SZEGEDI TUDOMÁNYEGYETEM
BÖLCSÉSZETTUDOMÁNYI KAR
NEVELÉSTUDOMÁNYI DOKTORI ISKOLA
INFORMÁCIÓS ÉS KOMMUNIKÁCIÓS TECHNOLÓGIÁK AZ OKTATÁSBAN DOKTORI PROGRAM

MAGYAR ANDREA

# SZÁMÍTÓGÉP ALAPÚ ADAPTÍV ÉS LINEÁRIS TESZTEK ÖSSZEHASONLÍTÓ HATÉKONYSÁGVIZSGÁLATA

PHD ÉRTEKEZÉS TÉZISEI

TÉMAVEZETŐ:
MOLNÁR GYÖNGYVÉR PHD, HABIL EGYETEMI DOCENS

SZEGED
2015

# BEVEZETÉS

A 20. században leginkább elfogadott és elterjedt papír alapú (PP) tesztekre alapuló mérések fejlesztése egyre több korlátba ütközött, a papír-alapú tesztekre alapozott fejlesztés lehetőségei fokozatosan kimerültek (*Molnár* és *Magyar*, 2015). A továbblépéshez, a 21. században jelentkező új mérés-értékelési igények kielégítéséhez alapvető, minőségi változtatásra van szükség (*Scheuermann* és *Pereira*, 2008; *Beller*, 2013). A technológia rohamos fejlődésével a továbblépés iránya egyértelműen a számítógépes tesztelésre való átállás (*Scheuermann* és *Björnsson*, 2009; *Molnár*, 2010; *Csapó*, *Ainley*, *Bennett*, *Latour* és *Law*, 2012), mely számos új lehetőséget kínál a papír alapú teszteléssel szemben. Ilyen például a motiválóbb környezet (*Thompson* és *Pometric*, 2007), az azonnali kiértékelés lehetősége (*Wang*, 2010), az innovatív, multimédiás elemeket is tartalmazó dinamikusan változó itemek megjelenítése (*Greiff*, *Wüstenberg* és *Funke*, 2012), illetve a személyre szabott, adaptív tesztelés megvalósítása (*Eggen* és *Straetmans*, 2000).

Adaptív tesztelés (CAT – *Computerized Adaptive Testing*; *Weiss, 2011*) alkalmazása során a teszt feladatai nem előre meghatározott fix sorrendben követik egymást, hanem azok egy feladatbankból kerülnek kiválasztásra a tesztmegoldó képességszintjéhez igazítva, a korábbi feladaton nyújtott teljesítménye alapján. Például feladatszintű adaptivitás esetén 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.

Ez a típusú feladatadás és tesztösszeállítás a hagyományos, mindenki számára azonos itemeket azonos sorrendben tartalmazó, lineáris tesztekkel szemben a teljesítmények sokkal finomabb mérését teszi lehetővé (*Linacre*, 2000; *Magyar* és *Molnár*, 2013). Jelentős mértékben megnő a tesztelés során kinyerhető itemekre és személyekre vonatkozó információ nagysága (*Molnár*, 2013; *Magyar* és *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 teremt 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 és Way*, 2007), ezzel párhuzamosan a teszt megoldásához szükséges idő is jelentős mértékben rövidül, utóbbi átlagosan felére csökken (*Frey* és *Seitz*, 2009; *Frey, Seitz* és *Kröhne*, 2011).

Miután jelenleg a számítógép alapú tesztelére való átállás fázisában vagyunk, indokoltak a médihatást tanulmányozó összehasonlító hatékonyságvizsgálatok, melyek egyrészt longitudinális kutatásokban kapnak kiemelkedő szerepet, ahol szükséges a korábbi papíralapú adatfelvételek eredményeinek számítógép alapú teszteredményekkel való összehasonlíthatósága, másrészt abban az esetben, amikor a kétféle médiumon való tesztelés alternatív módon párhuzamosan folyik.

Adaptív tesztek bevezetése során számos esetben végeztek adaptív és a lineáris változat összehasonlító hatékonyságvizsgálatot (*Al-A'ali*, 2007; *Brossman* és *Guille*, 2014; *Crotts*, *Zenisky*, *Sireci* és *Li*, 2013; *Frey, Seitz* és *Kröhne,* 2011; *Guille, Becker, Zhu, Zhang, Song* és *Sun*, 2011; *Hambleton* és *Xing*, 2006; *Jodoin*, *Zenisky* és *Hambleton*, 2006; *Kingsbury* és *Hauser*, 2004; *Olea*, *Revuelta*, *Ximénez* és *Abad*, 2000; *Pyper* és *Lilley*, 2010; *Rotou, Patsula, Manfred* és *Rizavi*, 2003; *Thompson* és *Way*, 2007; *Vispoel, Hendrickson* és *Bleiler*, 2000; *Zheng*, 2012), az eddigi kutatások azonban főként szimulált adatbázisokkal dolgoztak. Empirikus kutatások elsősorban az egyetemista korosztály körében folytak, melyek a legtöbb esetben kis mintán végzett pilotvizsgálatok voltak.

Ezt a hiányt pótolja az itt bemutatott kutatássorozat, melynek fő célja az adaptív tesztelés hatékonyságának vizsgálata az 1-8. évfolyamos tanulók körében. A kutatás elsősorban az adaptív tesztelés technikai jellemzőire fókuszál, a számítógépes adaptív tesztek kidolgozásának menetét, alkalmazását és működését veszi górcső alá, továbbá azt vizsgálja, hogy az általános iskolai korosztály körében osztálytermi környezetben milyen előnyöket jelenthet alkalmazásuk.

## AZ ÉRTEKEZÉS ELMÉLETI FORRÁSAI

Az adaptív tesztek szigorú algoritmus szerint működnek (*Linacre*, 2000; *Magyar*, 2012). A kezdő item/részteszt kiválasztása többféleképpen történhet: amennyiben rendelkezésre áll képességszintre vonatkozó előzetes információ a tesztelt személyről, akkor egy becslési algoritmus felhasználásával már a teszt kezdő iteme személyre szabott lehet, ha nem, akkor az itembankból vagy annak részhalmazából történhet véletlenszerű kiválasztással. Rendszerint egy közepes nehézségű itemmel indul a tesztelés. Az adott item megoldása után egy újabbat választ ki a rendszer. Ha a személy jól válaszolt, egy nehezebbet kap, amennyiben hibázott, akkor könnyebbet. A program algoritmusa biztosítja, hogy minden soron következő item a személy képességeihez mért legyen. A megoldott itemeket a gép értékeli, és dönt arról, hogy szükséges-e új item kiválasztása, vagy a tesztelés véget ért. A tesztelés végén a tanuló azonnali visszajelzést kap elért eredményéről (*Csapó*, *Molnár* és *R. Tóth*, 2008; *Eggen,* 2004). Ez alapján a CAT individualizált teszt. A tanulók különböző itemekkel kezdhetik és folytathatják a tesztet, és különböző lehet a megoldott itemek száma is, így a CAT dinamikus, személyre szabott, a tanuló képességszintjéhez igazodó tesztelési mód (*Weiss, 2011*).

Az adaptív tesztek számos változata létezik (*van der Linden*, 2008), az itemalapú tesztektől a lineáris alteszteket alkalmazó többszakaszos tesztekig, az alapelvet tekintve azonban mindegyik adaptív teszt hasonlóan épül fel. Az itemalapú adaptív tesztek számos előnyük ellenére hátrányokkal és korlátokkal is rendelkeznek. A legtöbb itemalapú tesztben nincs lehetőség a feladatok utólagos áttekintésére és javítására, a véletlenszerű itemkiosztás miatt az előző item információt szolgáltathat a következő item számára, valamint az itemek elhelyezkedése is befolyásolhatja a megoldást; ugyanaz az item a tesztben elfoglalt helyétől függően lehet könnyebb, vagy nehezebb (*Wainer* és *Kiely*, 1987; *Wainer*, 2000; *Linacre,* 2000).

