

Károli Gáspár Református Egyetem

Bölcsészet- és Társadalomtudományi Kar

Pszichológiai Intézet

TDK dolgozat

2024

Témavezetők: Gergely Bence, Dr. Takács Szabolcs

A feladatnehézségek változása a minta elemszám függvényében Rasch modell esetén

Hári Anna

M4YV98

Pszichológia BA

III. félév

Tartalom

Absztrakt	3
Bevezetés.....	4
Szakirodalmi bevezetés	6
A tesztelés története.....	6
Klasszikus tesztelmélet	6
Modern tesztelmélet	7
Számítógépes adaptív tesztelés	9
Rasch modell	11
Becslési módok	12
Conditional Maximum Likelihood.....	12
Marginal Maximum Likelihood	13
Expactation-Maximization	13
Módszer	15
Elemzési adatok leírása	15
Szoftver és csomagok.....	16
Elemzés R-ben.....	16
Eredmények.....	18
Diszkusszió.....	26
Irodalomjegyzék.....	28

Absztrakt

A modern tesztelméleti modellen alapú becslési módszerek nagy jelentőséggel bírnak az oktatási és pszichológiai mérésekben (Wainer & Dorans, 2000; Zanon és mtsai., 2016). Eredményeik nem csak a tanulókról adnak visszajelzést, hanem befolyással is vannak az oktatáspolitikai döntésekre (Auxné Bánfi és mtsai., 2017; OECD, 2012). Ezek a mérések azonban csak akkor működnek jól, ha a feladatok előre jól bemértek, vagyis nehézségi szintjük a tesztelés előtt pontosan meghatározott. Ezzel biztosítják, hogy a mérés eredményei megbízhatóan tükrözik a tanulók valós képességeit. Az IRT modell számos különböző móddal becsülhető, azonban egyik becslési mód sem tökéletes, mindegyik különböző feltételrendszeren alapszik, amelyeknek teljesülését nem tudjuk az adatelemzés előtt (Robitzsch, 2021). Kutatásunk célja az volt, hogy a három leggyakoribb becslési módot, valós adaton alapuló szimulációs eljárásban, különböző mintaméreteken (20, 50, 100, 500) megvizsgáljunk, megnézzük, milyen feltételek mellett, hogy működnek, hol működnek jól, vagy tapasztalhatók esetleges torzítások, az egyes mintaméreteken melyik a legjobb módszer. Vizsgáltuk a feladatok nehézségének becslését, valamint a teljes adaton végzett becslésekhez hasonlított átlagos négyzetes eltérést és a válaszokra vonatkozó prediktív pontosságot számoltunk az összehasonlításhoz. A feladatok nehézségének becslésekor, mind a négy mintanagyság esetében, a három becslési mód értékei egyszerre változtak, hasonló trendet követtek, ahogy az egyes feladatok nehézségi szintje nőtt, úgy arányosan a módszerek becslései is növekedtek, kisebb eltérésekkel. Itt a CML becslés konzisztensen felülbecsülte a nehézségeket, mely nem változott a különböző feltételek mellett. A mintaméret növekedésével egyre megbízhatóbbnak bizonyultak az egyes becslések, azonban a hiba nagysága a 100 és 500 fős minták között már jóval kevésbé csökkent, mint a kisebb mintaméretek között. 20 főnél a CML becslés hibája kisebb volt, mint a másik kettőé, azonban ez a nagyobb mintaszám esetében megfordult. 50 főnél az EM bizonyult a legjobbnak, míg 100 és 500 fős mintán az MML becslési mód adta a legpontosabb becslést.

