha a feladatok nehézségi szintje és az azokat megoldó diákok képességszintje azonos. Minél távolabb van egymástól ez a két érték, annál kisebb a tesztelés során kinyert információ nagysága.

Az empirikus vizsgálat eredményei

A tesztek reliabilitása

Az eredmények kiterjeszthetőségének, általánosíthatóságának körét első szinten jól jellemzi a teszt reliabilitásmutatójának értéke, aminek meghatározására rögzített formátumú teszt esetén a Cronbach-α-t, adaptív teszt esetén ennek kiterjesztését, a személyszeparációs reliabilitásmutatót használtuk. Az adaptív teszt WLE (*Weighted Likelihood Estimate*) személyszeparációs reliabilitásmutatója 0,85, ami magasabb, mint a rögzített formátumú teszt megbízhatósági mutatója (Cronbach-α=0,83). A reliabilitásmutatók alapján megállapítható, hogy a kidolgozott itembank megbízhatóságát tekintve alkalmas 5–8. évfolyamos diákok induktív gondolkodásának, e gondolkodás fejlettségi szintjének meghatározására.

A diákok rögzített formátumú teszten mutatott teljesítménye (átlag=500, szórás=100) és az adaptív teszt alapján számolt képességszintje (átlag=489, szórás=100) erősen korrelált egymással ($r=0,82$, $p<0,01$). A két tesztkörnyezetben egymástól függetlenül meghatározott, azonos diákra vonatkozó képességszintek átlagosan azonosnak tekinthetők, ugyanakkor a korrelációs együttható nagysága eltérésekre is utal.

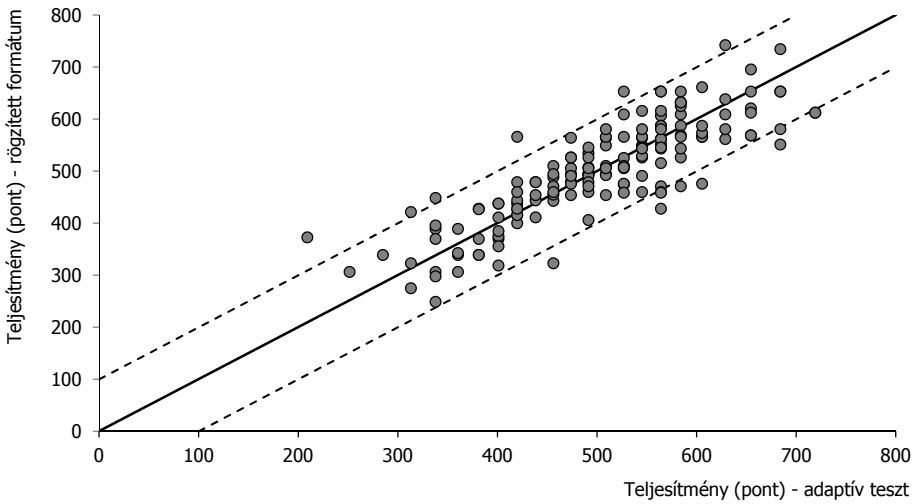
A becsült képességszintek összehasonlítása évfolyamonként és személyenként

A diákok rögzített, illetve adaptív tesztkörnyezetben mutatott teljesítményének alapstatisztikai mutatóit évfolyamonkénti bontásban a 2. táblázat mutatja. Az átlagosan legmagasabb képességszintű, 8. évfolyamos diákok esetén volt szignifikáns különbség a rögzített, illetve adaptív tesztkörnyezetben becsült képességszint között. Az 500 pont körül, azaz átlagosan teljesítő diákok között nem, miután a rögzített formátumú teszt is, hagyományos módon, elsősorban az ő képességszint-mérésüket célozza.