Ezek közül több probléma kiküszöbölését oldja meg a többszakaszos adaptív teszt (MST – *Multi Stage Test*; *Magyar*, 2013), mely egyesíti magában a hagyományos lineáris és az adaptív tesztek tulajdonságait, egyrészt a kérdéseket a tanuló képességszintjéhez igazítja, másrészt lehetőséget ad az itemek sorrendjének előzetes meghatározására (*Amstrong, Jones, Koppel* és *Pashley*, 2004; *Molnár*, 2013). Szerkezetét tekintve a hagyományos lineáris és az itemalapú adaptív tesztek között félúton helyezkedik el (*Jodoin, Zenisky* és *Hambleton*, 2006; *Patsula*, 1999; *Zheng*, 2012). A teszt több szakaszban itemek helyett modulokat tartalmaz, 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).

Az MST nagy változatosságot enged a szakaszok, a szakaszokon belül a modulok, és a modulokon belül az itemek számát illetően (*Davis*, 2005; *Yan*, *von Davier* és *Lewis*, 2014). 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 szakaszok számának növelésével azonban egyre összetettebbé válik a teszt adminisztráció a mérési precizitás arányos növekedése nélkül (*Amstrong* és mtsai, 2004; *Hendrickson*, 2007). Ezen meggondolások alapján a leggyakrabban 2-4 szakaszos tesztek alkalmazása terjedt el (*Zenisky*, *Hambleton* és *Luecht*, 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 szakaszonként (*Amstrong* és mtsai, 2004).

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; *Amstrong*, 2002; *Zenisky, Hambleton* és *Luecht*, 2010).

Az elágazási szabályokhoz szorosan kapcsolódik a modulok és a teljes teszt pontozása. Minden egyes tanuló képességszintje egy-egy szakasz végeztével becslésre kerül. Gyakran a becsült

képességpontokat NC (number-correct) pontszámokká konvertálja az algoritmus (*Keng*, 2008; *Yan*, *von Davier* és *Lewis*, 2014). Azonban, míg a modulok pontozására elegendő az NC pontozás, a teljes teszt pontozására nem megfelelő, mivel a tanulók statisztikailag különböző itemeket kapnak (*Zenisky*, *Hambleton* és *Luecht*, 2010). Ezért a teljes teszt pontozására az item-alapú adaptív teszteknél használatos módszerek alkalmazhatók a többszakaszos tesztek esetén is, a megfelelő valószínűségi tesztelméleti modellt alkalmazva (*Keng*, 2008).

Az itemalapú adaptív tesztekkel összehasonlítva a többszakaszos tesztek számos előnnyel rendelkeznek (*Magyar* és *Molnár*, 2013). A modulok előre tervezhetőek és szerkeszthetőek, így nagyobb kontrolt biztosítanak a teszt adminisztráció számára. Ezáltal kiküszöbölhető, 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 (*Zheng*, 2012). 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*, *Hendrickson* és *Bleiler*, 2000). Az itemalapú adaptív tesztekhez képest jóval kevesebb adminisztrációt és számítógépes számításokat igényelnek (*Hendrickson*, 2007; *Zheng*, 2012).

Előnyei mellett *Hendrickson* (2007) hangsúlyozza, hogy a többszakaszos tesztek bizonyos hátrányokkal is rendelkezhetnek. Általában több itemre van szükség az itemalapúval 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. További hátránya, hogy a teszt csak az adott modul végén érhet véget, így a teszt hossza kevésbé flexibilis az itemalapú adaptív tesztekhez képest (*Zheng*, 2012). Ezen hátrányok ellenére, az MST mégis egyensúlyt képvisel a pontosság, adaptivitás, gyakorlati használhatóság és az itemek feletti kontroll tekintetében (*Zenisky*, *Hambleton* és *Luecht*, 2010).

Amennyiben egy teszt többféle verziója létezik, fontos, hogy a különböző verziókon szerzett pontszámok összehasonlíthatóak legyenek egymással. Különösen lényeges ez longitudinális kutatások esetében, másrészt pedig abban az esetben, amikor a kétféle médiumon való tesztelés alternatív módon párhuzamosan folyik (*Way*, *Davis* és *Fitzpatrick*, 2006; *Paek*, 2005). A professzionális teszt standardok (*APA,* 1986; *AERA*, *APA* és *NCME*, 1999; *Wang, Jiao, Young, Brooks* és *Olson*, 2008) is hangsúlyozzák a különböző médiumokon elért pontszámok összehasonlíthatóságának fontosságát. Az összehasonlító kutatások fő fókuszában a tesztek mérési pontosságának összehasonlítása áll, és annak a feltárása, hogy az adaptív tesztelésre való átállás milyen hatással van a tesztelési folyamatra (idő, itemszám) és a különböző képességű egyének eredményeire.

Adaptív és papír alapú tesztek összehasonlítása különösen nagy kihívás (*Wang* és *Kolen*, 2001). Mivel a vizsgázók személyre szabott tesztet kapnak, különbségek lehetnek az itemek tartalmában, az itemek elhelyezkedésében és nehézségében, valamint a pontozásban. Ezek a tényezők jelentősen befolyásolhatják az összehasonlíthatóságot, melyeket a médiahatás mellett szintén javasolt figyelembe venni (*Wang* és *Kolen*, 2001; *Kolen*, 1999-2000). *Wang* és *Kolen* (2001) arra hívják fel a figyelmet, hogy ahhoz, hogy a CAT verzió összehasonlítható legyen a papír alapú verzióval, a CAT szempontjából nagymérvű korlátozottságot jelent, hiszen ez esetben az adaptív tesztfejlesztés során nem lehet az összes, a számítógép adta lehetőséget kihasználni. Mivel azonban jelenleg a papír alapú tesztelésről a számítógépen való tesztelésre való átállás fázisában vagyunk, és a trendvizsgálatok szempontjából szükség van az előző eredményekkel való összehasonlíthatóságra, ezért különösen indokoltak az ilyen irányú vizsgálatok (*Wang* és *Kolen*, 2001; *Pásztor-Kovács, Magyar, Hülber, Pásztor* és *Tongori*, 2013; *Wan*, *Keng*, *McClarty* és *Davis*, 2009).

Az adaptív tesztekkel foglalkozó kutatások küzül számos fókuszában az MST tesztek különböző konstrukcióinak viszgálata állt. A különböző típusok változatos formában fordultak elő az alkalmazott teszt tartalmától, az itembank méretétől, valamint az itemek jellemzőitől függően. A legtöbb esetben egy közepes modullal indult a tesztelés, majd 2-5 ágon folytatódott 2-6 szakasz alkalmazásával. A szakaszok és modulok számának növelése növelte a teszt precizitását. A legtöbb kutatás kettőnél több szakaszt alkalmazását javasolja, mivel így kiküszöbölhető a tanulók esetlegesen hibás szintre történő

besorolása, mert a későbbi szakaszokon korrigálhatóak az eltérések. Nagyon sok szakasz esetén azonban a mérési precizitás arányos növekedése nélkül indokolatlanul megnövekszik a teszt hossza. Gyakori típus az 1-3 (*Rotou* és mtsai, 2003), 1-2-3-4, 1-2-4 (*Zheng*, 2012), 1-3-3 (*Keng*, 2008), 1-2-2, 1-3-3, 1-2-3, 1-3-2 (*Jodoin*, *Zenisky* és *Hambleton*, 2006), 5-5-5-5-5-5 (*Crotts* és mtsai, 2013) és 1-3-3-3-3 (*Brossman* és *Guille*, 2014) szerkezetű MST. Több mérés esetén, amennyiben az alkalmazott itembank mérete engedte, a tesztbiztonság növelése érdekében több ekvivalens tesztváltozat, illetve modul került összeállításra, melyek random módon kerültek kiosztásra.

Az első vizsgálatoknál a tesztek összehasonlítására a klasszikus tesztelméleti módszereket használták (ANOVA elemzések, átlagok összehasonlítása; *Vispoel*, *Hendrickson* és *Bleiler*, 2000; *Olea* és mtsai, 2000) a későbbiekben azonban általánossá vált a valószínűségi tesztelméleti módszerek, eljárások alkalmazása, úgymint item- és tesztinformációk összehasonlítása. A tesztek mérési precizitásának mutatója a reliabilitás és a mérési hiba (SE - *standard error*), és szimulációk esetén gyakran használt mutató a valódi és a mért képességpontok korrelációs mérőszáma, valamint az RMSE (*Root Mean Square Error*), illetve AAD (*Average Absolute Difference*), melyek a valódi és a mért képességpontok eltéréseinek különböző mutatói (*Keng*, 2008).