Bevezetés

A pszichológiai és oktatási mérések mindennapi életünk részévé váltak. Segítségükkel mérjük fel és értékeljük az emberek egyes tulajdonságait, mint a képesség vagy a tudás. De vajon mi áll ezeknek a teszteknek a háttérében? Hogyan tudjuk biztosítani, hogy ezek a mérések megbízhatóak legyenek és pontosan mérjenek? Erre talán a legegyszerűbb megoldás a tesztelmélet háttérének és működésmódjának megismerése. Itt felvetődhet az a kérdés, mi szükség van a statisztikai modellek vagy a pszichometria mélyebb ismeretére ahhoz, hogy méréseket végezzünk, hisz ott vannak a tesztekhez az instrukciók, amik szerint használni kell őket. Ez azonban olyan, mintha valaki motorkerékpárt vezet anélkül, hogy ismerné annak működését. Valóban igaz, e tudás nélkül is tudunk vele közlekedni. Viszont jól jöhetnek ezek az ismeretek, ha optimális teljesítményt kívánunk elérni, netán meg is menthetnek minket, ha valami meghibásodik, de elengedhetetlenek, amennyiben kedvünk támad elkészíteni egy ilyen járművet (Hidegkuti & Balázs, 2015). Hasonlóan ehhez a gondolatmenethez a tesztelmélet ismerete nélkül nehéz azt biztosítani, hogy valóban jól, és azt mérjük, amit szeretnénk. Egy fontos aspektus a tesztek készítésénél a feladatok nehézségének pontos megállapítása, ugyanis, ha ez változik, az torzíthatja a diákok teljesítményéről kapott képet, ronthatja a vizsgarendszerek következetességét, amit bizonyítanak az 1995 és 1998 közötti matematikai érettségi vizsgák elemzéséből származó empirikus adatok. Ezek rávilágítottak arra, hogy az egyes évek vizsgáin tapasztalt teljesítménybeli különbségek nem a tanulók képességszintjének változásának tudható be, hanem a feladatok nehézségbeli eltérésének (Tompai, 2001). A klasszikus tesztelmélet mellett egyre inkább előtérbe kerül a moderntesztelmélet vagy item-válasz elmélet (Item Response Theory; IRT), melynek fókuszában elsősorban az egyes feladatok, vagy ahogy majd a későbbiekben is nevezzük őket, itemek állnak. Az IRT lényegében matematikai modellek családját jelenti (Embretson & Reise, 2013; Hambleton & Swaminathan, 1985; van der Linden & Hambleton, 1997), segítségével meg tudjuk határozni, hogy az egyes feladatok mennyire illeszkednek az adott tanuló képességéhez. Így biztosítani tudjuk a pontosabb mérést és egyben kivédeni azt is, hogy túl nehezek vagy túl könnyűek legyenek a kapott kérdések (Wainer & Dorans, 2000; Weiss & Kingsbury, 1984; Zheng, 2014). Ezen a modellen alapul a mai oktatási mérésekben már egyre inkább alkalmazott számítógépes adaptív tesztelés (computerized adaptive testing továbbiakban: CAT), ahol talán még nagyobb

jelentősége van annak, hogy az itemeket milyen pontosan tudjuk előzőleg bemérni, ugyanis itt már nem minden tanuló azonos feladatsort kap maga elé, itt dinamikusan változnak a feladatok a korábbi válaszok függvényében (Wainer & Dorans, 2000). Az itemek bemérésénél a minta mérete, illetve a populáció összetétele, befolyásolja a reliabilitást. Közérthetőbb megfogalmazásban a mérés pontosságát meghatározza mennyien oldják meg a feladatokat és ők milyen képességűek. A helyes nehézségi szint meghatározásához az egyes itemeknél általában sok tanulóra van szükségünk, akik feladatok hosszú sorait oldják meg, ami magától értetődően idő-, és energiaigényes folyamat (Ul Hassan & Miller, 2019; Zheng, 2014). Továbbá nem elhanyagolható az sem milyen becslési módot használunk. Minden modell más feltételek mellett becsül jól, illetve kezd el torzítani (van der Linden, 2016). Olyan nagyszabású oktatási mérések, mint a PISA (Programme for International Student Assessment), melyben a matematika, a szövegértés és a természettudományok területén mérik fel 15 éves diákok teljesítményét, valamint az Országos kompetenciamérés (OKM) hazánkban rendszeresen végzett felmérés is ilyen, előre bemért itemekkel dolgozik (Auxné Bánfi és mtsai., 2017; OECD, 2012). Itt a pontos bemérések nem csak a tanulók képességének helyes felmérése miatt fontosak, hanem azért is, mert az innen származó adatok közvetlen hatással vannak az oktatáspolitikai döntéshozatalokra, hisz ezekre alapozva valósulnak meg az újítások, a tantervi követelmények és az értékelési rendszerek változtatása, amelyek az oktatási rendszerek fejlődését és javítását segítik elő (Breakspear, 2012; Tóth, 2015). Kutatásunk célja, hogy a legelterjedtebb IRT alapú módszereket megvizsgáljuk különböző szituációkban, segítve azt, hogy a bemérési folyamat optimálisabb lehessen. Arra törekszünk, hogy megértsük, milyen feltételek mellett, melyik becslési mód adja a legpontosabb eredményt az itemek nehézségéről, így kevesebb diák bevonásával, rövidebb idő alatt, gazdaságosabban lehessen elvégezni a feladatok bemérését. Mindehhez valós mérésből származó adatokat használtunk, ami azért fontos, mert lehetővé teszi, hogy a modell kimeneteleit összehasonlítsuk, eldöntsük mennyire tükrözik a valóságot, ami elengedhetetlen a modell megbízhatóságának biztosításához, ezzel szemben a mesterségesen generált, idealizált adatokon való becslések általánosíthatósága megkérdőjelezhető, hisz ezeknél feltételezzük, hogy a modell illeszkedik a szimulált mintára (Kleijnen, 1995; Sargent, 2010).

