EÖTVÖS LORÁND TUDOMÁNYEGYETEM
PEDAGÓGIAI ÉS PSZICHOLÓGIAI KAR

Doktori disszertáció tézisei

Takácsné Kárász Judit

Adaptív teljesítménymérési algoritmusok kidolgozása az Országos kompetenciamérés adatainak felhasználásával

DOI-azonosító: 10.15476/ELTE.2024.139

Neveléstudományi Doktori Iskola
A Doktori Iskola vezetője: Dr. habil. Zsolnai Anikó, egyetemi tanár

Oktatás-tanulás-egyenlőtlenségek program
Programvezető: Dr. habil. Lénárd Sándor, egyetemi docens

Témavezetők:
Dr. habil. Nahalka István CSc, ny. egyetemi docens
Dr. habil. Széll Krisztián László, egyetemi docens

Budapest, 2024

Tartalom

1. Bevezető

A ’80-as években a közmenedzsment modell (*New Public Management*) bevezetésével a közfeladatok ellátása – így az oktatás is – a világ számos országában decentralizáltabb lett (Hood, 1991). Ezzel együtt az üzleti világ eredményességet és hatékonyságot mérő eszközei és az elszámoltathatósági rendszerek is kialakításra kerültek, így az ezredforduló után sorban indultak a nemzetközi tanulóiteljesítmény-mérések, melyek országok vagy oktatási rendszerek eredményességét hasonlítják össze különböző célok, területek és korosztályok mentén.

Az OECD PISA (OECD, 2023) 2000 óta háromévente a 15 éves korosztály szövegértési, matematikai és természettudományos műveltségét méri. A vizsgálat célja felmérni, hogy a tanulók rendelkeznek-e a munka világában elengedhetetlen önálló tanuláshoz szükséges kompetenciákkal. Az IEA PIRLS (Mullis & Martin, 2019) 2001 óta ötévente vizsgálja a 4. évfolyamosok szövegértését az olvasástanulás végén, míg az IEA

TIMSS (Mullis et al., 2021) 1995 óta négyévente a tantervhez igazodva teszteli 4. és 8. évfolyamon a tanulók matematika és természettudomány képességeit.

Ezek a nemzetközi tanulóiteljesítmény-mérések 2010-től kezdődően fokozatosan a számítógépes, majd adaptív mérés irányába fordultak. A PISA 2015-ben (OECD, 2017), a TIMSS 2019-ben (Mullis & Martin, 2017) főként számítógépes felületen került felvételre, azonban a papír-ceruza módot is lehetett választani, a trendek számítása érdekében, kizárólag híd itemekkel. A PISA 2018-ban többszakaszos adaptív teszteléssel mérte az akkor kiemelt szövegértés területet (OECD, 2019a), a PIRLS 2021-ben egyszerre vezette be a TIMSS-nél kifejlesztett számítógépes mérést és a különböző célú és nehézségű tesztek (digitalPIRLS, ePIRLS, PIRLS Literacy) egy rendszerben történő kiközvetítését, melyet csoport adaptív mérésnek nevez, és amit 2023-ban a TIMSS is alkalmazott (Mullis et al., 2021).

A PISA mérés módszertana alapján fejlesztett és 2001 óta működő magyarországi Országos kompetenciamérést (OKM) az Oktatási Hivatal szervezi. Évente méri a tanulók szövegértését és matematikai eszköztudását, eredetileg a 6. és 10. évfolyamon majd később a 8. évfolyammal kiegészülve (Balázsi et al., 2014). Célja objektív teljesítménymutatók biztosítása a köznevelési intézményeknek, fenntartóknak és az oktatáspolitikai döntéshozóknak, valamint a mérési kultúra terjesztése (Csíkos & Vidákovich, 2012). A doktori kutatás kezdetekor az OKM még papír-ceruza mérés volt, 2022-ben azonban számítógépes mérésként szervezték meg (Balázsi et al., 2021). 2024-re a mért területek kiegészültek a természettudománnyal, a digitális kultúrával, és a mérési felület magába olvasztotta az addig különálló nyelvi méréseket. A mért évfolyamok kiegészültek a 4–11. évfolyamokra (Oktatási Hivatal, 2022). A számítógépes mérésre való átállás utat nyithat a további fejlesztés, az adaptív mérésmódra történő váltás előtt.

Az adaptív mérés (Magyar, 2012) során a tanuló által kitöltött feladatok teszt közben pontozásra kerülnek, és a modern tesztelméleti modellek (IRT) segítségével becsült képességpont alapján, a nehézség szerint legjobban illeszkedő következő tesztrész kerül kiosztásra. A többszakaszos adaptív tesztek (MST) esetében itemcsoportok, a számítógépes adaptív tesztelés (CAT) esetében minden item után megtörténik a teszt irányítása. Ezáltal egyéni tesztutak keletkeznek, a megfelelő itemekből pontosabb képességbecslés és/vagy rövidebb teszt következik. Egy ilyen teszt előkészítése minden szempontból költséges, azonban a fejlesztés első szakasza szimulációs vizsgálatokkal jól előkészíthető (Thompson & Weiss, 2011).

Kutatásom motivációja és egyben célkitűzése, hogy ilyen előzetes vizsgálatokkal előkészítse az OKM, azon belül a matematika terület számítógépes adaptív megvalósítását, felhasználva a mérés hosszú története alatt felhalmozódott nagy mennyiségű empirikus adatot. A következőkben röviden összegzem a disszertáció méréselméleti hátterét, majd a kutatási kérdések után ismertetem a papír-ceruza – számítógépes – adaptív mérések közötti átmenetekre vonatkozó eredményeimet. Az egyes eredmények egyszersmind hidat képeznek a papír-ceruza mérés adatai és a hipotetikus adaptív mérésre vonatkozó megállapítások között. Vizsgálatom nóvum a hazai szakirodalomban, hasonló kutatás korábban a többszakaszos adaptív tesztelés esetében történt (Magyar & Molnár, 2015).

2. Méréselméleti háttér

A tesztek egyik jóságmutatója, hogy a mérés mennyire pontosan becsüli a jelenséget (reliabilitás), ezen belül vizsgálhatjuk, hogy az egyes itemek mennyire működnek együtt, avagy az itemek együttese mennyire sikeres a tesztalanyok megkülönböztetésében (belső konzisztencia) (Nagybányai-Nagy, 2006). A belső konzisztencia mérésére több mutató is létezik, ezek közül a három legismertebb a klasszikus tesztelmélet alapján működő KR-20 (Kuder & Richardson, 1937), a Cronbach-alfa (Cronbach, 1947, 1951), valamint a modern tesztelmélet, azon belül a Rasch-modell esetében alkalmazható személyszeparációs mutató (Wright & Masters, 1982). Amikor tehát különböző tesztek szimulációit hasonlítják össze, vagy megfelelő pontosságú teszt fejlesztése a cél, jellemzően ezek valamelyikét használják. Bár a Cronbach-alfa a legelterjedtebb, adaptív mérés esetén nem alkalmazható, mivel a különböző tesztutak miatt bizonyos itemek közötti kapcsolat nem számítható.

A CAT (Weiss & Kingsbury, 1984) nem más, mint az adaptív tesztelés, az IRT módszerek és az interaktív számítógépes felmérésvezetés kombinációja. A számítógép egy kezdő feladat után az összes korábbi válasz alapján megbecsüli a válaszoló képességfejlettségét, majd a becsült értékhez leginkább illő következő itemet választja. A teszt ezen ciklusa addig tart, amíg valamilyen megállítási kritérium nem teljesül. A CAT-nak tehát hat szerkezeti eleme van:

1) *IRT modell,* mely szerint az itemek jellemzőit és a válaszoló képességét a válasz megoldási valószínűségét leíró egyenlet köti össze. A válaszok alapján az itemek paraméterei és a válaszoló képességfejlettsége megbecsülhető. Az egyparaméteres

Rasch-modellben (Rasch, 1960) az itemeket a nehézségük különbözteti meg egymástól, a kétparaméteres modellben az itemek a diszkrimináló képességükben (meredekség) is különböznek, a háromparaméteres modell pedig figyelembe veszi azt, hogy az alacsony képességfejlettségű kitöltők hajlamosak lehetnek véletlenszerűen válaszolni a számukra nehezebb feladatokra (DuToit, 2003). Az OKM a háromparaméteres IRT modellt alkalmazza.