A vizsgálatok megerősítik, hogy az adaptív tesztek jóval pontosabb képességmérést tesznek lehetővé, csökken a tesztelési idő, illetve az alkalmazott itemek száma. Az adaptív tesztek minden képességszinten több információt szolgáltatnak, mint a lineáris változat, és a mérési hiba is szignifikánsan csökken adaptív tesztkiosztás esetében. A legtöbb vizsgálat eredménye szerint a többszakaszos teszt precizitása valamivel elmaradt az item, illetve a tesztlet alapú tesztekétől, viszont a tesztfejlesztés folyamán alkalmazható nagyobb adminisztratív kontroll előnyössé teszi ezt a teszttípust, ezért papír alapú tesztelésről való átállás esetén a legpreferáltabb típusú adaptív tesztkonstrukció.

## AZ EMPIRIKUS VIZSGÁLATOK KONCEPCIÓJA

### *CÉLOK*

A kutatás fő célja az adaptív tesztelés hatékonyságának vizsgálata hagyományos, lineáris tesztelési móddal való összehasonlítása során az 1-8. évfolyamos tanulók körében.

A vizsgálat alcéljai:

1) a korábban papír alapon alkalmazott tesztek online formára konvertálása;

2) a számítógép alapú alapú itemekből osztálytermi környezetben használható lineáris, illetve adaptív tesztrendszerek kialakítása;

3) az adaptív és lineáris tesztek mérési pontosságának összehasonlítása;

4) a becsült képességszintek évfolyam és személyszintű összehasonlítása;

5) a kétféle tesztkörnyezetben elért helyes válaszok arányának összehasonlítása;

6) az adaptív tesztelés során kiosztott itemek, illetve résztesztek nehézségi szintjének, ennek változásmintázatainak jellemzése;

7) a lineáris és az adaptív tesztelés során kinyert információ és a mérési hiba nagyságának összehasonlítása képességszint szerinti bontásban.

### *KUTATÁSI KÉRDÉSEK*

A felvázolt célok alapján a következő kutatási kérdésekre keressük a választ:

1) A korábban papír alapon alkalmazott tesztek átkonvertálhatóak-e online tesztformátumra?

2) A számítógép alapú alapú itemekből kialakítható-e osztálytermi környezetben megbízhatóan alkalmazható adaptív tesztrendszer?

3) Az adaptív tesztrendszerek pontosabb mérést tesznek-e lehetővé, mint a lineáris tesztek?

4) Van-e különbség az adaptív, illetve lineáris tesztekkel becsült képességszintek között évfolyam, illetve személyszinten?

5) Hogyan alakul a kétféle tesztkörnyezetben a tanulók által elért helyes válaszok aránya?

6) Mely nehézségi szintű itemek/résztesztek szerepelnek leggyakrabban az adaptív tesztkiosztás során?

7) Hogyan alakul a kinyert információk és mérési hibák mértéke a kétféle tesztkörnyezetben?

### *Hipotézisek*

1) a korábban papír alapon alkalmazott tesztekből kifejleszthető többszakaszos adaptív tesztelésre alkalmas tesztrendszer;

2) az adaptív tesztrendszerek hatékonyan és megbízhatóan alkalmazhatóak az 1-8. évfolyamos korosztály diagnosztikus mérésére;

3) az adaptív rendszer a képességek pontosabb mérését teszi lehetővé;

4) a becsült képességszintek nem különböznek jelentős mértékben online adaptív és lineáris tesztelés esetén;

5) adaptív tesztelés esetén az alacsonyabb képességtartományban magasabb a helyes válaszok aránya, mint lineáris tesztkörnyezetben, átlag feletti tanulóknál viszont fordítva, kisebb arányban fordulnak elő helyes válaszok;

6) a tanulók többsége átlagos képességszintű, ezért adaptív tesztkiosztás esetében az átlagos nehézségű itemek/résztesztek szerepelnek leggyakrabban;

7) a tesztekből kinyert információ minden képességszinten szignifikánsan magasabb, a mérési hiba viszont szignifikánsan alacsonyabb az adaptív tesztrendszer esetében, mint lineáris teszteknél.

### *A vizsgálatsorozat menete*

A hatékonyságvizsgálatokat pilotmérések és nagymintás mérések alkalmazásával valósítottuk meg 2012 és 2014 között több részletben. Három pilotvizsgálatot és két nagymintás vizsgálatot folytattunk le, melyeket különböző képességeket mérő, korábban papír alapon alkalmazott tesztek itemeinek felhasználásával állítottuk össze. Az első pilotvizsgálat induktív gondolkodást, a második problémamegoldó képességet mérő tesztek felhasználásával történt. A harmadik mérés alkalmával szóolvasás készséget mérő tesztrendszert konvertáltunk online formára, melynek működését először kismintás, majd ezt követően nagymintás méréssel teszteltünk.

A két pilotmérés (induktív gondolkodás és problémamegoldás) alkalmával a papír alapon bemért paramétereket használtuk az adaptív tesztváltozatok összeállításánál, a szóolvasás mérésénél viszont külön nagymintás méréssel paramétereztük az itemeket, ezt követően állítottuk össze az adaptív tesztrendszert és valósítottuk meg a hatékonyságmérő nagymintás mérést. A mérések során az adaptív teszt mellett ugyanabból az itembankból összeállított lineáris tesztváltozat biztosította a mérések összehasonlíthatóságát. Az alkalmazott adaptív tesztek minden esetben többszakaszos tesztek voltak, az itembank méretétől és összetételétől függően különböző szerkezeti struktúrában.

### *A vizsgálat során használt mérőeszközök bemutatása*

A vizsgálathoz felhasznált mérőeszközök olyan képesség-, illetve készségterületek mérésére alkalmasak, melyek kiemelt szerepet játszanak az általános iskolai korosztály készség, illetve képességfejlődésében. Mindhárom mérőeszköz eredetileg papír alapon került kidolgozásra, és több alkalommal használták őket nagymintás mérések keretében a vizsgált korosztály készség, illetve képességmérésére, melyek során a tesztek nagyon megbízhatóan működtek, mindegyik teszt megbízhatósági mutatója Cronbach-alpha=0,80 felett volt. A papíralapú mérések alapján a tesztek minden iteme paraméterezésre került, a nehézségi paraméterek lefedték a vizsgált korosztály képességskáláját, ezért ezeket a paramétereket használtuk a kismintás mérések alkalmával is. A felhasznált tesztek kiválasztásának további indoka, hogy ezek a tesztek tartalmaztak elég itemet az adaptív teszteléshez megfelelő méretű itembank kialakításához. További fontos szempont volt, hogy az itemek számítógépes formátumra való konvertálása csekély változtatással megoldható volt, ezáltal a médiahatás nem befolyásolta jelentős mértékben a tesztek validitását.

Az induktív gondolkodást és a problémamegoldó képességet mérő tesztek online adaptív formára történő konvertálásakor az eredeti, papír alapon bemért paraméterek alapján állítottuk össze

a teszteket. A harmadik mérés során a mérési precizitás növelése érdekében online paraméterezést hajtottunk végre az itembank összes itemére vonalkozóan, és az így kapott paraméterek alapján állítottuk össze a pilot, illetve nagymintás mérés során használt adaptív tesztrendszert. Az induktív gondolkodást mérő teszt esetében 1-3-3-3 szerkezetű MST teszkonstrukciót alkalmaztunk, a problémamegoldó képesség esetében az 1-2-3 szerkezetű tesztekre esett a választásunk, mivel az előzetes szakirodalmi kutatások során kisebb itembankok esetén ezek megvalósítása tűnt a legmegfelelőbbnek. A harmadik kutatás esetén a szóolvasó készség mérése kapcsán az itembank nagy mérete és az eredeti tesztstruktúra komplikáltabb szerkezet alkalmazását indokolta, ezért ebben az esetben az 1-4-5-5 szerkezetű MST mellett döntöttünk.