2. táblázat. Rögzített, illetve adaptív tesztkörnyezetben mutatott teljesítmények átlaga és szórása évfolyamonkénti bontásban

Évfolyam	*N*	*Rögzített formátumú teszt*		*Adaptív teszt*		*t*	*p*
		átlag	szórás	átlag	szórás		
5.	22	439	90	432	95	-0,55	0,59
6.	44	510	94	500	96	-1,35	0,18
7.	51	498	106	491	99	-0,76	0,45
8.	41	524	93	547	86	2,60	0,01

A 2. ábra a két tesztkörnyezetben nyújtott teljesítmények diákszintű összehasonlítását ábrázolja. Ha a diák képességszintje tesztkörnyezettől függetlenül számszerűen ugyanannak bizonyult, akkor a diákot reprezentáló alakzat a folytonos vonalon helyezkedik el. Amennyiben megállapított képességszintje nem különbözik egymástól szignifikánsan rögzített és adaptív környezetben, az őt reprezentáló jel a szaggatott vonalakon belül található. A szaggatott vonalak által képzett sávon kívül elhelyezkedő diákok számára a rögzített vagy az adaptív tesztkörnyezet bizonyult kedvezőbbnek. Előfordulásuk elenyésző a mintában, azaz különböző mérési hiba alkalmazása mellett, de szignifikanciaszinten belüli, közel azonos képességszint-becslést végeztünk rögzített, illetve adaptív tesztkiosztással.

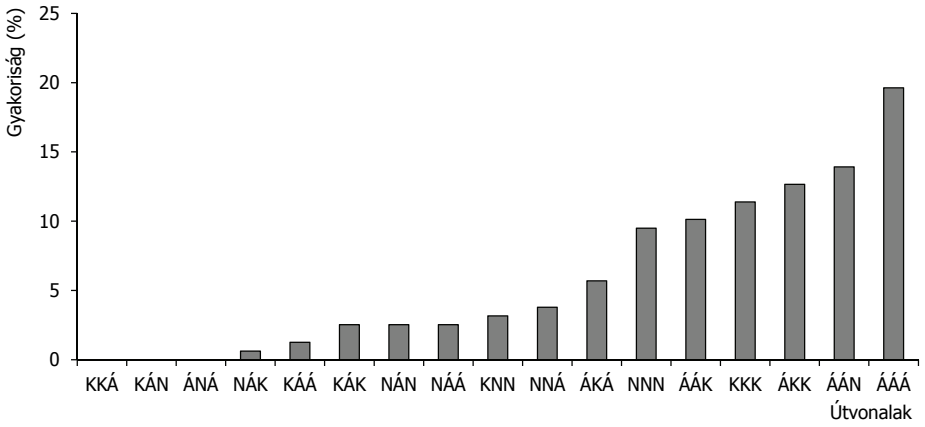


2. ábra
A rögzített formátumú és az adaptív teszten nyújtott teljesítmények összehasonlítása diákonkénti bontásban

A rögzített és az adaptív tesztelés során kiosztott résztesztek nehézségi szintjének változásmintázata

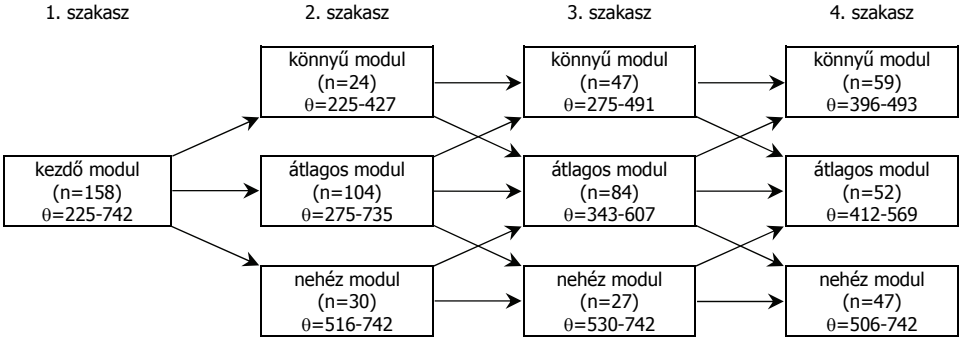
Az adatfelvétel során a többszakaszos adaptív teszt esetében a négy szakaszból összeállítható 17 különböző teszt közül 14-et osztottunk ki (3. ábra). Az esetek ötödében a résztesztek nyújtott teljesítmények alapján kizárólagosan az átlagos nehézségi szintű feladatokból álló tesztet közvetítettük ki. 11%-ban a kizárólagosan könnyű és közel 10%-ban a kizárólagosan nehéz résztesztekből álló, nehézségi szint tekintetében homogén teszteket vettük fel. Mindezek alapján megállapítható, hogy a diákok 40%-a a kezdő résztesztek nyújtott teljesítménye alapján egyértelműen besorolható volt a három képességsáv egyikébe. Egyetlen egy diák esetében fordult elő, hogy két nehézségi szintet is ugrott a tesztelés folyamán. A kezdő modul után megállapított képességszintje a legmagasabb képességtartományba sorolta őt, ugyanakkor a tesztelés végére átkerült az átlagosnál alacsonyabban teljesítő diákok csoportjába.