2) *Itembank vagy feladatbank*. Kvalifikációs (sikeres/sikertelen) jellegű mérések esetében az itemek nehézségének elsősorban a határpont környékét, képességfejlettség mérés esetén a populációt jellemző teljes képességskálát fedniük kell, és célszerű, ha magas a diszkrimináló értékük (azaz a meredekségük). Az itembank nagysága a várható minta nagyságától függ, néhány tíz itemtől néhány száz itemig terjedhet (Magyar, 2014; Weiss & Kingsbury, 1984). Nagyobb vagy többször használt mérések esetén az itemek egy része sok tesztkitöltő előtt ismertté válik, ami veszélyeztetheti a mérés biztonságát.

3) *Kezdő vagy belépési érték*. A belépési érték (első képességbecslés) lehet a teljes mintapopuláción ugyanaz, de ha rendelkezésre áll valamilyen előzetes információ, akkor lehet személyre szabott.

4) *Itemkiválasztási eljárás*. A legáltalánosabb módszerek a maximum információ, a becsült képességponthoz legközelebbi nehézség szerinti választás, illetve a bayesi megközelítés, melyek általában nagyon hasonló eredményre jutnak. Az itemkiválasztási eljárásokkal kapcsolatos vizsgálatok jellemzően különböző módszereket hasonlítanak össze a teszt befejezéséhez szükséges itemszám és/vagy a képességbecslés pontossága szerint, tipikusan szimulációs módszerekkel (Ito & Segall, 2013).

5) *Képességbecslési eljárás*. Az aktuális válaszmintázat alapján a képességpont és esetleg a képességpont konfidencia-intervallumának becslése. A becslésre maximum likelihood és Bayes becslési módszereket használnak, esetleg ezek valamilyen kombinációját.

6) *Megállítási kritérium*. A teszt céljainak megfelelően lehet feltétel több megállítási kritérium egyikének teljesülése, például meghatározott számú item megválaszolása; a képességbecslés hibája bizonyos szint alá csökken (minden válaszadóra egyformán pontos képességbecslés); a képességfejlettség konfidencia-intervalluma alapján a teszt kitöltője besorolható valamilyen teljesítményszintre (klasszifikáció); letelt a kérdések megválaszolására szánt maximális idő.

A CAT számos előnnyel bír a lineáris tesztekhez képest. A tesztek várhatóan lényegesen rövidebbek, a képességbecslés a skála szélein pontosabb (Weiss, 2011). A

képességfejlettséghez illeszkedő nehézségű kérdés javíthat a tesztkitöltés belső motivációján, bár ez inkább a képességskála alsó részén tapasztalható, amennyiben a kitöltőket informálták a CAT működéséről (Wise, 2014). A CAT hátránya lehet, hogy visszalapozásra, a korábbi feladatok javítására nincs lehetőség, ami növelheti a stresszt, azonban Akhtar és munkatársai (2023) metaanalízisükben sem nagyobb motivációra, sem nagyobb stresszre nem találtak egyértelmű eredményt.

A nagy létszámú CAT előfeltétele a számítógépes mérés, vagyis felvetődik a papír-ceruza teszt és számítógépes változata közötti különbség, a médiahatás lehetősége (Buerger et al., 2019). A különbség megmutatkozhat a mért konstruktumok közötti különbségben, a teljesítménypontok szisztematikus eltérésében vagy a szöveges válaszok jellemzőiben. További különbséget okozhat az el nem ért és kihagyott itemek eltérő pontozása.

A CAT során az azonnali kiértékelés miatt nyílt végű, önálló szövegalkotást és képzett kódolót igénylő feladatok sem alkalmazhatók, de legalábbis a teszt közben a képességbecslésben, tesztirányításban nem vesznek részt. Amennyiben ezek az itemek a jelenség más aspektusait mérik, mint a zárt végű, automatikus kódolású itemek, az szintén a lineáris és adaptív tesztek közötti különbséghez vezet.

3. Kutatási célok, kérdések

Kutatásomban az alábbi fő- és alkérdésekre keresem a választ.

1) Az OKM papír-ceruza méréseiből származó adatok relevánsan felhasználhatók-e a számítógépes adaptív mérés tervezésére?

   a. Mi a számítógépes mérési környezet hatása a mérés eredményére? Kell-e médiahatásra számítani, és ha igen, hogyan kezelhető? (Fishbein et al., 2018)
   b. A nyílt végű itemek elhagyása mellett is azonos marad-e az OKM mérés tartalmi kerete? Kizárólag zárt végű itemeket alkalmazva milyen eltéréseket tapasztalnánk a tanulók képességpontjának becslésében?

2) Az eredeti méréssel megegyező mérési pontosság mellett mi az adaptív teszteléssel elérhető legrövidebb teszthossz? (Weiss, 2011)

3) Az OKM papír-ceruza méréseiből származó adatok alapján mely adaptív mérési elemek valószínűsítik a mérés céljának sikeresebb megvalósítását (a matematikai eszköztudás területen)? (Thompson & Weiss, 2011)

a. A papír-ceruza teszttel azonos itemszám mellett az adaptív teszt esetében csökken-e a tanulói képességfejlettség-becslés hibája?
b. Az OKM adatain alapuló, számítógépes adaptív tesztet imitáló szimulációk alátámasztják-e, hogy lényegesen rövidebb idő alatt (kevesebb itemmel) a papír-ceruza teszt pontosságának megfelelő pontossággal meghatározható a diákok képességpontja/teljesítménye?
c. Milyen megállítási kritériumok milyen mérési céloknak felelnek meg az adaptív OKM kapcsán?
d. A megállítási kritériumok között van-e hierarchia, azaz léteznek-e olyan erős kritériumok, melyek teljesülése magával hozza a gyengébb feltételek teljesülését?
e. Az első 5–10–15–20 kérdés után változik-e még a diák teljesítménye? 5–10 kérdéses teszthossz mellett milyen teljesítménybecslések várhatók?

4. A kutatás módszertana

Kutatásom egy már létező teljesítménymérési rendszer, az Országos kompetenciamérés következő fejlődési lépcsőjének megalapozó vizsgálata, ennek értelmében alkalmazott kutatás, módszertanát tekintve elsősorban kvantitatív módszertanú. A médiahatás vizsgálatát (1a.) szisztematikus szakirodalmi áttekintéssel végeztem, mely alapvetően kvalitatív fókuszú vizsgálat. A nyílt itemek elhagyása (1b.) kvantitatív, empirikus vizsgálat, melyben IRT szerinti képességbecslést, korrelációs és kereszttáblás elemzéseket alkalmaztam. A számítógépes adaptív mérési technológia elméleti vizsgálata (2) a tanulói képességfejlettség becslésének mérési hibájával kapcsolatban kvalitatív matematikai levezetés. A különböző item kiválasztási és képességbecslési eljárások összehasonlítása (3a.-e.) kvantitatív és empirikus jellegű vizsgálatok, melyeket hibrid és Monte Carlo szimulációkkal vizsgáltam.

Az 1a.-b. kérdések vizsgálata matematika, szövegértés és természettudomány területeken történt. A hasonló tesztszerkezet a matematika terület eredményeinek a trianguláció elve szerint nagyobb érvényességet ad. A 2. kérdés vizsgálata területfüggetlen, a 3a.-e. kérdéseknél kifejezetten a matematika területre koncentráltam.

Adaptív mérések esetében a *szimulációs módszerek* olyan technikák, amelyek a nagy számítógépes kapacitáson és az IRT modellek formális matematikai egyenletein alapulnak. Lényege, hogy az előre meghatározott tanulói elméleti képességfejlettség és az ismert itemparaméterek segítségével az alkalmazott IRT modell alapján a számítógép

a helyes válasz valószínűségét kiszámítja, és egy 0 és 1 közötti véletlen számmal összehasonlítva szimulálja a helyes vagy helytelen választ. A képességpont becslése a szimulált válaszok alapján történik, az itemkiválasztási eljárás ehhez a becsléshez igazodva választja a következő elemet. A szimuláció előnye a valódi adatfelvétellel szemben, hogy nagyobb mintaelemszámra és számos különböző kondíció összehasonlítására ad lehetőséget (pl. Şahin & Weiss, 2015). A szimulációk csoportosíthatók aszerint, hogy mekkora mértékben használnak valós adatokat. A *Monte Carlo szimulációk* (Kehl, 2012) teljes egészében véletlenszám generátorral készült adatokat használnak. *Hibrid szimulációk* esetén a válaszok egy része valódi, más része pedig az IRT modell alapján generált. *Post-hoc szimulációról* akkor beszélhetünk, amikor minden item és minden kitöltő esetében rendelkezésre áll az itemre adott valódi válasz (Sari, 2020), mivel azokat lineáris teszt formájában kitöltötték.