Szakirodalmi bevezetés

A tesztelés története

Az egyéni különbségek felismerésének gyökerei az ókori Görögországig vezethetők vissza, ahol már Platón és Arisztotelész is foglalkozott az emberek közti különbségekkel (Rózsa és mtsai., 2016). Másik kultúrában, az ókori Kínában rendszeres vizsgákkal mérték a hivatalnokok alkalmasságát (Bowman, 1989; DuBois, 1970). A 20. századi Európa közoktatásában problémát jelentett az osztályokban a gyenge képességű diákok jelenléte. A gyerekek szelektálására Binet és Simon dolgoztak ki egy tesztet, mely fontos szerepet tölt be az intelligencia mérésének történetében. Az első világháború idején katonák kiválasztására létrehozott Army Alpha és Beta tesztek szintén hozzájárultak a pszichometriai tesztelés fejlődéséhez (Gregory, 2015b). Az 1926-ban bevezetett SAT (Scholastic Aptitude Test) az oktatási rendszerek szerves részévé vált, lehetővé téve a diákok képességeinek pontosabb mérését és az egyetemi teljesítmény előrejelzését (Gregory, 2015a; Wainer, 2000).

Klasszikus tesztelmélet

„Ami létezik, az valamilyen mennyiségben létezik” (Thorndike, 1918) és „ami mennyiségben létezik, az mérhető is” (McCall, 1939). A klasszikus tesztelmélet alapfeltevése szerint a személyiség összetett építőeleme, azaz a konstruktum, mérhető és rendelkezik egy valódi értékkel. Ezt a valódi értéket szeretnénk megbecsülni a mérések során, azonban minden mérés hibát is hordoz magában. Vagyis az az érték, amit a méréskor megfigyelünk a valódi érték mellett valamilyen mértékű hibát is tartalmaz (Nagybányai Nagy, 2006).

$$X = T + e$$

(X = megfigyelt érték; T = valódi érték, angolul: true score; e = valamilyen mértékű hiba, angolul: error score)

Annál megbízhatóbbnak tekinthetünk egy eszközt minél jobban és pontosabban mér, más szóval a mért értékek minél kevesebb hibát tartalmaznak, annál jobban tükröződik bennük a valódi érték. A hibákon belül elkülöníthetünk szisztematikus és véletlen hibákat. Az előbbiről akkor beszélünk például, ha véletlen ott maradnak a megoldások a papír szélén, így a tanulók képességét szisztematikusan felül mérné a teszt. A véletlen hibák azonban semmi ilyen

szisztematikusságot nem mutatnak, nem egy irányba torzítanak, hol lefelé hol felfelé módosítják a valódi értéket (Hidegkuti & Balázs, 2015; Nagybányai Nagy, 2006).

Három alapvető feltétel van, amire a klasszikus tesztelmélet épül (Lord és mtsai., 2008):

1. A hibák középértéke nulla, ami abból következik, hogy nem egy irányba torzítanak a véletlen hibák, így végső soron kiegyenlítik egymást.
2. A hiba nem korrelál a valódi értékkel, a hibák mértéke nem függ össze a valódi érték nagyságával.
3. A hibák sem korrelálnak egymással, függetlenek egymástól (Hidegkuti & Balázs, 2015; Nagybányai Nagy, 2006).

Az alapelvek leírása Spearman nevéhez kötődik, ahogy a reliabilitás fogalma is. (Guttman, 1945).