A diákok kétharmadának viselkedését jól jellemzi a hat leggyakoribb útvonal, melyek között a három azonos nehézségi szintű modulból álló tesztek mellett szerepel két olyan útvonal, ahol a két átlagos nehézségi szintű részteszt után a teljesítmények alapján a záró modulon a nehéz, illetve a könnyű részteszt irányába ágazott el a rendszer. A négy szakaszból álló adaptív tesztelési modell előnye a háromszakaszos modellhez képest az átlagos képességszinthez közeli, ugyanakkor azt vagy nem elérő, vagy kicsit túlteljesítő tanulók pontosabb képességszint-meghatározásában mutatkozott meg.



3. ábra
Az adaptív tesztrendszeren belül a második, harmadik és negyedik szakaszban kiosztott útvonalak gyakorisága (K: könnyű, Á: átlagos, N: nehéz modul)

A szakaszokon belül a tanulók modulonkénti eloszlását képességszint és gyakoriság szerinti bontásban a 4. ábra szemlélteti. A kezdő modulon mutatott teljesítmény alapján a tanulók háromötöd része a teszt második szakaszában közepes nehézségű résztesztet kapott, majd a teszt harmadik szakaszában mutatott teljesítmények alapján a teszt negyedik szakaszában közel azonos módon oszlottak el az átlagos (34%), az átlagnál alacsonyabb (37%) és az átlagnál magasabb (29%) képességszintű diákok. A rögzített formátumú teszttel ellentétben, ahol állandó volt a teszten belüli könnyebb, átlagos és nehezebb feladatok aránya, az adaptív feladatkiosztás során a magasabb képességszintű diákok nagyobb arányban kaptak nehezebb, míg az alacsonyabb képességszintű diákok könnyebb feladatokat. A teszt utolsó szakaszában, a diákokat közel harmadolva azonos átfedéssel, egyértelműen kialakult a három képességsáv.

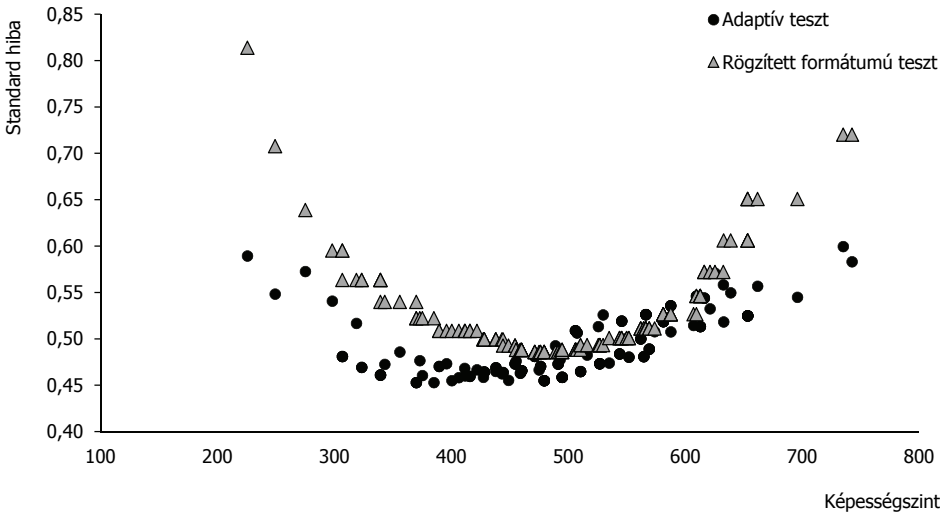


4. ábra
A tanulók gyakorisági és képességszint szerinti eloszlása a szakaszokon és a modulokon belül