A CAT eljárás szimulációját az R (R Core Team, 2016) környezetben működő, nyílt forráskódú programcsomag, a catR (Magis et al., 2017b) segítségével végeztem, ami ingyenes felhasználású és programozható, azaz adaptálható a kutatáshoz.

5. Eredmények

5.1. Papír-ceruzáról számítógépes adatfelvételre – médiahatás vizsgálat

Az OKM esetében nincs médiahatás vizsgálatról elérhető publikáció, ezért ezt a kérdést a PISA, PIRLS és TIMSS mérések digitalizációjával, azon belül a médiahatással kapcsolatos hazai és nemzetközi tudományos publikációk és mérési dokumentumok szisztematikus szakirodalmi áttekintésével (Rother, 2007) vizsgáltam. A kutatás során a szisztematikus áttekintések és metaanalízisek esetében ajánlott PRISMA (Page et al., 2021) irányelveket követtem, azaz az adatbázisokban folytatott keresés célja a lehető legtöbb és legrelevánsabb forrás felfedése és szintetizálása előre jól meghatározott keresési és kizárási kritériumok alapján. A keresést 2021. december 2-án hajtottam végre.

A hazai keresést a MATARKA, MTMT és Arcanum Digitális Tudománytára adatbázisokon és az Oktatási Hivatal honlapján, a nemzetközi keresést az EBSCO, ERIC, JSTOR, ProQuest, Science Direct és Web of Science adatbázisokon, valamint az OECD és az IEA honlapján hajtottam végre. Befogadásra került minden 2010 után megjelent tudományos lektorált empirikus kvantitatív cikk, könyvfejezet vagy tanulmány, amely angol vagy magyar nyelven jelent meg, a mérések eredeti adatait vagy azokhoz szorosan

kapcsolódó saját vizsgálat adatait dolgozta fel, és kifejezetten a médiahatás vizsgálatára irányult.

A hazai keresés 375 itemből 2 releváns forrást talált. Mindkettő valamely mérés összefoglalója, melyhez további 4 mérési dokumentumot találtam. A dokumentumok röviden tájékoztatnak a számítógépes mérés bevezetéséről, illetve megemlítik a teljesítménypontok ebből következő lehetséges eltérését. A nemzetközi keresés 1262 tételt azonosított. Az áttekintés 20%-a és a teljes szöveg válogatása másodkódolással történt, a Cohen-kappa alapján legalább jelentős egyezés volt a kódolók között. A kritériumoknak végül 8 tétel felelt meg (1. melléklet), ehhez csatlakozott 24 mérési dokumentum. A mérési dokumentumok közül a PISA 2015 és a TIMSS 2019 dokumentumai relevánsak, a PISA 2018 nem ad többlet információt a médiahatásról, a PIRLS esetében pedig nem volt a kijelölt időszakban számítógépes mérés vagy ennek előkészítésére vonatkozó információ.

A PISA médiahatás vizsgálatát a 2015-ös mérés próbamérése során végezték (OECD, 2016). A papír-ceruza és számítógépes mérés konstruktumát egyezőnek találták, a trend itemek kb. 90%-a legfeljebb nehézségében különbözött. Az itemek szintjén mindkét irányú médiahatást (könnyebb vagy nehezebb, mint a papír-ceruza változat) találtak, a korrekciót ezért item-szinten alkalmazták. A kapcsolódó 6 cikk magas mérési minőségű az eljárás és az alkalmazott statisztikák szempontjából, azonban jellemzően Németországhoz, illetve a 2012-es és 2015-ös próbamérésekhez kapcsolódnak. A vizsgálatok hasonló eredményre jutnak, a teljesítménypontokban 10–20 teljesítménypontnyi nem szisztematikus médiahatást találnak. A szöveges válaszok elemzése valamivel hosszabb és nagyobb információtartalmú válaszokat jeleznek a számítógépes tesztek esetében.

A TIMSS 2019 esetében az egyik cikk maga a médiahatásvizsgálat (Fishbein et al., 2018). A médiahatás az itemek esetében elhanyagolható, a teljesítménypontok esetében 7–14 pontnyi szisztematikus különbséget talált, a számítógépes mérést nehezebbnek mutatva. Ennek alapján a mérések konstruktumai megegyeznek, azonban a trendek esetében évfolyamonként és mérési területenként eltérő korrekciót alkalmaztak. Országonként sem találtak szisztematikus médiahatást, ugyanakkor a másik találat, a holland alminta elemzése szerint a közös korrekció valamelyest felülbecsüli az eredményt (Robitzsch et al., 2020).

5.2. Lineáristól az adaptív mérés felé – a nyílt itemek szerepe

Az adaptív mérés előfeltétele, hogy a tanulók által adott válaszokat a rendszer azonnal pontozza, a képességfejlettséget ez alapján becsülje. A papír-ceruza OKM tesztek megközelítőleg harmada nyílt végű, képzett kódoló munkáját igénylő feladat. A nyílt itemek nem egyes gondolkodási műveletek vagy tartalmi területek pontosabb mérése miatt szerepelnek, hanem a mérés egészének változatosságához járulnak hozzá (Balázsi et al., 2014). A számítógépes mérésekből fokozatosan kivezették a nyílt végű itemeket, azonban megmaradtak az automatikusan kódolható, rövid szöveges választ (egy szó vagy szám) igénylő feladatok, azonban ezeket az itemtípusokat jelen vizsgálatban nyílt itemnek tekintettem.

A vizsgálat az OKM 2017. évi tanulói adatain történt. A nem értékelhető füzetek és mentesülő tanulók adatainak eltávolítása után a teljes tesztre számított képességbecslés mellé csak a zárt itemek alapján is képességpontot számítottam, majd ezeket az OKM szerinti 7 + 1 képességszintre osztottam be. Az eredmények az OKM módszertanának megfelelően súlyozással (N_6 = 86151, N_8 = 80886, N_{10} = 76550) készültek.

A teljes és a csak zárt itemekből álló tesztek korrelációs elemzése alapján a szövegértés és a matematika képességpontok egymással 0,664 – 0,777 közötti szinten korrelálnak, függetlenül attól, hogy az adott területet teljes teszttel vagy csak zárt itemekkel mértük, ami közepes vagy erős kapcsolatot mutat (Vargha, 2015) a két terület között, függetlenül a nyílt itemek használatától. A zárt kérdésekből számított képességpontok és a teljes tesztből számított képességpontok mindkét területen 0,9 feletti, de 1-nél kisebb korrelációt mutatnak, ami nagyon erősnek számít. Ez alapján nincs eltérés a teljes teszt és a csak zárt itemekből álló teszt által mért konstruktumok között.

A képességszintek összehasonlításával ellenőriztük a képességbecslések közötti eltérést. Mindkét területen a zárt itemek alkalmazásának centráló hatása jelenik meg, azaz a legalsó szinteken (1. alatti és 1. szint) álló tanulók könnyebben kapnak egy szinttel magasabb, míg a magasabb képességszinteken állók egy szinttel alacsonyabb besorolást. Az 1. képességszint alatti tanulók 40%-a, az 1. képességszinten levők harmada került besorolásra eggyel magasabb képességszintre. Matematikából a magasabb szintek torzítása már az 5. és 6. szinteken észrevehető, míg szövegértés területen a 6. és 7. szint érintett. Két szintnyi tévedés mindössze a tanulók kevesebb, mint 1%-a esetében volt.