Modern tesztelmélet

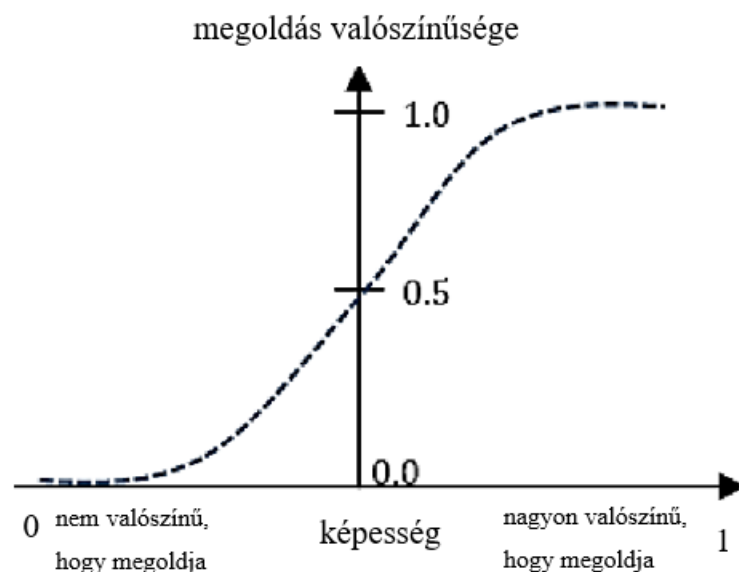
A modern tesztelmélet vagy más néven látens vonáselmélet vagy item-válasz elmélet (Item Response Theory; IRT) története 1916-ban kezdődött. Binet és Simon voltak az elsők, akik a teljesítményszinteket egy független változó függvényében ábrázolták, és ezeket az ábrákat használták a tesztfejlesztés során. Ezzel elindult az a folyamat, amely később az IRT kialakulásához vezetett. Az IRT korai szakirodalmához legbefolyásosabban hozzájáruló Frederic M. Lord megalkotott egy modellt, amely lehetővé tette, hogy a tesztek eredményei jobban tükrözzék a vizsgált képességeket. Ezt Birnbaum tovább egyszerűsítette, hogy még könnyebben alkalmazható legyen a tesztelési gyakorlatban. A 20. század második felében Georg Rasch új típusú modelleket fejlesztett ki, amelyek jelentős hatással voltak a tesztelési módszerekre (Hambleton & Swaminathan, 1985).

Az IRT matematikai modellek családjára utal, ami paradigmaváltást jelentett az oktatási és pszichológiai tesztek értékelésében. Elsősorban az egyes itemek állnak a fókuszában nem pedig az itemekből összeállított tesztek. Az elmélet alapjául szolgál, hogy közvetlenül megfigyelhető, manifeszt változók segítségével próbálunk mérni nem megfigyelhető, látens tulajdonságot, képességet. Feltételezi, hogy a látens változó és a mért itemek egy nem megfigyelhető kontinuumon helyezkednek el, így főként az egyén ezen való elhelyezésére fókuszál, modellezi hogyan alakul a helyes válasz valószínűsége a látens képesség függvényében (Hidegkuti & Balázs, 2015; Thompson, 2024). Általában a feladat megoldásának valószínűségét ICC (Item

Characteristic Curve) segítségével ábrázolják. Ezen görbe ábrázolásánál a vízszintes tengely a képesség, a függőleges tengely a feladat vagy item megoldásának valószínűsége. Ahogy az 1. ábrán is látszik a gyenge képességűek kis, az átlagos képességűek már kétes, a nagyon jó képességű tanulók viszont már nagy valószínűséggel oldják meg az aktuális feladatot. Tehát az ICC görbék általában szigmoid alakúak, és azt mutatják, hogy a képesség növekedésével nő a helyes válaszadás valószínűsége (Petra & Aziz, 2020).

1. ábra

ICC görbe



A teljesítménytesztek során az egyes itemekre a válasz lehet helyes vagy helytelen, vagyis az adott tanuló vagy megoldotta vagy nem oldotta meg a feladatot. Az ilyen itemeket, változókat kétértékű vagy dichotóm változóknak nevezzük. Az IRT modellek ilyen dichotóm változókra adott válaszokat modelleznek (Hidegkuti & Balázs, 2015). Az IRT tehát egy árnyaltabb képet ad a tesztelés eredményeiről, hisz alkalmazkodik az itemek nehézségéhez és az egyének képességbeli különbségeihez. Emellett tesztelhetővé válik az, hogy a teszt valóban egy dimenziót mér-e, illetve nem feltételezzük, hogy minden feladat ugyanolyan jól méri a valós képességet (Cai és mtsai., 2016; Embretson & Reise, 2013; Hambleton & Swaminathan, 1985; van der Linden & Hambleton, 1997).