A rögzített és az adaptív tesztelés során kinyert információ és a mérési hiba nagyságának összehasonlítása

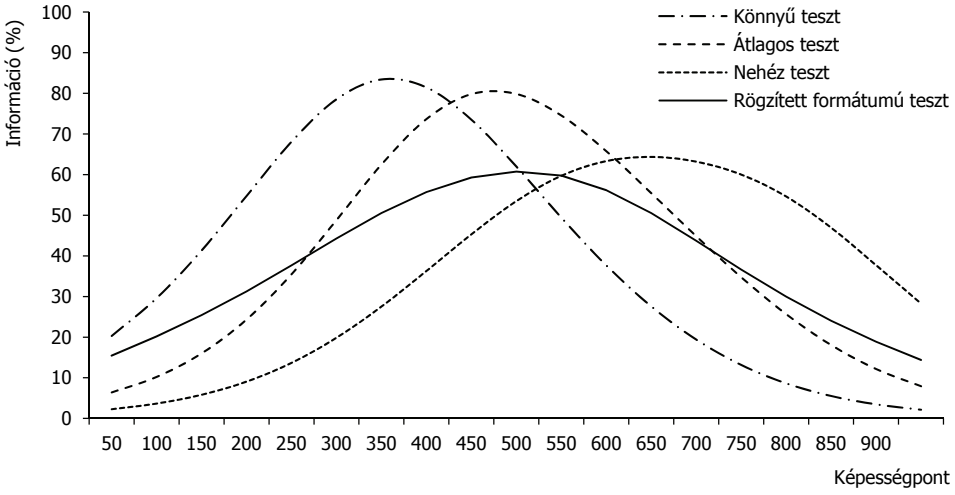
A kutatás felépítése diákszinten lehetővé teszi a becsült képességszintek pontosságának összehasonlítását. Diákonkénti bontásban a rögzített, illetve az adaptív tesztkörnyezetben történt képességszint-becslés során elkövetett hiba nagyságát összehasonlítva (*Wang*, 2001, 2010) megállapítható, hogy a rögzített formátumú teszt alapján történt képességszint-becslés hibáinak nagysága diákszinten átlagosan nagyobb ($t=-7{,}54$, $p<0{,}01$; se_átlag=0,53), mint ugyanazon diákok adaptív tesztkörnyezetben történt képességszint-becslésének hibája (se_átlag=0,49). A teljes minta vonatkozásában pontosabban, kisebb mérési hibával történt adaptív tesztkörnyezetben a diákok képességszintjének becslése. Hipotézisünk alapján azonban a hiba nagysága nem egyenletesen oszlik el a teljes képességskálán: különböző mintázat várható az alacsonyabb, az átlagos és a magasabb képességtartományban.

Összehasonlítva a rögzített formátumú és az adaptív teszten elért eredmények sztenderd hibáinak változását (5. ábra), hipotéziseinknek megfelelően, az alacsonyabb és a magasabb képességtartományban nagyobb hibával mér a rögzített formátumú teszt, mint az átlagos képességszintű diákok körében. Ez azt jelenti, hogy az adaptívteszt-algoritmus alkalmazásának előnye ezen képességtartományok esetén a legjelentősebb, átlagos képességszintű diákok mérése során közel azonosnak bizonyult a két tesztkörnyezetben becsült képességszintek sztenderd hibáinak alakulása.



5. ábra
Az adaptív és a rögzített formátumú lineáris teszt standard hibáinak alakulása a tanulók képességszintjének függvényében

A pontosság egy másik mutatója a tesztelés során kinyert információ nagysága, amit jelen esetben a (rész)teszteken nyújtott teljesítmények alapján felrajzolt tesztinformációs görbék segítségével jellemzünk. A 6. ábra görbéi grafikusan szemléltetik, hogy már akár a kizárólagosan könnyű, átlagos, illetve nehéz modul részteszjeiből összeállított tesztek (a 17 tesztváltozatból csak 3) is több információt szolgáltattak a tesztelés során, mint az egyetlen, sokféle nehézségű feladatot tartalmazó rögzített formátumú teszt. A görbéket a 4. ábrán ismertetett képességszintekkel egybevetve megállapítható, hogy minden képességszinten több információt tudtunk kinyerni adaptív tesztek alkalmazásával (a görbék minden esetben a rögzített formátumú teszt által adott információs függvény felett futnak az érintett képességtartományokban). A többletinformáció már abban az esetben is kimutatható, ha a diákokat az első részteszten nyújtott teljesítményük alapján három csoportba soroljuk, majd az adott képességtartományhoz közeli nehézségi szintű feladatokból állítjuk össze a tesztet.