5.3. Adaptív mérés tervezése – elméleti optimum

Az adaptív tesztek mintanagyságára és teszthosszára vonatkozó eredmény azon az ötleten alapul, hogy adott mérés céljai, jóságmutatóinak elérendő szintje előre meghatározható, ezek alapján a mérés szervező által meghatározott elemei tervezhetők. Az alábbiakban a tervezés kiindulási pontja a teszt reliabilitás mutatója, azaz minden tesztkitöltésnél bizonyos mértékű mérési hiba elérése a cél, akár az itemekről (S_1), akár a tanulókról (S_2) legyen szó. Adaptív mérések esetében tervezhető a teszt nehézsége, azaz a soron következő feladat megoldási valószínűsége (p), ami jellemzően 50%, azonban ettől eltérő értékek is lehetségesek. Tegyük fel, hogy az itemek és tanulók kategorizálása K képességszint használatával történik, ez legyen a mérés finomsága. A reliabilitás mérésére a KR-20 és Wright formulái szolgálnak, így a teszt dichotóm itemek és a Rasch-modell használatára épül.

A levezetés eredménye alapján (1) az item elvárt pontossága (S_1), a mérés finomsága (K) és a teszt nehézsége (p) alapján számítható az item próbaméréséhez szükséges mintanagyság (N), valamint (2) a képességbecslés elvárt pontossága (S_2), a teszt nehézsége (p), a minta nagysága (N) és a mérés finomsága (K) alapján számítható a teszt várható hossza (L).

(1) $$S_1 = \sqrt{\frac{N}{s_i(N-s_i)}} = \frac{1}{\sqrt{N}}\left[\frac{K}{\sqrt{p(K-p)}}\right],$$

(2) $$S_2 = \sqrt{\frac{1}{Lpq}}\sqrt{1 + \frac{\left(\frac{L}{L-1}\right)\left(\frac{N-L}{N}\right)^2 \ln^2\left(\frac{K-p}{p}\right)}{2{,}89}}.$$

Az adaptív mérésekre jellemző kiegyensúlyozott megoldottságú itemek (p = ½) esetén az egyenletek tovább pontosíthatók:

(3) $$N = \left[\frac{1}{S_1^2}\frac{4K^2}{2K-1}\right],$$

(4) $$D = \frac{2{,}89}{(2K-1)},$$

$$A = \frac{2{,}89S_2^2}{4(2K-1)} = \frac{S_2^2}{4}D,$$

$$0 = L^3 - L^2(AN^2 + 2N) + L(N^2 + DN^2 + AN^2) - DN^2.$$

A próbamérés szükséges nagyságát és a tesztek várható hosszát növeli a nagyobb pontosság és a képességszintek nagyobb száma, valamint a közepestől eltérő nehézség. Például az OKM esetében, ahol a mérés finomsága 7 + 1 képességszint, kiegyensúlyozott tesztet feltételezve, egy item nehézségének 0,2 hibával terhelt beméréséhez 400 fős

próbamérést kell végezni, feltéve, hogy valamilyen képességbecslés alapján a képességskála egyenletes mintáján osztjuk ki, azaz nem a középső szintekről kerül ki a legtöbb kitöltő. Ugyanezen mérés esetén, a képességbecslés szóráshoz mérten 0,5-ös (100 pontos) hibájával, 100 000 fős minta mellett várhatóan minden teszt 58 item alatt véget ér. Kevesebb, 5 képességszinttel elegendő lehet tesztenként 44 feladat.

5.4. Lehetséges adaptív stratégiák összehasonlítása pontosság és megbízhatóság alapján

Kutatásom utolsó fázisában szimulációs vizsgálatokat végeztem, hogy összehasonlítsam néhány képességbecslési és itemkiválasztási módszer hatékonyságát. Hibrid szimulációt futtattam abban az értelemben, hogy a 2008–2019. évi mérések jól működő dichotóm itemei (625 item), és azok háromparaméteres modell szerint számított paraméterei alkották az itembankot. A szimulációk így előzetes információt szolgáltatnak arról is, hogy az évek során felhalmozott itemek digitalizált változatai megfelelők lehetnek-e egy adaptív méréshez.

A tanulók elméleti képességfejlettségét a képességskála finom felosztása adta. 800 és 2200 pont között 50 pontos lépésközzel minden osztásközön kétszáz mérést szimuláltam. A szimulált tesztek belépési értéke a 6. évfolyamos országos átlag, 1500 pont volt. A képességbecslés Bayes-változatai esetében 1500 pont átlagú és 200 pont szórású prior eloszlást határoztam meg. A vizsgált képességbecslési eljárások a maximum likelihood (ML) (Lord, 1980), a Bayes-modal (BM) (Birnbaum, 1969) és az expected a-posteriori (EAP) (Bock & Mislevy, 1982) eljárások voltak. Az itemkiválasztási eljárások közül a Maximum Fisher információ (MFI) (Birnbaum, 1968), a legközelebbi nehézség (bOpt) (Urry, 1970) és a legközelebbi maximális információ (thOpt) (Barrada et al., 2006) kritériumokat hasonlítottam össze. Két lehetséges mérési célt és ehhez illeszkedő megállítási kritériumot alkalmaztam. Az első esetben 50 item hosszú tesztek mellett a becslés hibájának nagyságát vizsgáltam. A második esetben egységes mérési minőség, 60 pontnyi mérési hiba elérése volt a cél, ekkor a tesztek hosszát lehetett összehasonlítani.

Az eredmények alapján a képességfejlettség és a képességbecslés átlagos különbsége szerint a maximum likelihood becslések hozták a legjobb eredményt. A rögzített teszthossz esetén képességskála szélein akár 100 ponttal kisebb eltérést mutattak, mint a Bayes-becslések, a rögzített hiba esetén mindössze 50 pont volt a különbség.

Az itemkiválasztási módszerek közül a maximum Fisher információ szerinti kiválasztás valamivel jobb eredményt hozott, mint a legközelebbi nehézség és a legközelebbi maximális információ kiválasztási módszerek. Ezzel a módszerrel az 1200 és 1900 pont közötti tartományban a jelenlegi lineáris tesztnél rögzített teszthossz mellett pontosabb, rögzített hiba mellett rövidebb tesztek születtek.

A jelenlegi itemekből álló itembank további vizsgálatára Monte Carlo szimulációt alkalmaztam, ahol hasonló, de egyenletes eloszlású nehézséggel és az OKM-nél elvártnál átlagosan 0,5 logittal nagyobb meredekséggel rendelkező, 300 elemű itembankot szimuláltam. A mintanagyság 90 000 fő volt, a kombinált megállítási kritérium az előző vizsgálathoz illeszkedően az 50 item vagy a 60 pontnyi hiba elérése volt. A szimuláció során MFI kiválasztási módszert és BM és EAP becslést alkalmaztam kétparaméteres IRT modell mellett. Tesztbiztonsági szempontként 20%-os kitettségi korlátot állítottam be.

Az eredmények alapján a tesztek átlagosan 19–20 item után fejeződtek be, alig volt 30 itemnél hosszabb teszt. Az elméleti és a becsült képességpont közötti korreláció igen magas (r = 0,95). Az elméleti és a becsült érték közötti eltérés (a mérési hiba) szórása, vagyis a standard hiba 60 pont körüli (*RMSE* = 60,44), az eltérések átlaga -0,39 pont, szisztematikus eltérés tehát nincs. Az itemek kitettségét illetően, 44 item a lehető legnagyobb számban lett kiosztva, ezek jellemzően a meredekebb itemek közül kerültek ki, ugyanakkor 65 itemet egyáltalán nem használt föl a szimuláció, ezek jellemzően kevésbé jól diszkrimináló itemek voltak.

6. A kutatás összegzése, korlátai és kitekintés

Kutatásomban az Országos kompetenciamérés fejlesztésének egy lehetséges irányát, a digitalizációra épülő számítógépes adaptív tesztelés módszertani kérdéseinek vizsgálatát tűztem ki célul. Olyan megelőző vizsgálatok elvégzését, melyek előkészítik az egyik hazai tanulóiteljesítmény-mérési rendszer esetében a szakmai és mérésmódszertani szempontból sikeres papír-ceruza – számítógépes adaptív adatfelvétel átmenetet.

A papír-ceruza teszt és a számítógépes mérés közötti átmenet a kutatás ideje alatt, 2022-ben megvalósult, azonban a nemzetközi nagymintás tanulóiteljesítmény-mérések médiahatással kapcsolatos eredményeink összegzése továbbra is hiánypótlónak számít. A PISA és TIMSS mérések médiahatással kapcsolatos eredményei némiképpen különböznek: míg a TIMSS esetében minimális, területenként és évfolyamonként különböző nagyságú, de a számítógépes mérést szisztematikusan nehezebbnek mutató

médiahatást találtak, addig a PISA esetében az itemek kis részénél találtak kismértékű, különböző irányú eltérést. A szisztematikus szakirodalmi áttekintés eredménye alapján az OKM esetében sem kell jelentős médiahatásra számítani, azonban kisebb eltérés a nehézség paraméterekben vagy a képességpontokban lehetséges, akár területenként, akár egyes itemek esetében. Ennek feltárására javasolt empirikus vizsgálat lefolytatása.