### *A MINTÁK ÖSSZEÁLLÍTÁSÁNAK SZEMPONTJAI*

A kutatás célja szerint az általános iskolai korosztály körében vizsgáltuk az adaptív tesztelés bevezethetőségeinek lehetőségeit. Ezért a mérések során törekedtünk arra, hogy a rendelkezésre álló tesztek lehetőségeit kihasználva, minél nagyobb mértékben átfogjuk az általános iskolai korosztályt. Mivel a bemutatott tesztek alkalmasnak bizonyultak tág életkori határokon belül történő képességmérésekre, ezért ki tuduk terjeszteni a méréseket az alsó és a felső korosztály körére is. A mérésekben az általános iskolák 1-8. évfolyamos tanulói közül összesen 8165-en vettek részt. Az induktív és a problémamegoldó gondolkodás mérésére kidolgozott tesztekkel a felső tagozatosok vizsgálatát céloztuk meg. A szóolvasás teszt feladatainak megoldásába elsősorban az alsó tagozatos korosztályt vontuk be. A nagymintás hatékonyságvizsgálathoz szűkebb életkori intervallumot, a 4-5. évfolyamot választottuk, az életkori fejlettségből eltérő jellemzők kiküszöbölése végett.

### *ADATFELVÉTEL*

A tesztek minden esetben az eDia rendszer segítségével készültek és kerültek kiközvetítésre. A tanulók online módon, a saját iskolájukban, saját internethálózatukon keresztül oldották meg a feladatokat. A mérésekre minden esetben egy tanítási óra (45 perc) állt a diákok rendelkezésére. A segítő pedagógusok részletes mérési útmutatót kaptak az adatfelvétel lebonyolításának részletes leírásával. Mindegyik tesztet két változatban oldották meg a tanulók, az első fázisban véletszerűen kaptak adaptív vagy lineáris tesztet, a második fázisban fordítva történt a tesztkiosztás, aki az első fázisban lineáris tesztet oldott meg, az adaptívat kapott, és fordítva. A két fázis között minimum két, maximum négy hét eltérés volt. A mérés végén a rendszer visszajelzést adott a tanulók teljesítményéről. A lineáris teszteknél a teszten elért százalékos teljesítményt jelezte vissza a rendszer, az adaptív teszteknél viszont képességpontot számolt a program.

### *AZ ADATOK ELEMZÉSE*

A kapott adatok elemzése során elsősorban a tesztek technikai működését vizsgáltuk, a kutatás fő célja az itemjellemzők és az itemek viselkedésének összehasonlítása volt a kétféle tesztkörnyezetben. Az elemzések elvégzése egyrészt klasszikus tesztelméleti módszerekkel, másrészt a valószínűségi tesztelmélet alkalmazásával történt. A klasszikus tesztelméleti elemzéseket az SPSS program, a valószínűségi tesztelméleti elemzéseket a ConQuest program segítségével végeztük. A valószínűségi tesztelmélet lehetővé tette az összes feladat egy közös nehézségi skálán történő elhelyezését és az egyes itemek valószínűségi alapon történő, populációfüggetlen elemzését. Az elemzéseket parciális kredit modellel végeztük. Az itemek paraméterezése az egyparaméteres Rasch-modell segítségével történt (*Rasch*, 1960). A logitegységben kapott diákokra vonatkozó képességpontokat 500 pontos átlagú és 100 pontos szórású skálára transzformáltuk.

A lineáris tesztek mérési pontosságának meghatározásához a teszt megbízhatósági mutatója, a Cronbach-alpha reliabilitásmutató szolgál. A Cronbach-alpha viszont csak olyan esetben számítható, amikor a tesztelésben résztvevő minden személy minden feladatot megold, azaz nincs hiányzó adat. Adaptív teszteknél viszont a tanulók az itembanknak csak egy részhalmazát oldják meg, így a Cronbach-alpha az itembank szintjén nem számítható. Ezért az adaptív tesztek megbízhatóságának jellemzésére a valószínűségi tesztelmélettel számítható WLE személy-szeparációs reliabilitásmutatót alkalmaztuk

(*Linacre*, 1997; *Clauser* és *Linacre*, 1999). A mérési pontosság további mutatójaként a teszt információt és a standard hiba mértékét használtuk (*Weiss*, 2013), melyeket szintén a Rasch-modell segítségével számítottunk. A tesztinformációs görbék a tesztből kinyert információ nagyságát a tesztet megoldó tanulók átlagos képességszintje és az itemek nehézségi szintje közötti különbségek segítségével jellemzik. A kinyert információ nagyságát akkor tekintettük maximálisnak, ha a feladatok nehézségi szintje és az azokat megoldó diákok képesség-szintje azonos volt. Minél távolabb volt egymástól ez a két érték, annál kisebb volt a tesztelés során kinyert információ nagysága.

Mivel mindkét tesztváltozatot minden tanuló megoldotta, a becsült képességszinteket és a helyes válaszok arányát illetően személyszintű összehasonlításra is lehetőség nyílt. A változók közötti kapcsolatokat korrelációkkal vizsgáltuk. A különbözőségek szignifikanciaszintjének meghatározását egymintás t-próbával, illetve varianciaanalizissel (ANOVA) végeztük. A két teszten mért különbözőségek jellemzésére a különbség mértékét szórásegységben kifejező mutatót, a Cohen d-t használtuk (*Cohen*, 1988).

### *Eredmények összefoglalása, értelmezése*

Az induktív gondolkodás pilotmérése során mindkét tesztverzió esetén a reliabilitásmutatók értékei megfelelőek voltak, tehát mindkét tesztváltozat megbízhatóan alkalmazható a korosztály induktív képességének mérésére. Az adaptív teszten mért személyszeparációs reliabilitása magasabb volt (0,85), mint a lineáris teszté (0,83). A tanulók adaptív és lineáris tesztkörnyezetben elért eredményei magasan korreláltak egymással ($r=0,82$, $p<0,01$), a személyszintű bontás szerint a legidősebb, 8. évfolyamos tanulók esetében volt szignifikáns eltérés a kétféle teszten elért teljesítmények tekintetében. Az alacsonyabb évfolyamos diákok eredményei nem különböztek jelentős mértékben a különböző tesztkörnyezetekben. Az adaptív teszt 17 tesztváltozata közül hat útvonal fordult elő a legnagyobb hányadban, ezek között is az átlagos nehézségű modulok előfordulási aránya volt a legmagasabb. Mivel a tanulók többsége átlagos képességű volt, ez megfelel az elvártnak. A könnyű és a nehéz modulok elsősorban a gyengébb és a magasabb képességű tanulók elkülönítésében vettek részt. A tesztelés során megvalósult a gyengébb és a magasabb képességszintű tanulók elkülönítése, és a tesztelés végére a minta közel azonos arányban történő eloszlásával alakult ki a három képsségsáv.

A tesztelés során a teljes minta szintjén több információt nyertünk ki, szignifikánsan pontosabb képességszint-meghatározást végeztünk az adaptívteszt-algoritmus alkalmazásával, mint a hagyományos, lineáris módszerű teszteléskor. A kinyert információ százalékos nagyságát összehasonlítva, a lineáris teszt átlagosan 60%-os információt szolgáltatott, addig az adaptív tesztelés során kinyert átlagos információ nagysága 76% volt. A személyszintű összehasonlítás alapján 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 problémamegoldó gondolkodás mérésére lefolytatott vizsgálatban mindkét teszt reliabilitása megfelelő volt, azonban az adaptív teszt reliabilitása magasabb volt (0,83), mint a lineáris teszté (0,80), mely az adaptív tesztrendszer pontosabb képességszint meghatározását jelzi. A tanulók kétféle tesztkörnyezetben elért eredményei erősen korreláltak ($r=0,71$, $p<0,01$), a t-próba eredményei szerint nem különböztek egymástól szignifikáns mértékben ($t=-0,03$, $p=0,98$). Az évfolyamszintű összehasonlítás szerint sem volt különbség a kétféle tesztkörnyezetben elért eredmények között.