6. ábra
Az adaptív technikával összeállított és a rögzített formátumú, azonos nehézségű részteszteket tartalmazó tesztek információs függvényei

Összegzés

A vizsgálat során ugyanazon mintán ugyanazon konstruktum mérése kapcsán összehasonlítottuk a számítógépes, rögzített formátumú tesztelés és a többszakaszos adaptív tesztkiosztás során becsült képességszinteket, a becslés pontosságát, valamint a kinyert információ nagyságát. A kutatás célja annak vizsgálata volt, hogy a hagyományos rögzített formátumú tesztekről az adaptív tesztelésre való átállás biztosítja-e és milyen mértékben a nagyobb mérési precizitás elérését. A kutatásban egy – ugyanazon a mintán fel-

vett – 28 itemes rögzített formátumú és egy összességében 28 itemes, ám 1-3-3-3 szerkezetű négyszakaszos adaptív teszten elért teljesítményeket, a becsült képességszinteket, azok sztenderd hibáit, a tesztelés során kinyert információ nagyságát, valamint a tesztek jóságmutatóit hasonlítottunk össze.

Az eredmények alapján megállapítható, hogy a teljes minta szintjén több információt nyerünk ki a tesztelés során, szignifikánsan pontosabb képességszint-meghatározást végzünk adaptívteszt-algoritmus alkalmazásával, mint hagyományos, rögzített formátumú teszteléskor. A kinyert információ százalékos nagyságát összehasonlítva, míg a rögzített teszt átlagosan 60%-os információt szolgáltatott, addig az adaptív tesztelés során kinyert átlagos információ nagysága 76% volt. Az eltérés elsősorban az alacsony és a magas képességszintű tanulók esetében volt jelentős, előbbi esetén közel 34%, utóbbi során közel 24%-kal több volt az adaptív tesztből kinyert információ mennyisége. A képességszintek becslése során elkövetett hiba nagysága is ezzel párhuzamosan csökkent az adaptívteszt-algoritmus alkalmazása során. A diákok képességszintjéhez igazodó tesztelési eljárás hatékonysága már abban az esetben is jelentős, ha egy rövid, megfelelő jóságmutatókkal rendelkező kezdő résztesztен elért eredmény alapján három képességcsoportba soroljuk a diákokat, majd mindenki a saját képességcsoportjának képességtartományában legtöbb információt szolgáltató, rögzített formátumú tesztet kap.

A kutatás eredményei több szempontból korlátozottak – a minta és az itembank mérete, struktúrája, az alkalmazott itemek tartalmi és pszichometrikus jellemzői – mely korlátok nagyban befolyásolhatják a képességszintek becslését, ezért további kutatások szükségesek különböző méretű és tartalmi lefedésű itembankok felhasználásával a kinyert információ mértékének pontosabb meghatározására. A kutatás egyedisége, hogy az adaptív tesztelés hatékonyságát vizsgáló legtöbb kutatással szemben nem szimulált adatbázison, hanem empirikus adatok segítségével hasonlítottuk össze a rögzített és az adaptív tesztkörnyezetben becsült képességszintek alakulását, továbbá az azonos minta alkalmazása lehetővé tette a diákszintű összehasonlítást is. Az eredmények alátámasztották a szimulációs kísérletekben is tapasztaltakat, miszerint jelentős mértékű mérési precizitás érhető el akár már három nehézségi szintet megkülönböztető adaptívteszt-algoritmus alkalmazásával a hagyományos lineáris tesztekhez képest.

A kutatást a TAMOP 3.1.9/11 kutatási program, az Oktatáselméleti Kutatócsoport és az MTA-SZTE Képességfejlődés Kutatócsoport támogatta.

Irodalom

Al-A'ali, M. (2007): Implementation of an improved adaptive testing theory. *Educational Technology & Society*, **10**. 4. sz. 80–94.