A teljes teszt és csak zárt végű itemekből számított képességpontok összehasonlítása alapján a tesztek ugyanazt a jelenséget mérik. A nyílt itemek szerepe a legalacsonyabb és a legmagasabb képességfejlettségű diákok megkülönböztetésében, vagyis a leszakadók és a tehetségek azonosításában játszhat szerepet, ez azonban, többek között az egyéni hibák nagysága miatt, nem célja az OKM-nek. Az alsóbb képességszintek feljebb értékelése hátterében meghúzódhat a tippelés, ami a zárt itemeket érinti, illetve a válaszadási hajlandóság, ami a nyílt itemek esetében alacsonyabb lehet. A magasabb képességszintű tanulók lejjebb értékelése hátterében a nyílt itemek nagyobb nehézsége, esetleg az ilyen típusú feladatok összetettsége állhat. Mivel a papír-ceruza teszt esetén nyíltnak számító rövid szöveges választ igénylő item automatikus kódolású itemnek számít számítógépes környezetben, ezért lehetségesnek tartom, hogy a vártnál még kisebb eltérésre lehet számítani. Ugyanakkor indokolt a nyílt végű itemek további vizsgálata, hasonló nehézségű, tartalomban megfelelő automatikus kódolású itemek fejlesztése.

Az elméleti levezetés hidat teremt a matematikai egyenletek és a gyakorlat, mind a szimulációk, mind a valódi tesztek tervezése felé. Mivel jellemzően ideális körülményeket feltételez, nem számol például a teszt elején a képességbecslés nagy bizonytalanságával, ezért eredménye leginkább elméleti alsó korlátnak tekinthető. Ugyanakkor a Rasch-modell alkalmazása miatt a többparaméteres modellek jobb eredményeket is hozhatnak, azonban ezek elméleti vizsgálata jelenleg túlságosan bonyolult. Amennyiben a megállítási kritériumok meghatározása vagy itemkiválasztási lépés során a képességszinteket alkalmaznák, javasolt lehet a képességszintek számának drasztikus csökkentése a teszt rövidítésének érdekében.

A médiahatással és a nyílt itemek vizsgálatával kapcsolatos eredmények alapján a papír-ceruza lineáris tesztek során összegyűlt item és tanulói szintű adatok feltételezhetően jól használhatók a számítógépes vagy adaptív tesztek tervezésére. A szimulációs eredmények alapján a maximum likelihood képességbecslés és a maximális Fischer-információ szerinti kiválasztás eredményezheti a legpontosabb és leggyorsabb teszteket. Feltételezhető azonban, hogy a papír-ceruza lineáris tesztekhez fejlesztett,

inkább a képességskála középső részére koncentráló és a tanulók szélesebb spektrumát mérő itemek mellett szükséges a képességskála széleire fókuszáló és/vagy meredekebb itemek fejlesztése. A szimulációk érvényességét tovább javítaná a valódi tanulói válaszok felhasználása vagy a kezdőérték más információkat is integráló személyre szabottabb megválasztása, ezek tehát lehetséges továbblépési irányok. Szintén következő lépés az itembank egyéb sajátosságainak feltérképezése, mely történhet az Oktatási Hivatalon kívül, valamint az itemfejlesztés, melyet a tesztbiztonsági szempontok miatt továbbra is a Hivatalon belül javaslok elvégezni.

Az OKM elkötelezett az adaptív mérési módszer fejlesztése mellett mind szakmai (Balázsi et al., 2021), mind oktatáspolitikai (Karkó, 2023) oldalról, ezért a lineáris tesztről az adaptív mérésre történő áttérés vizsgálata aktuális és releváns. Az OKM-ből származó információk széleskörű használata miatt a téma vizsgálata társadalmi hasznossággal bír. Ugyanakkor az adaptív mérés, bár rövidebb teszteket és pontosabb képességbecslést eredményezhet, e két lehetséges cél közül várhatóan az egyiket tudja csak megvalósítani. Éppen ezért a további fejlesztések előtt szükséges lenne a mérés céljának pontosabb meghatározása. A teszt rövidítése az iskolák és tanulók terheit csökkentené, a középső képességszinteken mindenképpen, azonban az egyéni eredmények pontossága várhatóan nem lenne nagyobb. A képességbecslés egységes pontossága inkább az egyéni eredmények nagyobb érvényességét és a tanárok értékelésének lehetőségét jelenti, miközben a középső szinteken valószínűleg szintén valamivel rövidebb teszteket eredményezne.

Irodalomjegyzék

Akhtar, H., Silfiasari, Vekety, B., & Kovacs, K. (2023). The Effect of Computerized Adaptive Testing on Motivation and Anxiety: A Systematic Review and Meta-Analysis. *Assessment, 30*(5), 1379–1390. https://doi.org/10.1177/10731911221100995

Balázsi, I., Balkányi, P., Balogh, V. K., Gyapay, J., Ostorics, L., Palincsár, I., Rábainé Szabó, A., Suhajda, E., Szepesi, I., Szipőcsné Krolopp, J., & Velkey, K. (2021). *Folytonosság és változás az Országos kompetenciamérés szövegértés és matematika tartalmi kereteiben* (Köt. 1). Oktatási Hivatal. https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/digitalis_orszmer/OKMtartalmikeret_Szovegertes_Matematika.pdf

Balázsi I., Balkányi P., Ostorics L., Palincsár I., Rábainé Szabó A., Szepesi I., Szipőcsné Krolopp J., & Vadász C. (2014). *Az Országos kompetenciamérés tartalmi keretei—Szövegértés, matematika, háttérkérdőívek*. Oktatási Hivatal. https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/orszmer2014/AzOKMtartalmikeretei.pdf

Barrada, J. R., Mazuela, P., & Olea, J. (2006). Maximum information stratification method for controlling item exposure in computerized adaptive testing. *Psicothema*, *18*(1), 156–159.

Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In F. M. Lord & M. R. Novick (Szerk.), *Statistical theories of mental test scores* (o. 397–479). Addison-Wesley.

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology*, *6*(2), 258–276. https://doi.org/10.1016/0022-2496(69)90005-4

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP Estimation of Ability in a Microcomputer Environment. *Applied Psychological Measurement*, *6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Buerger, S., Kroehne, U., Koehler, C., & Goldhammer, F. (2019). What makes the difference? The impact of item properties on mode effects in reading assessments. *Studies in Educational Evaluation*, *62*, 1–9. https://doi.org/10.1016/j.stueduc.2019.04.005

Cronbach, L. J. (1947). Test "reliability": Its meaning and determination. *Psychometrika*, *12*(1), 1–16. https://doi.org/10.1007/BF02289289

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), Article 3. https://doi.org/10.1007/BF02310555

Csíkos C., & Vidákovich T. (2012). A matematikatudás alakulása az empirikus vizsgálatok tükrében. In Csapó B. (Szerk.), *Mérlegen a magyar iskola* (o. 83–130). Nemzeti Tankönyvkiadó.

DuToit, M. (Szerk.). (2003). *IRT from SSI: Bilog-MG, Multilog, Parscale, Testfact*. Scientific Software International.

Fishbein, B., Martin, M. O., Mullis, I. V. S., & Foy, P. (2018). The TIMSS 2019 Item Equivalence Study: Examining Mode Effects for Computer-Based Assessment and Implications for Measuring Trends. *Large-Scale Assessments in Education*, *6*. https://doi.org/10.1186/s40536-018-0064-z

Hood, C. (1991). A Public Management for All Seasons? *Public Administration, 69*(1), 3–19. https://doi.org/10.1111/j.1467-9299.1991.tb00779.x

Ito, K., & Segall, D. O. (2013). A Comparison of Four Methods for Obtaining Information Functions for Scores From Computerized Adaptive Tests With Normally Distributed Item Difficulties and Discriminations. *Journal of Computerized Adaptive Testing, 1*(5).