A vizsgálat kitért az adaptív és a lineáris teszteken elért helyes válaszok összehasonlítására. Az évfolyamszintű összehasonlítás szerint a 8. évfolyam kivételével mindegyik évfolyamon magasabb volt a helyes válaszok száma az adaptív tesztkörnyezetben, mint a lineáris teszt esetén. A képességszint szerinti bontást vizsgálva, a képességszint növekedésével emelkedett a helyes válaszok száma, azonban az átlag alatti tanulók az adaptív teszten több helyes választ adtak, mint a lineáris teszten. A magas képességű tanulók esetén fordítva alakult, a lineáris teszten tapasztaltnál kisebb arányban fordultak elő helyes válaszok. Az adaptív teszt hat modulja összesen négy különböző teszt kiosztására

adott lehetőséget. Az esetek kevéssel több, mint a felében a tanulók a könnyű modulokon haladtak végig, és a teszt végére is közel a minta fele a könnyű modulon végzett. A résztesztek személyszintű összehasonlítása szerint az adaptivitás előnye elsősorban a magas képességszintű tanulók elkülönítésénél mutatkozott meg.

Az adaptív teszt esetében a minta teljes szintjén több volt a kinyerhető információ mennyisége, és ezzel összhangban kisebb a standard hiba mértéke, mint a lineáris teszt esetén. Az adaptív tesztkörnyezet előnye a magasabb képességtartományban volt számottevő, itt átlagosan 20-25%-kal több volt a kinyert információ mennyisége.

A szóolvasó készség vizsgálatára lefolytatott pilotmérés eredményei alapján megállapítható, hogy a rendszer mind évfolyamonkénti, mind diákonkénti bontásban helyesen működött. Az alacsonyabb képességszintű diákok tipikusan a könnyebb, a magasabb képességszintűek a nehezebb klasztereket kapták a tesztelés során. Ennek következtében az adatfelvétel során kinyert információ mennyisége javult, mert minden egyes diák a képességszintjéhez relatív közel álló feladatokat kapott a teszt utolsó moduljában. Az utolsó két modul tekintetében 31 tanuló esetén nem változott a kapott modul szintje a harmadikról a negyedik szakaszba való lépésnél, ami a tanulók egy ötödét jelenti, tehát mindenképpen indokolt volt a négy szakasz alkalmazása.

A nagymintás mérés eredményei alapján az adaptív tesztek pontosabb képességbecslést tettek lehetővé a vizsgált évfolyamok tekintetében. A feladatok nehézségi indexei lefedték a vizsgált korosztály képességszintjét, tehát alkalmasak voltak a korosztály képességszintjének becslésére. A tesztben szereplő dimenziók szorosan korreláltak egymással, ugyanígy a kétféle tesztkörnyezetben elért eredmények is erős összefüggést mutattak, vagyis nem különbözött jelentős mértékben a tanulók kétféle tesztkörnyezetben elért eredménye. A teszteken elért helyes válaszok arányát összehasonlítva az adaptív teszt az alacsonyabb képességszintű tanulók esetén magasabb helyes válaszokat mutatott, a magas képességtartományban viszont fordítva, viszonylag kevesebb jó válasz született adaptív tesztkiosztással, mint a lineáris teszttel. Ez azt mutatja, hogy az adaptív kiosztás nagyobb sikerélményt jelentett a gyengébb képességű tanulók számra és kihívást a magas képességű tanulók részére.

A tanulók jelentős része átlagos képességű volt, ennek megfelelően a közepes nehézségű résztesztek szerepeltek a legnagyobb gyakorisággal az adaptív tesztkiosztás esetén, azonban a harmadik, illetve még a negyedik szakaszban is sok tanuló esetén módosult a szint, ami indokolja az öt különböző nehézségi szintű modul alkalmazásának szükségességét.

A mérés során összehasonlítottuk a tesztinformációkat, illetve a mérési hibák nagyságát, és mindkét esetben a teljes képességskála tekintetében az adaptív teszt esetében mértünk magasabb tesztinformációt és kisebb mérési hibát.

## HIPOTÉZISEK IGAZOLÁSA

A disszertáció célja annak vizsgálata volt, hogy a hagyományos lineáris tesztekről az adaptív tesztelésre való átállás biztosítja-e és milyen mértékben a nagyobb mérési precizitás elérését. A vizsgálat során felvetett hipotézisek és beigazolódásuk:

(1) A korábban papír alapon alkalmazott tesztekből kifejleszthető többszakaszos adaptív tesztelésre alkalmas tesztrendszer.

A kutatás során három, korábban papír alapon alkalmazott tesztrendszer online formára való konvertálása történt. A kifejlesztett tesztek mindegyik esetben megfelelő jóságmutatókkal rendelkeztek, a tanulók kétféle tesztkörnyezetben elért eredményei erős korrelációt mutattak, és a t-próba eredményei szerint csak egy-egy évfolyam tekintetében volt az elért eredmények között szignifikáns különbség. A mérések során beigazolódott első hipotézisünk, miszerint a korábban papír alapon alkalmazott rendszerek átkonvertálhatóak számítógépes adaptív verzióra.

(2) Az adaptív tesztrendszerek hatékonyan és megbízhatóan alkalmazhatóak az 1-8 évfolyamos korosztály diagnosztikus mérésére.

A tesztek hatékonyságának és megbízhatóságának egyik mutatója a reliabilitásmutató. Mivel adaptív tesztkiosztás esetén többféle tesztváltozat kerül kiközvetítésre, és a tanulók az itemeknek csak bizonyos részhalmazát oldják meg, ezért a Cronbach-α reliabilitásmutató nem volt használható.

Ehelyett az adaptív teszt esetén ennek kiterjesztését, a személyszeparációs reliabilitásmutatót használtuk. Az adaptív teszt WLE (*Weighted Likelihood Estimate*) személyszeparációs reliabilitásmutatói mindegyik teszt esetén megfelelőnek bizonyultak, így beigazolódott a második hipotézis, mely szerint a kidolgozott adaptív tesztek megbízhatóságukat tekintve alkalmasak az 1–8. évfolyamos diákok képességeinek diagnosztikus mérésére.

(3) Az adaptív rendszer a képességek pontosabb mérését teszi lehetővé.

A tesztek mérési pontosságát jól jellemzi a mérési hiba nagysága, valamint a kinyert információ mértéke (*Wang* és *Kolen*, 2001; *Wang*, 2010; *Molnár*, 2013). A tesztekből kinyerhető információ nagyságát a tesztet megoldó tanulók átlagos képességszintje és az itemek nehézségi szintje közötti különbségek segítségével jellemeztük. A kinyert információ nagyságát akkor tekintettük maximálisnak, ha a feladatok nehézségi szintje és az azokat megoldó diákok képességszintje azonos volt. Minél távolabb volt egymástól ez a két érték, annál kevesebb volt a tesztelés során kinyert információ mértéke. Mindhárom mérés esetén az adaptív tesztből kinyerhető információk mértéke szignifikánsan nagyobb volt, mint a különböző nehézségű itemekből álló lineáris teszt esetén, tehát az adaptív változatok pontosabb, precízebb mérés valósítottak meg, mint a lineáris tesztváltozatok. Hasonlóan, a mérési hibák szignifikánsan alacsonyabbak voltak az adaptív tesztkiosztások esetén, mely szintén igazolja az adaptív tesztek pontosabb működését. A mérési eredmények tehát beigazolták a harmadik hipotézisünket.

(4) A becsült képességszintek nem különböznek jelentős mértékben online adaptív és lineáris tesztelés esetén.

Mindhárom esetben megvizsgáltuk az adaptív és a lineáris tesztkiosztás esetén becsült képességszintek közötti eltérések nagyságát. A kétféle tesztkiosztás esetén az elért eredmények magas korrelációs együtthatókat mutattak, amely jelzi, hogy a kétféle tesztkiosztás hasonlóan sorolta be a tanulókat. A t próbák eredményei szerint csak egy-egy évfolyam tekintetében volt különbség a teszteken elért eredmények között, ami arra enged következtetni, hogy általában nem volt jelentős különbség a kétféle képességbecslés között, viszont bizonyos korosztály, illetve képességszintű tanulók esetében előnyösebb lehet egyik, vagy másik tesztváltozat, mely indokolja az előzetes vizsgálatok szükségességét.

(5) Adaptív tesztelésnél az alacsonyabb képességtartományban magasabb a helyes válaszok aránya, mint lineáris tesztkörnyezetben, átlag feletti tanulók esetén viszont fordítva, kisebb arányban fordulnak elő helyes válaszok.