Amstrong, R. D. (2002): *Routing rules for Multiple-Form Structures*. (Computerized Testing Report 02-08). Law School Admission Council. http://www.lsac.org/lsacresources/research/ct/pdf/ct-02-08.pdf. Utolsó letöltés: 2013. április 07.

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

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

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.

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.

Frey, A., Seitz, N. N. és Kröhne, U. (2011): Reporting differentiated literacy results in PISA by using multidimensional adaptive testing. In: Prenzel, M., Kobarg, M., Schöps, K. és Rönnebeck, S. (szerk.): *Research in the context of the Programme for International Student Assessment*. Springer, Berlin. 1–33.

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.

Kingsbury, G. G. és Hauser, C. (2004): Computerized adaptive testing and no child left behind. Előadás, Annual Meeting of the American Educational Research Association, San Diego, CA.

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

Magyar Andrea (2012): Számítógépes adaptív tesztelés. *Iskolakultúra*, **22**. 6. sz. 52–60.

Molnár Gyöngyvér (2011): Az információs-kommunikációs technológiák hatása a tanulásra és oktatásra. *Magyar Tudomány*, 9. sz. 1038–1047.

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

Thompson, T. és Way, D. (2007): Investigating CAT designs to achieve comparability with a paper test. In: Weiss, D. J. (szerk.): *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*. www.psych.umn.edu/psylabs/CATCentral/. Utolsó letöltés: 2013. április 04.

Tian J., Miao, D. és Zhu Xia, G. J. (2007): An introduction to the computerized adaptive testing. *US-China Education Review*, **4**. 1. sz. 72–81.

Tóth Zoltán (2005): A tudásszerkezet és a tudás szerveződésének vizsgálata a tudástér-elmélet alapján. *Magyar Pedagógia*, **105**. 1. sz. 59–82.

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

Wang, T. és Kolen, M. J. (2001): Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement*, **38**. 1. sz. 19–49.

Wang, H. (2010): Comparability of computerized adaptive and paper-pencil tests. *Test, measurements and research services bulletin.* http://www.pearsonassessments.com/NR/rdonlyres/057A4A04-9DCB-4B68-9CB0-3F32DDF396F6/0/Bulletin_13.pdf. Utolsó letöltés: 2013. április 07.

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. (szerk.): *Elements of adaptive testing.* Springer, New York. 355–372.

ABSTRACT

ANDREA MAGYAR AND GYÖNGYVÉR MOLNÁR: COMPARING THE EFFICACY OF COMPUTERIZED ADAPTIVE AND FIXED-ITEM TESTING

With the rapid spread of computers in the past two decades, linear paper-and-pencil tests are gradually being replaced by computer-based assessment. The most advanced form is computerized adaptive testing, in which the test is adapted to examinees' ability level by only administering items of appropriate difficulty. The aim of this paper is to compare the effectiveness of fixed-item and adaptive tests from an assessment perspective by: (1) relating differences in student level achievement; (2) outlining item difficulties of delivered tests; and, finally, (3) comparing measurement error and test information functions in linear and adaptive test environments. The samples from the pilot study were drawn from children in Years 5 and 8 at Hungarian primary schools (N=158). A fixed-item test was administered to half of the participants; the other part took four-stage adaptive tests (1-3-3-3 structure). Two weeks later, the types of test were switched. Both tests measured inductive reasoning. A one-parameter Rasch model was used for the analyses. The reliability of the adaptive tests proved to be higher (Cronbach-α=.85) than that of the fixed test form (Cronbach-α=.83). The adaptive test provided consistently higher information at every skill level than that of the fixed-form test. The standard error of the four-stage test was significantly lower, especially in upper and lower ability levels. The study provided a promising step towards more precise educational assessment in using multistage testing with even three stages besides traditional linear test forms.

Magyar Pedagógia, **113**. Number 3. 181–193. (2013)

Levelezési cím / Address for correspondence: Magyar Andrea, SZTE Neveléstudományi Doktori Iskola, Molnár Gyöngyvér, SZTE Neveléstudományi Intézet, MTA-SZTE Képességfejlődés Kutatócsoport H–6722 Szeged, Petőfi S. sgt. 30–34.