Karkó, Á. (2023). Mindig minden változik: Az emberek, a társadalom és az oktatás. *Új köznevelés, 79*(7), 3–5.

Kehl D. (2012). Monte-Carlo-módszerek a statisztikában. *Statisztikai Szemle, 90*(6), 521–543.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160. https://doi.org/10.1007/BF02288391

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Inc.

Magis, D., Yan, D., & von Davier, A. A. (2017). *Computerized Adaptive and Multistage Testing with R*. Springer International Publishing. https://doi.org/10.1007/978-3-319-69218-0

Magyar A. (2012). Számítógépes adaptív tesztelés. *Iskolakultúra, 22*(6), Article 6.

Magyar A. (2014). Adaptív tesztek készítésének folyamata. *Iskolakultúra, 14*(4), 26–35.

Magyar A., & Molnár G. (2015). A szóolvasási készség online mérésére kidolgozott adaptív és lineáris tesztrendszer összehasonlító hatékonyságvizsgálata. *Magyar Pedagógia, 115*(4), Article 4. https://doi.org/10.17670/MPed.2015.4.403

Mullis, I. V. S., & Martin, M. O. (Szerk.). (2017). *TIMSS 2019 Assessment Frameworks*. TIMSS & PIRLS.

Mullis, I. V. S., & Martin, M. O. (Szerk.). (2019). *PIRLS 2021 Assessment Frameworks*. TIMSS & PIRLS; Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/pirls2021/frameworks/

Mullis, I. V. S., Martin, M. O., & von Davier, M. (Szerk.). (2021). *TIMSS 2023 Assessment Frameworks*. TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2023/frameworks/pdf/T23_Frameworks.pdf

Nagybányai-Nagy O. (2006). A pszichológiai tesztek reliabilitása. In Rózsa S., Nagybányai-Nagy O., & Oláh A. (Szerk.), *A pszichológiai mérés alapjai: Elmélet, módszer és gyakorlati alkalmazás* (o. 103–116). Bölcsész Konzorcium.

OECD. (2016). Annex A6 The PISA 2015 field trial mode-effect study. In *PISA 2015 Results (Volume I): Excellence and Equity in Education*. OECD. https://doi.org/10.1787/9789264266490-en

OECD. (2017). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*. OECD. https://doi.org/10.1787/9789264281820-en

OECD. (2019a). *PISA 2018 Assessment and Analytical Framework*. OECD Publishing. https://doi.org/10.1787/b25efab8-en

OECD. (2019b). PISA 2018 Technical Report—Chapter 2 Test Design and Test Development. In *PISA 2018 Assessment and Analytical Framework*. https://www.oecd.org/pisa/data/pisa2018technicalreport/

OECD. (2023). *PISA 2022 Assessment and Analytical Framework*. OECD Publishing. https://doi.org/10.1787/dfe0bf9c-en

Oktatási Hivatal. (2022, augusztus 17). *A digitális országos mérések általános leírása*. https://www.oktatas.hu/kozneveles/meresek/digitalis_orszagos_meresek/altalanos_leiras

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., & Moher, D. (2021). Updating guidance for reporting systematic reviews: Development of the PRISMA 2020 statement. *Journal of Clinical Epidemiology*, *134*, 103–112. https://doi.org/10.1016/j.jclinepi.2021.02.003

R Core Team. (2016). *R: A Language and Environment for Statistical Computing* [Software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Robitzsch, A., Luedtke, O., Goldhammer, F., Kroehne, U., & Koeller, O. (2020). Reanalysis of the German PISA Data: A Comparison of Different Approaches for Trend Estimation With a Particular Emphasis on Mode Effects. In *Frontiers in Psychology* (Köt. 11). FRONTIERS MEDIA SA. https://doi.org/10.3389/fpsyg.2020.00884

Rother, E. T. (2007). Systematic literature review X narrative review. *Acta Paulista de Enfermagem*, *20*, v–vi. https://doi.org/10.1590/S0103-21002007000200001

Şahin, A., & Weiss, D. J. (2015). Effects of Calibration Sample Size and Item Bank Size on Ability Estimation in Computerized Adaptive Testing. *Educational Sciences: Theory & Practice*, *15*(6), 1585–1595. https://doi.org/10.12738/estp.2015.6.0102

Sari, H. İ. (2020). Testing Multistage Testing Configurations: Post-Hoc vs. Hybrid Simulations. *International Journal of Psychology and Educational Studies*, *7*(1), 27–37. https://doi.org/10.17220/ijpes.2020.01.003

Thompson, N. A., & Weiss, D. A. (2011). A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment, Research, and Evaluation, 16*(1), 1–9. https://doi.org/10.7275/WQZT-9427

Urry, V. W. (1970). *A Monte Carlo investigation of logistic test models* [Nem publikált PhD-értekezés, Purdue University]. https://files.eric.ed.gov/fulltext/ED058317.pdf

Vargha, A. (2015). *Matematikai statisztika*. Pólya Kiadó.

Weiss, D. J. (2011). Better Data From Better Measurements Using Computerized Adaptive Testing. *Journal of Methods and Measurement in the Social Sciences*, *2*(1), 1–27. https://doi.org/10.2458/v2i1.12351

Weiss, D. J., & Kingsbury, G. G. (1984). Application of Computerized Adaptive Testing to Educational Problems. *Journal of Educational Measurement*, *21*(4), 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Wise, S. L. (2014). The Utility of Adaptive Testing in Addressing the Problem of Unmotivated Examinees. *Journal of Computerized Adaptive Testing*, *2*(1), 1–17.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Mesa Press.

Publikációk a témában

T. Kárász J., Nagybányai Nagy O., Széll K., & Takács S. (2022). Cronbach-alfa: Vele vagy nélküle? *Magyar Pszichológiai Szemle*, *77*(1), 81–98. https://doi.org/10.1556/0016.2022.00004

T. Kárász J., & Széll K. (2023). Hogyan térnek el a papír-ceruza és számítógépes teszteredmények? - Szisztematikus szakirodalom áttekintés a PISA, TIMSS és PIRLS mérésekkel kapcsolatos tapasztalatokról. *Iskolakultúra*, *33*(3), 51–73.

T. Kárász, J., Széll, K., & Takács, S. (2023). Closed formula of test length required for adaptive testing with medium probability of solution. *Quality Assurance in Education*, *31*(4), 637–651. https://doi.org/10.1108/QAE-03-2023-0042

T. Kárász J., & Takács S. (2021). Adaptív tesztek minimális hosszának, hibájának, értékelési szintjének és a megoldók számának összefüggései—Általános megoldási aránnyal. *Alkalmazott Matematikai Lapok, 38*(1), 39–58. https://doi.org/10.37070/AML.2021.38.1.04

T. Kárász, J., & Takács, S. (2023). Use of open and closed items in automation of evaluation systems. *Alkalmazott Pszichológia, 25*(3), 33–54. https://doi.org/10.17627/ALKPSZICH.2023.3.33

Publikációk a témán kívül

Péter P., Szivák J., Rapos N., & T. Kárász J. (2021). A pályakezdő tanárok tanulásának jellemzői. *Pedagógusképzés, 20*(3), 5–28. https://doi.org/10.37205/TEL-hun.2021.3.01

Rapos N., Szivák J., Tókos K., T. Kárász J., & Lénárd S. (2021). Az optimális tanulási környezet támogatásának intézményi lehetőségei vezetői és tanári nézőpontból. In Fehérvári A., Paksi B., & Széll K. (Szerk.), *Számít-e az iskola?* (o. 87–104). Eötvös Loránd Tudományegyetem; MTMT. https://m2.mtmt.hu/api/publication/32187118

T. Kárász J. (2019). Hibabecslési eljárások véletlen jelenségek paramétereinek becslésére. *Psychologia Hungarica Caroliensis, 7*(2), 104–114. https://doi.org/10.12663/PSYHUNG.7.2019.2.7.