Mindkét tesztkiosztás esetén a képességszint növekedésével párhuzamosan nőtt a helyes válaszok száma is. Az adaptív tesztkiosztás esetén azonban a növekedés mértéke eltérő volt a lineáris tesztelés során tapasztaltaktól. Az alacsony képességtartományban jelentős mértékben megnőtt a helyes válaszok aránya, aminek az a magyarázata, hogy az alacsonyabb képességszintű diákok könnyebb feladatokat kaptak, melyekkel könnyebben megbírkóztak, így nagyobb sikerélményre tehettek szert. A magas képességtartományban az induktív tesztet kivéve fordítva történt, a magasabb képességszintű tanulók nehezebb feladatokat kaptak, így kevesebb helyes válasz született, ezzel együtt nagyobb kihívást jelentett számukra a teszt megírása. Az induktív tesztelésnél mindegyik képességtartományban nőtt a helyes válaszok aránya. Az eredmények részlegesen igazolták ötödik hipotézisünket.

(6) A tanulók többsége átlagos képességszintű, ezért adaptív tesztkiosztás esetében az átlagos nehézségű itemek/résztesztek szerepelnek leggyakrabban.

Az adaptív tesztelés során kiosztásra került útvonalakat megvizsgáltuk, minden esetben a közepes nehézségű modulok szerepeltek leggyakrabban. Mivel a tanulók többsége átlagos képességszintű, ez megfelelt az elvárásainknak. A tesztbiztonság növelése érdekében ezért célszerű a közepes nehézségű modulok számát növelni. A hatodik hipotézünk beigazolódott.

(7) A tesztekből kinyert információ minden képességszinten szignifikánsan magasabb, a mérési hiba viszont szignifikánsan alacsonyabb az adaptív tesztrendszer esetében, mint lineáris teszteknél.

Mindegyik kutatás esetében elemeztük a tesztekből kinyert információ mennyiségét. Az adaptív tesztrendszer alkalmazásával kinyert információ adatfelvételtől függetlenül magasabb volt, mint a lineáris tesztekből kinyert információ nagysága. Az eltérések mértéke azonban változó volt a

különböző képességtartományokban. Az induktív gondolkodás fejlettségi szintjét célzó mérés esetén az alacsony és a magas képességtartományokban volt a kinyert információ mértéke jelentősen több, mint a lineáris teszt esetében. A problémamegoldó képességet mérő tesztek esetében a rendszer elsősorban a magas képességű egyéneket különítette el, tehát a kinyert információ mennyisége a magas képességsávban volt jelentősen több. A szóolvasás mérése kapcsán megállapítható, hogy az adaptív és a lineáris teszteken mért információk eltérése egyenletesen oszlott el a teljes képességskálán, minden képességszinten több volt az adaptív tesztből kinyert információ mennyisége, mint a lineáris teszt esetében. A mérési hiba mértéke is ennek megfelelően alakult mindegyik mérés esetében. A hetedik hipotézis részben igazolódott be, a minta nagyságától függően a különböző képességszinteken más-más mértékű lehet a kinyerhető információ mennyisége és a mérési hiba nagysága. A kutatás eredményei szerint a kisebb minták esetében elsősorban az alacsony és a magas képességtartományban jelentősen magasabb az adaptív tesztből kinyerhető információ mértéke, nagyobb mintán ez kiegyenlítődik, és a képességtartomány minden szintjén egyenletesen több információ nyerhető ki az adaptív tesztből, mint a lineáris változatból. A mérési hiba mértéke is ennek megfelelően alakul.

## Összefoglalás

A 21. században jelentkező mérési-értékelési igények egyértelműen a számítógépes tesztelés felé jelölik ki a fejlődés irányvonalát. A számítógépes tesztek számos új lehetőséget kínálnak a képességmérésre, segítségükkel lehetővé válik az azonnali kiértékelés, új, innovatív itemtípusok kerülhetnek kidolgozásra, új, eddig nem, vagy csak nehezen mérhető képességterületek pontos és hatékony mérésére adanak lehetőséget. A papíralapú tesztek számítógépes formára való konvertálása több szinten megvalósulhat. A ma létező leginnovatívabb forma a számítógépes adaptív tesztelés. Adaptív tesztelési technika esetén az itemek, illetve résztesztek egy pontosan bemért, paraméterezett itemeket tartalmazó itembankból kerülnek kiközvetítésre, és minden tanuló a képességszintjének legmegfelelőbb itemeket, illetve részteszteket kapja. Ez a tesztelési mód a hagyományos, lineáris tesztelési technikához képest a képességek sokkal pontosabb és hatékonyabb mérését teszi lehetővé. Mivel a tanulók saját képességszintjükhöz illeszkedő feladatokat kapnak, a teszt feladatai egyformán kihívást jelentenek számukra, ezáltal a teszt minden iteme egyforma mértékben járul hozzá a személy képességszintjének meghatározásához, így sokkal pontosabb képességszint meghatározásra nyílik lehetőség.

Az adaptív teszteknek számos típusa létezik, az egyik legpreferáltabb típus a többszakaszos adaptív tesztstruktúra, mely során több szakaszban itemek helyett itemcsoportok, azaz modulok kerülnek kiosztásra, melyek különböző nehézségi szintű rövid fix tesztek. A teszttípus egyesíti magában a hagyományos lineáris és az adaptív tesztek tulajdonságait, mivel egyrészt a kérdéseket a tanuló képességszintjéhez igazítja, másrészt lehetőséget ad az itemek modulon belüli sorrendjének előzetes meghatározására. A modulok előre tervezhetőek és szerkeszthetőek, így nagyobb kontrolt biztosítanak a teszt adminisztráció számára, így kiküszöbölhetővé válik, hogy az itemek egymásnak információt szolgáltassanak. 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. Az itemalapú adaptív tesztekhez képest jóval kevesebb adminisztrációt és számítógépes számítást igényelnek.

Papíralapú tesztekről adaptív tesztelésre való átállás során fontos megvizsgálandó kérdés, hogy az átállás biztosítja-e az elvárt szintű mérésmetodikai javulást, és hatékonyabb képességmérést. A vonatkozó nemzetközi vizsgálatok szerint az adaptív tesztek reliabilitása magasabb, mint a lineáris teszteké, a kinyerhető információ mennyisége szintén több, a mérési hiba viszont alacsonyabb, azaz összességében a lineáris teszteknél jóval pontosabb mérést tesznek lehetővé. A legtöbb vizsgálat eredménye szerint a többszakaszos teszt precizitása valamivel elmaradt az item, illetve a tesztlet alapú tesztekétől, viszont a tesztfejlesztés folyamán alkalmazható nagyobb adminisztratív kontroll előnyössé teszi ezt a teszttípust, ezért papír alapú tesztelésről való átállás esetén a legpreferáltabb típusú adaptív

tesztkonstrukció. A nemzetközi kutatásokon elért eredmények elsősorban szimulált adatbázisokra alapozottak, csak néhány pilotmérés történt empirikus adatok felhasználásával, egyetemista hallgatók bevonásával.

Kutatásunk célja a papír alapú tesztelésről adaptív tesztelésre való átállás feltételeinek megvizsgálása volt empirikus vizsgálat keretében az általános iskolai korosztály körében. A kutatás során több képességterület, valamint különböző szerkezetű adaptív változat esetében az adaptív tesztek hatékonyságának, mérési precizitásának összehasonlítása történt a hagyományos, lineáris tesztkörnyezetben folyó teszteléssel szemben.

Az adatfelvétel több fázisban zajlott. Különböző képességterületeket érintő három pilotmérést és két nagymintás mérést végeztünk el. A mérések az 1-8. évfolyamos tanulók körében történtek három különböző képességterület vonatkozásában. A pilotmérések alkalmával a tanulók induktív gondolkodásának, problémamegoldó képességének és szóolvasó készségének fejlettségi szintjét mértük, mely méréseket két nagymintás adatfelvétel követett, melyek során szóolvasó készség fejlettségi szintjének mérése valósult meg. A felhasznált teszteket minden esetben előzőleg papír alapon alkalmazott tesztek online formára történő konvertálásával dolgoztuk ki az eDia rendszerben. A tesztek felvétele osztálytermi környezetben történt, egy tanítási óra, azaz 45 perc alatt. A személyszintű összehasonlítás biztosítása érdekében minden tanuló kétféle tesztváltozatot oldott meg, egy hagyományos lineárisat és egy adaptívat. Mivel az adaptív tesztelés mérésmetodikai hátterét a valószínűségi tesztelmélet biztosítja, az elemzések során az egyparaméteres Rasch-modellt alkalmaztuk.