T. Kárász, J., Nagybányai-Nagy, O., Takács, N., & Takács, S. (2022). Egy felsőoktatási e-learning tananyagfejlesztés értékelése. *Educatio, 31*(2), 303–312. https://doi.org/10.1556/2063.31.2022.2.10

Takács, R., T. Kárász, J., Takács, S., Horváth, Z., & Oláh, A. (2021). Applying the Rasch model to analyze the effectiveness of education reform in order to decrease computer science students' dropout. *Humanities and Social Sciences Communications, 8*(1), 1–8. https://doi.org/10.1057/s41599-021-00725-w

Takács, R., T. Kárász, J., Takács, S., Horváth, Z., & Oláh, A. (2022a). Oktatási reform hatékonyságának vizsgálata – Tantárgyak nehézségi elemzése IRT-modell segítségével programtervező informatikus hallgatók körében. *Magyar Pszichológiai Szemle, 77*(2), 209–229. https://doi.org/10.1556/0016.2022.00014

Takács, R., T. Kárász, J., Takács, S., Horváth, Z., & Oláh, A. (2022b). Successful Steps in Higher Education to Stop Computer Science Students from Attrition. *Interchange* 53, 637–652. https://doi.org/10.1007/s10780-022-09476-2

Takács, R., Takács, S., T Kárász, J., Horváth, Z., & Oláh, A. (2021). Exploring Coping Strategies of Different Generations of Students Starting University. *Frontiers in Psychology, 12*. https://www.frontiersin.org/article/10.3389/fpsyg.2021.740569

Takács, R., Takács, S., T. Kárász, J., Oláh, A., & Horváth, Z. (2023). The impact of the first wave of COVID-19 on students' attainment, analysed by IRT modelling method. *Humanities and Social Sciences Communications, 10*(1), Article 1. https://doi.org/10.1057/s41599-023-01613-1

Takács, R., Takács, S., T. Kárász, J., Oláh, A., & Horváth, Z. (2024). Applying Q-methodology to investigate computer science teachers' preferences about students' skills and knowledge for obtaining a degree. *Humanities and Social Sciences Communications, 11*(1), 1–10. https://doi.org/10.1057/s41599-024-02794-z

Tókos K., Rapos N., Szivák J., Lénárd S., & T. Kárász J. (2020). Osztálytermi tanulási környezet vizsgálata. *Iskolakultúra, 30*(8), 41–61. https://doi.org/10.14232/ISKKULT.2020.8.41

Tókos, K., Takácsné Kárász, J., Rapos, N., Lénárd, S., & Szivák, J. (2023). Classroom learning environments and dropout prevention in Hungary. *European Journal of Education, 58*(4), 741–758. https://doi.org/10.1111/ejed.12591

Konferenciamegjelenések a témában

T. Kárász, J., & Széll, K. (2022a). MODE EFFECT AND ITEM EQUIVALENCE IN LARGE-SCALE INTERNATIONAL STUDENT ASSESSMENTS - A SYSTEMATIC LITERATURE REVIEW. In *EDULEARN22 Proceedings* (https://dx.doi.org/10.21125/edulearn.2022.1275; o. 5399–5399). IATED. https://library.iated.org/view/TKARASZ2022MOD

T. Kárász, J., & Széll, K. (2022b). Nagymintás nemzetközi tanulói teljesítménymérések szisztematikus szakirodalmi áttekintése az elektronikus adatfelvétel szemszögéből. In D. Molnár & D. Molnár (Szerk.), *XXV. Tavaszi Szél Konferencia Absztraktkötet* (o. 584). Doktoranduszok Országos Szövetsége (DOSZ).

T. Kárász, J., Széll, K., & Takács, S. (2022a). Adaptív tesztelés során szükséges teszthossz zárt formulája közepes megoldottsági valószínűség mellett. In D. Molnár & D. Molnár (Szerk.), *XXV. Tavaszi Szél Konferencia Absztraktkötet* (o. 711). Doktoranduszok Országos Szövetsége (DOSZ).

T. Kárász, J., Széll, K., & Takács, S. (2022b). CLOSED FORMULA OF REQUIRED ITEM NUMBER FOR ADAPTIVE TESTING WITH MEDIUM PROBABILITY OF ITEM SOLUTION. In *EDULEARN22 Proceedings* (https://dx.doi.org/10.21125/edulearn.2022.1284; o. 5432–5432). IATED. https://library.iated.org/view/TKARASZ2022CLO

T. Kárász, J., & Takács, S. (2019b). Nyílt és zárt végű itemek közötti kapcsolatok az Országos kompetenciamérés (2017-es) adatainak elemzése nyomán. In BME GTK (Szerk.), *I. Szakképzés és Oktatás: Ma – Holnap konferencia. Fejlődés és partnerség: Absztraktkötet* (o. 122–123). BME Gazdaság- és Társadalomtudományi Kar.

T. Kárász J., & Takács S. (2021a). Adaptív tesztek minimális hosszának, hibájának, értékelési szintjének és a megoldók számának összefüggései – általános megoldás. In Molnár G. & Tóth E. (Szerk.), *A neveléstudomány válaszai a jövő kihívásaira* (o. 525). MTA Pedagógiai Tudományos Bizottsága, SZTE Neveléstudományi Intézet. http://edu.u-szeged.hu/onk2021/download/ONK_CES_2021_Absztrakt_Kotet_-_Book_of_Abstarcts.pdf

T. Kárász, J., & Takács, S. (2021b). Adaptív tesztek minimális hosszának, hibájának, értékelési szintjének és a megoldók számának összefüggései – általános megoldási aránnyal. In Sass, J. (Szerk.), *Út a reziliens jövő felé. A Magyar Pszichológiai Társaság XXIX. Országos Tudományos Nagygyűlése* (o. 194–195). Magyar Pszichológiai Társaság. http://mptnagygyules.hu/wp-content/uploads/2021/08/MPT_kivonatkotet_2021_0825.pdf

Takácsné Kárász, J. (2023a). Adaptív teszt képességbecslési és feladatkiválasztási módszereinek összehasonlítása szimulációs módszerrel az Országos kompetenciamérés adatain. In Kasik L. & Gál Z. (Szerk.), *19. Pedagógiai Értékelési Konferencia Absztraktkötet* (o. 66). Szegedi Tudományegyetem Neveléstudományi Doktori Iskola. https://www.edu.u-szeged.hu/pek2023/download/PEK_2023_CEA_2023_absztraktkotet.pdf

Takácsné Kárász, J. (2023b). Adaptív teszt képességbecslési és feladatkiválasztási módszereinek összehasonlítása szimulációs módszerrel az Országos kompetenciamérés adatain. In A. Bajzáth, K. Csányi, & J. Győri (Szerk.), *Elkötelezettség és rugalmasság: A neveléstudomány útjai az átalakuló világban* (o. 423). MTA Pedagógiai Tudományos Bizottság, ELTE Pedagógiai és Pszichológiai Kar. https://onk2023.ppk.elte.hu/download/onk_absztraktok_VEGSO-10-26.pdf

Takácsné Kárász, J. (2023c). Adaptív tesztműködtetési eljárások összehasonlítása szimulációs módszerekkel az Országos kompetenciamérés adatain. In G. Kulcsár & V. D. Horváth (Szerk.), *Találkozás a változásban—Változások a találkozásban: A Magyar Pszichológiai Társaság XXX. Országos Tudományos Nagygyűlése—Kivonatkötet* (o. 45–46). Magyar Pszichológiai Társaság. https://mpt.hu/wp-content/uploads/2023/09/Kivonatkotet_2023.pdf

Konferenciamegjelenések a témán kívül

Gergely, B., T. Kárász, J., & Takács, S. (2019). Különböző mérési modellek az Országos kompetenciamérésben: Mi állhat a hibák hátterében? In BME GTK (Szerk.), *I. Szakképzés és Oktatás: Ma – Holnap konferencia. Fejlődés és partnerség: Absztraktkötet* (o. 66–67). BME Gazdaság- és Társadalomtudományi Kar.

Gergely, B., T. Kárász, J., & Takács, S. (2021). Hol a hiba? Többdimenziós IRT modellek alkalmazása az Országos kompetenciamérésben. In Sass, J (Szerk.), *Út a reziliens jövő felé. A Magyar Pszichológiai Társaság XXIX. Országos Tudományos Nagygyűlése* (o. 192–193). Magyar Pszichológiai Társaság. http://mptnagygyules.hu/wp-content/uploads/2021/08/MPT_kivonatkotet_2021_0825.pdf

Kispál, S., Gergely, B., T. Kárász, J., & Takács, S. (2021). Hátrányban vannak-e a halmozottan hátrányos helyzetűek az országos kompetenciamérésben? In Sass, J (Szerk.), *Út a reziliens jövő felé. A Magyar Pszichológiai Társaság XXIX. Országos Tudományos Nagygyűlése* (o. 193). Magyar Pszichológiai Társaság. http://mptnagygyules.hu/wp-content/uploads/2021/08/MPT_kivonatkotet_2021_0825.pdf

Koltói, L., Harsányi, S. G., Nagybányai-Nagy, O., & Takácsné Kárász, J. (2019). Családi háttér és iskolai teljesítmény—Születni tudni kell? In E. Lippai (Szerk.), *Összetart a sokszínűség* (o. 164). Magyar Pszichológiai Társaság.

Kövesdi, A., Kovács, D., & T. Kárász, J. (2021). Diszgráfia és írási nehézség előfordulása 6., 8., 10. Osztályos gyermekek körében 2012-2018 időszakban. In Sass, J (Szerk.), *Út a reziliens jövő felé. A Magyar Pszichológiai Társaság XXIX. Országos Tudományos Nagygyűlése* (o. 153–154). Magyar Pszichológiai Társaság. http://mptnagygyules.hu/wp-content/uploads/2021/08/MPT_kivonatkotet_2021_0825.pdf

Kövesdi A., Kovács D., & Takácsné Kárász J. (2019). Az SNI-vel és BTM-mel diagnosztizált 6, 8, 10. Osztályos gyermekek iskolai teljesítménye. In Lippai E. (Szerk.), *Összetart a sokszínűség* (o. 165). Magyar Pszichológiai Társaság.

Nádor, A., & T. Kárász, J. (2021). Teljesítményingadozás és annak háttértényezői az országos kompetenciamérés tükrében. In Sass, J (Szerk.), *Út a reziliens jövő felé. A Magyar Pszichológiai Társaság XXIX. Országos Tudományos Nagygyűlése* (o. 154–155). Magyar Pszichológiai Társaság. http://mptnagygyules.hu/wp-content/uploads/2021/08/MPT_kivonatkotet_2021_0825.pdf

Nyitrai E., Takács N., & Takácsné Kárász J. (2019). Szülői bevonódás és iskolai teljesítmény. In Lippai E. (Szerk.), *Összetart a sokszínűség* (o. 163). Magyar Pszichológiai Társaság.

Péter P., Szivák J., Rapos N., & T. Kárász J. (2022). A kezdő tanárok szakmai fejlődése és tanulása. In Steklács J. & Molnár-Kovács Z. (Szerk.), *21. Századi képességek, írásbeliség, esélyegyenlőség. Absztraktkötet* (o. 65). MTA Pedagógiai Tudományos Bizottság – PTE BTK Neveléstudományi Intézet. https://konferencia.pte.hu/sites/konferencia.pte.hu/files/ONK_absztraktkotet_2022.pdf

Péter, P., Szivák, J., Rapos, N., & T. Kárász, J. (2023a). A kezdő tanárok szakmai fejlődése és tanulása az eredményes tanulás modellje mentén. In A. Bajzáth, K. Csányi, & J. Győri (Szerk.), *Elkötelezettség és rugalmasság: A neveléstudomány útjai az átalakuló világban* (o. 381). MTA Pedagógiai Tudományos Bizottság, ELTE Pedagógiai és Pszichológiai Kar. https://onk2023.ppk.elte.hu/download/onk_absztraktok_VEGSO-10-26.pdf

Péter, P., Szivák, J., Rapos, N., & T. Kárász, J. (2023b). Professional Development And Learning Of Novice Teachers. In E. P. Chaw, F. N. Barcın, L. A. Erdei, A. O.

Pongor-Juhász, & E. Kopp (Szerk.), *ATEE Annual Conference 2023: Teacher Education on the Move* (o. 273–275). Association for Teacher Education in Europe (ATEE). https://ateeannual2023.elte.hu/wp-content/uploads/2023/09/ATEE%20Annual%20Conference%202023_Book%20of%20abstracts.pdf

Smohai M., Simon G., & Takácsné Kárász J. (2019). Az Országos kompetenciamérés során felvett szabadidős sporttevékenységre irányuló adatok elemzése. In Lippai E. (Szerk.), *Összetart a sokszínűség* (o. 164). Magyar Pszichológiai Társaság.

T. Kárász J., Nagybányai-Nagy O., Takács N., & Takács S. (2021). Egy elsőéves egyetemi gyakorlat átalakítása a távolléti oktatás igényeinek és lehetőségeinek fényében. In Buda A. & Kiss E. (Szerk.), *Interdiszciplináris Pedagógia a bizonytalanság korában* (o. 55). Debreceni Egyetem Nevelés- és Művelődéstudományi Intézet.

T. Kárász, J., & Takács, S. (2019a). Kevert mérési területek pilot vizsgálata meredekségi és nehézségi paraméterek elemzésével az Országos kompetenciamérés 2017-es adatai alapján. In BME GTK (Szerk.), *I. Szakképzés és Oktatás: Ma – Holnap konferencia. Fejlődés és partnerség: Absztraktkötet* (o. 123). BME Gazdaság- és Társadalomtudományi Kar.

Tókos, K., Rapos, N., Szivák, J., Lénárd, S., & T. Kárász, J. (2021). An Examination of Classroom Learning Environments. In *(Re)imagining & Remaking Teacher Education* (o. 223–224). Association for Teacher Education in Europe (ATEE). https://drive.google.com/file/d/1QdMROSsN5gsBXdIDPiAd7xACYqiVhd7i/view

Tókos K., T. Kárász J., Rapos N., Szivák J., & Lénárd S. (2021). Az osztálytermi tanulási környezet vizsgálata. In Molnár G. & Tóth E. (Szerk.), *A neveléstudomány válaszai a jövő kihívásaira* (o. 437). MTA Pedagógiai Tudományos Bizottsága, SZTE Neveléstudományi Intézet. http://edu.u-szeged.hu/onk2021/download/ONK_CES_2021_Absztrakt_Kotet_-_Book_of_Abstarcts.pdf

Mellékletek

1. melléklet

A médiahatás vizsgálatára irányuló szisztematikus szakirodalmi áttekintés

Fishbein, B., Martin, M. O., Mullis, I. V. S., & Foy, P. (2018). The TIMSS 2019 Item Equivalence Study: Examining Mode Effects for Computer-Based Assessment and Implications for Measuring Trends. L*arge-Scale Assessments in Education, 6*. https://doi.org/10.1186/s40536-018-0064-z

Hamhuis, E., Glas, C., & Meelissen, M. (2020). Tablet assessment in primary education: Are there performance differences between TIMSS' paper-and-pencil test and tablet test among Dutch grade-four students? *British Journal of Educational Technology, 51*(6), 2340–2358. https://doi.org/10.1111/bjet.12914

Jerrim, J. (2016). PISA 2012: How do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy & Practice, 23*(4), 495–518. https://doi.org/10.1080/0969594X.2016.1147420

Jerrim, J., Micklewright, J., Heine, J.-H., Salzer, C., & McKeown, C. (2018). PISA 2015: How big is the 'mode effect' and what has been done about it? *Oxford Review of Education, 44*(4), 476–493. https://doi.org/10.1080/03054985.2018.1430025

Kroehne, U., Buerger, S., Hahnel, C., & Goldhammer, F. (2019). Construct Equivalence of PISA Reading Comprehension Measured With Paper-Based and Computer-Based Assessments. *Educational Measurement: Issues & Practice, 38*(3), 97–111. https://doi.org/10.1111/emip.12280

Robitzsch, A., Luedtke, O., Goldhammer, F., Kroehne, U., & Koeller, O. (2020). Reanalysis of the German PISA Data: A Comparison of Different Approaches for Trend Estimation With a Particular Emphasis on Mode Effects. *Frontiers In Psychology, 11*. FRONTIERS MEDIA SA. https://doi.org/10.3389/fpsyg.2020.00884

Zehner, F., DIPF, Kroehne, U., Hahnel, C., & Goldhammer, F. (2020). PISA reading: Mode effects unveiled in short text responses. *Psychological Test and Assessment Modeling, 62*(1), 85–105. Publicly Available Content Database.

Zehner, F., Goldhammer, F., Lubaway, E., & Sälzer, C. (2019). Unattended consequences: How text responses alter alongside PISA's mode change from 2012 to 2015. *Education Inquiry, 10*(1), 34–55. https://doi.org/10.1080/20004508.2018.1518080