Az elemzések során a vonatkozó nemzetközi kutatásoknak megfelelően összehasonlítottuk a tesztek jóságmutatóit, a becsült képességszinteket adaptív és lineáris tesztkörnyezetben, megvizsgáltuk a kétféle tesztelés során kiosztott résztesztek nehézségi szintjének változásmintázatát és a helyes válaszok arányát, valamint összevetettük a kinyert információkat és a mérési hibák nagyságát mindkét tesztelési módra vonatkozóan.

Az eredmények igazolták hipotéziseinket, és a szakirodalmi kutatásokkal összhangban kimutatták, hogy az előzőleg papír alapon használt tesztek adaptív formára konvertálhatók, és hatékonyan alkalmazhatók az általános iskolai korosztály képességszintjének a meghatározására. Az adaptív tesztelési technika esetén az alacsonyabb képességgel rendelkező tanulók több helyes megoldást tudtak produkálni, így több sikerélményt jelentett számukra a tesztelés, a magasabb képességszintű tanulók számára pedig kihívást jelentett az adaptív tesztek megoldása. Az adaptív tesztek a képességek pontosabb mérését tették lehetővé, kedvezőbb jóságmutatókkal rendelkeztek, alkalmazásukkal a teljes minta szintjén több információt nyertünk ki, szignifikánsan pontosabb képességszint-meghatározást végeztünk, mint hagyományos, lineáris tesztelés esetén. A minta nagyságától függően azonban a különböző képességszinteken más-más mértékű volt a kinyerhető információ mennyisége és a mérési hiba nagysága. A kisebb minták esetében elsősorban az alacsony és a magas képességtartományban volt jelentősen magasabb az adaptív tesztből kinyerhető információ mértéke, nagyobb mintán ez kiegyenlítődött, és a képességtartomány minden szintjén egyenletesen több információt nyertünk ki az adaptív tesztből, mint a lineáris változatból. A képességszintek becslése során elkövetett hiba nagysága is ezzel összecsengett, és a becsült hiba nagysága a minta méretének függvényében változott.

A kutatás egyedisége, hogy az adaptív tesztelés hatékonyságát vizsgáló legtöbb kutatással szemben nem szimulált adatbázison, hanem empirikus adatok segítségével hasonlította össze a lineáris és az adaptív tesztkörnyezetben becsült képességszintek alakulását, továbbá az azonos minta alkalmazása lehetővé tette a diákszintű összehasonlítást is. A mérésekben a 6-14 éves korosztály vett részt, mely szintén egyedinek mondható az adaptív tesztekkel foglalkozó kutatások között. Az eredmények alátámasztották a szimulációs kísérletekben is tapasztaltakat, miszerint jelentős mértékű mérési precizitás érhető el adaptívteszt-algoritmus alkalmazásával a hagyományos lineáris tesztekhez képest.

A kutatási eredményeink általánosíthatóságának korláta, hogy az adatfelvételek során három teszt esetében vizsgáltuk az adaptív tesztelésre való átállás lehetőségét, az alkalmazott itembankok mérete, az itemek típusa különböző volt. Ezek a jellemzők befolyásolhatják a képességszint becslés

soránt kinyert és kinyerhető információ mennyiségét, ezért a kinyert információ mértékének pontosabb meghatározásához további kutatások szükségesek a különböző méretű és tartalmi lefedésű itembankok felhasználásával.

A kutatás gyakorlati jelentősége, hogy a kifejlesztett tesztek osztálytermi környezetben, tanórák keretében felhasználhatóak, azonnali visszajelzést biztosítanak a tanulók és a pedagógus számára. A tanulók képességbecslése pontosabbá válik, ami elsősorban a kritériumorientált mérések esetében jelentős mértékben befolyásolhatja a tanulók elért eredményeit. A tanulók különböző feladatokat kapnak, ami által növekszik a tesztbiztonság. Mivel a feladatok előre paraméterezett itembankból kerülnek kiközvetítésre, a tanulók eredményei közös képességskálán jellemzhetőek, így anélkül, hogy minden tanuló mindegyik feladatot megoldaná, nagy valószínűséggel megmondható, hogy a többi feladaton hogyan teljesítene.

## IRODALOM


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

American Educational Research Association, American Psychological Association, és National Council on Measurement in Education. (1999): *Standards for educational and psychological testing.* American Educational Research Association, Washington, DC.

American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessment (1986): *Guidelines for computer-based tests and interpretations.* Author, Washington, DC.

Amstrong, R. D. (2002): *Routing rules for multiple-form structures.* (Computerized Testing Report 02-08). Law School Admission Council.

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.** 3. sz. 147–164.

Beller, M. (2013): Technologies in large-scale assessments: New directions, challenges, and opportunities. In: *The Role of international large-scale assessments: Perspectives from technology, economy, and educational research.* Springer, Netherlands. 25–45.

Brossman, B. G., és Guille, R. A. (2014): A Comparison of multi-stage and linear test designs for medium-size licensure and certification examinations. *Journal of Computerized Adaptive Testing*, **2**. 2. sz. 18–36.

Clauser B. és Linacre J.M. (1999): Relating Cronbach and Rasch reliabilities. *Rasch Measurement Transactions.* **13.** 2. sz. 696–697.

Cohen, J. (1988): *Statistical power analysis for the behavioral sciences*. Lawrence Earlbaum Associates, Hillsdale, NJ.

Crotts, K. M., Zenisky, A. L., Sireci, S. G. és Li, X. (2013): Estimating measurement precision in reduced-length multi-stage adaptive testing, *Journal of Computerized Adaptive Testing.* **1.** 4. sz. 67–87.

Csapó, B., Ainley, J., Bennett, R. E., Latour, T. és Law, N. (2012): Technological issues for computer-based assessment. In: Griffin, P., McGaw, B. és Care, E. (szerk.): *Assessment and teaching of 21st century skills*. Springer, New York. 143–230.

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,* **18**. 3–4. sz. 3–16.

Davis, S. L. (2005): *Exploring a new methodology for setting performance level standards with computerized adaptive tests*. 35th Annual National Conference on Large-Scale Assessment. San Antonio, Texas, TX.

Eggen, T. J. H. M. (2004): *Contributions to the theory and practice of computerized adaptive testing*. Citogroep Arnhem, Netherlands.

Eggen, T. J. H. M. (2007): Choices in CAT models in the context of educational testing. In: Weiss, D. J. (szerk.): *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* http://publicdocs.iacat.org/cat2010/cat07eggen.pdf. Letöltés dátuma: 2013.12.12.

Eggen, T. J. H. M. és Straemans, G. J. J. M. (2000): Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement,* **60.** 5. sz. 713–734.

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. 103–133.

Greiff, S., Wüstenberg, S. és Funke, J. (2012). Complex problem solving. More than reasoning? *Intelligence*, **40.** 1–14.

Guille, R. A., Becker, K. A., Zhu, R. X., Zhang, Y., Song, H., és Sun, L. (2011): *Comparison of asymmetric early termination MST with linear testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Hambleton, R. K. és Xing, D. (2006): Optimal and nonoptimal computer-based test designs for making pass–fail decisions. *Applied Measurement in Education,* **19.** 3. sz. 221–239.

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, TX.

Kingsbury, G. G. és Hauser, C. (2004): *Computerizated adaptive testing and the No Child Left Behind.* Presented at the Annual Meeting of the American Educational Research Association. San Diego, CA.

Kolen, M. J. (1999-2000): Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment*, **6**. 73–96.

Linacre, J. M. (1997): Kr-20/Cronbach alpha or Rasch reliability: Which tells the truth? *Rasch Measurement Transactions*, **11**. 3. sz. 580–581.

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

Molnár Gyöngyvér (2010): Technológia-alapú mérés-értékelés hazai és nemzetközi implementációi. *Iskolakultúra*, **20.** 7–8. sz. 22–34.

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

Olea, J., Revuelta, J., Ximénez, M.C. és Abad, F. J. (2000): Psychometric and psychological effects of review on computerized fixed and adaptive tests. *Psicológica*, **21**. 1. sz. 157–173.

Paek, P. (2005): *Recent trends in comparability studies*. PEM Research Report 05–05.

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

Pyper, A. és Lilley, M. (2010): *A comparison between the flexilevel and conventional approaches to objective testing.* CAA Konferencia. University of Hertfordshire.

Rasch, G. (1960): *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research, Copenhagen.

Rotou, O., Patsula, L., Manfred, S. és Rizavi, S. (2003): *Comparison of multi-stage tests with computerized adaptive and paper and pencil tests.* Paper presented at the annual meeting of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME). Chicago, IL.

Scheuermann, F. és Björnsson, J. (2009, szerk.): *The Transition to computer-based assessment: New approaches to skills Assessment and implications for large-scale testing.* Luxemburg: Office for Official Publications of the European Communities, Luxembourg.

Scheuermann, F. és Pereira, G. A. (2008, szerk.): *Towards a research agenda on computer-based assessment.* Office for Official Publications of the European Communities, Luxembourg.

Thompson, N. A. és Prometric, T. (2007): A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research and Evaluation*, **12**. 1. sz. 1–13.

Thompson, T. és Way, D. (2007): *Investigating CAT designs to achieve comparability with a paper test.* Presented at the Applications and Issues Paper Session. Pearson.

van der Linden, W. J. (2008): Some new developments in adaptive testing technology. *Zeitschrift für Psychologie*, **216**. 1. sz. 3–11.

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). Erlbaum, Hillsdale, NJ.

Wainer, H. és Kiely, G. (1987): Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, **24**. 3. sz. 185–201.

Wan, L. Keng, L., McClarty, K. és Davis, L. (2009): Methods of comparability studies for computerized and paper-based tests. *Test, measurements and research services bulletin,* **12**. 10. sz. 1–4.

Wang, H. (2010): Comparability of computerized adaptive and paper-pencil tests. *Test, measurements and research services bulletin*, **13**. 1. sz. 1–7.

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, S., Jiao, H., Young, M. J., Brooks, T. és Olson, J. (2008): Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments. *Educational and Psychological Measurement,* **68**. 1. sz. 5–24.

Way, W. D., Davis, L. L. és Fitzpatrick, S. (2006): *Practical questions in introducing computerized adaptive testing for K-12 assessments*. Pearson, San Antonio.

Weiss, D. J. (2011): Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, **2**. 1. sz. 1–27.

Weiss, D. J. (2013): Item banking, test development, and test delivery. In Geisinger, K. F. (szerk.): *The APA handbook on testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology.* American Psychological Association, Washington DC. 185–200.

Yan, D., von Davier, A. A. és Lewis, C. (2014): *Computerized multistage testing: Theory and applications.* CRC Press, New York.

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.

Zheng, Y. (2012): *Multistage adaptive testing for a large-scale classification test: Design, heuristic assembly, and comparison with other testing modes*, ACT research report series.

## A DISSZERTÁCIÓ TÉMAKÖRÉHEZ KAPCSOLÓDÓ PUBLIKÁCIÓK

Gyöngyvér Molnár és Andrea Magyar (2015): *Comparing the measurement effectiveness of computerised adaptive testing and fixed item testing*. 16. EARLI Konferencia. Limassol, Cyprus, augusztus 25-29. Elfogadott absztakt.

Molnár Gyöngyvér és Magyar Andrea (2015): A számítógép alapú tesztelés elfogadottsága pedagógusok és diákok körében. *Magyar Pedagógia*, **115**. 1. sz. 49–66.

Magyar Andrea és Molnár Gyöngyvér (2014): A szóolvasási készség adaptív mérését lehetővé tevő online tesztrendszer kidolgozása. *Magyar Pedagógia*, **114**. 4. sz. 259-279.

Magyar Andrea, Molnár Gyöngyvér, Pásztor Attila, Pásztor-Kovács Anita, Pluhár Zsuzsa (2015): *21. században elvárt képességek számítógép alapú mérése.* XIV. Országos Neveléstudományi Konferencia. Debrecen, november 6-8.

Magyar Andrea és Molnár Gyöngyvér (2014): *A szóolvasási készség személyre szabott diagnosztikus mérését megvalósító online tesztrendszer kidolgozása*. XIV. Országos Neveléstudományi Konferencia, Debrecen, november 6-8.

Magyar Andrea és Szili Katalin (2014): *Application of Computer-Based Test to the Assessment of Reading Skills among Young Children.* VII. EARLI SIG 1: Assessment and Evaluation Conference. Madrid, augusztus 27-29.

Magyar Andrea (2014): Adaptív tesztek készítésének folyamata. *Iskolakultúra*, **24**. 4. szám. 26-33.

Andrea Magyar (2014): *Problem Solving Competence Assessment with Computerized Adaptive Testing and Fixed Item Testing among Young Children*. XVIII. JURE Konferencia. Nicosia, Ciprus, június 30-július 4.

Magyar Andrea és Szili Katalin (2014): *Computer-based assessment of word reading skills*. XII. Pedagógiai Értékelési Konferencia. Szeged, május 1-3.

Magyar Andrea (2014): *Szóolvasási készséget mérő adaptív tesztelésre alkalmas feladatbank fejlesztése.* VI. Oktatás-Informatika Konferencia. Budapest, február 7-8.

Magyar Andrea (2014): *Szóolvasási készséget mérő adaptív tesztelésre alkalmas feladatbank fejlesztése.* VI. Oktatás-informatika Konferencia tanulmánykötete. 404-412. http://www.eltereader.hu/media/2014/03/VI_OKTINF_Tanulmanykotet_READER.pdf

Pásztor-Kovács Anita, Magyar Andrea, Hülber László, Pásztor Attila és Tongori Ágota (2013) Áttérés online tesztelésre – a mérés-értékelés új dimenziói. *Iskolakultúra*, **23.** 11. sz. 86-100.

Magyar Andrea (2014): *A problémamegoldó gondolkodás vizsgálata adaptív tesztek alkalmazásával.* A VIII. Kiss Árpád Emlékkonferencia előadásainak szerkesztett változata Tartalmi összefoglalók. Debreceni Egyetem Neveléstudományi Intézete, Debrecen. 211-221.

Magyar Andrea és Molnár Gyöngyvér (2013): Adaptív és rögzített formátumú tesztek alkalmazásának összehasonlító hatékonyságvizsgálata. *Magyar Pedagógia*, **113**. 3. szám 181–193.

Szili Katalin és Magyar Andrea (2013): *A szóolvasó készség számítógép alapú mérése*. Beszédkutatás Konferencia. Budapest, november 14-15.

Magyar Andrea (2013): *Különböző típusú adaptív tesztek hatékonyságának összehasonlítása*. XIII. Országos Neveléstudományi Konferencia. Eger, November 7-9.

Magyar Andrea (2013): *Problémamegoldó gondolkodás vizsgálata adaptív tesztekkel*. XIII. Országos Neveléstudományi Konferencia. Eger, november 7-9.

Magyar Andrea (2013): *Problémamegoldó gondolkodás vizsgálata adaptív és lineáris tesztekkel.* VIII. Kiss Árpád Emlékkonferencia. Debrecen, szeptember 6-7.

Andrea Magyar (2013): *Comparative Study on Computerized Adaptive Testing and Fixed Item Testing.* 5th Szeged Workshop on Educational Evaluation. Szeged, április 15-16.

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. Letöltés ideje: 2014.06.20.

Magyar Andrea (2013): *Többszakaszos adaptív tesztek gyakorlati alkalmazása. XI. Pedagógiai Értékelési Konferencia*. Szeged, április 11-13.

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

Magyar Andrea (2012): *Számítógép alapú adaptív tesztek és fix tesztek összehasonlító vizsgálata.* XII. Országos Neveléstudományi Konferencia. Budapest, november 8-10.

Andrea Magyar (2012): *Comparative Studies on Computerized Adaptive Testing*. X. Pedagógiai Értékelési Konferencia., Szeged, április 26-28.