Számítógépes adaptív tesztelés

A papíralapú tesztek korlátai közé tartozik, hogy minden tesztalany ugyanazokat a kérdéseket kapja meg, nincs különbség a gyengébb és a kiemelkedően jó teljesítményű tanulók feladatsorai között. Azért, hogy pontosabban lehessen felmérni a tanulók képességét, jobban illeszkedjen hozzájuk a teszt, létrejött az 1970-es években a számítógépes adaptív tesztelés (computerized adaptive testing; CAT), amely az IRT elméletén alapul, és melynek szemléltetésére talán Wainer & Mislevy (2000) példája ¹a legalkalmasabb. Az adaptív tesztek is hasonlóan működnek, mint egy atlétika versenyen a magasugrók magasságának megválasztása. Minden egyes tesztalany különböző nehézségű kérdéseket kap, és a cél az, hogy az egyén képességeihez leginkább illeszkedő kérdéseket válasszuk ki. Az IRT egyik legnagyobb előnye, hogy lehetővé teszi a feladatok jellemzőinek pontos meghatározását, ami az adaptív tesztelésben elengedhetetlen. Az IRT modellezi a feladatok és a vizsgázók közötti kapcsolatot, és minden kérdésről megállapítja, hogy mennyire nehéz, mennyire jól különbözteti meg a vizsgázókat, és mekkora az esélye, hogy valaki találgatással helyesen válaszol. Így az adaptív tesztelés során mindig a legmegfelelőbb feladatot lehet kiválasztani az egyes vizsgázók számára, ahogy a magasugrásnál is megválasztják a megfelelő kezdőmagasságot az ugró képességeinek megfelelően (Wainer & Mislevy, 2000).

A CAT célja, hogy a kérdéseket a vizsgázó teljesítményéhez igazítsa, így mindenki a saját képességeinek megfelelő nehézségű feladatokkal találkozhat. Ugyanis azoknál a teszteknel, ahol egyformák a feladatok mindenki számára, a jó képességű tanulóknak nagyon sok feladatot kell megoldaniuk, ezzel szemben a rosszabb képességűeknek meg lehet, hogy csak néhány feladat megfelelő nehézségű. A megfelelő feladatok kiosztásával így tehát a tesztelés ideje is lerövidülhet. A rendszer működése mögötti elv viszonylag egyszerű: ha egy vizsgázó helyesen válaszol egy kérdésre, akkor magasabb lesz a képességbecslése, vagyis a következő kérdés

¹ Vegyünk egy futóversenyt, ahol a versenyzők képességeit szeretnénk összehasonlítani. Ha minden versenyző ugyanazon a pályán futna, egyszerűen csak azt néznénk meg, ki ért célba először. Aki a legtöbb helyes választ adja a teszten, azt tartjuk a legfelkészültebbnek. De mi van akkor, ha a versenyzők különböző napokon vagy különböző pályákon futnak? Ebben az esetben valamilyen módon össze kellene hasonlítani őket, függetlenül a különbségektől. Hasonló helyzetben van a tesztelés is, ha különböző tesztváltozatokat alkalmazunk. Az egyik megoldás erre az esetre, hogy minden pálya szabványos: ugyanolyan hosszú, sík, azonosak az időjárási körülmények, mint a szél erőssége és a hőmérséklet. Ez azt biztosítja, hogy a versenyzők a lehető leginkább azonos körülmények között versenyezzenek. Ez a módszer hasonló a standardizált tesztekhez, ahol a különböző tesztformák ugyanabból a feladatbankból merítenek, és statisztikai módszerekkel korrigálják az esetleges nehézségi különbségeket. Ez olyan, mint amikor különböző maratonok eredményeit hasonlítják össze, például a bostoni és a new york-i maraton időeredményeit, amelyek bár ugyanolyan hosszúak, a különböző körülmények miatt eltérő eredményeket hoznak. Az atlétikánál maradvány a magasugrás talán még jobban példaként vehető az adaptív tesztelésre. Az ugrók saját képességeik alapján választják ki a kezdő magasságot. A jobb ugrók átugorják az alacsonyabb magasságokat, és energiájukat a magasabb, nehezebb ugrásokra tartalékolják. Az alacsonyabb szinten lévő ugrók viszont korábban kezdik a versenyt. Ha valaki nem tud átugrani 3 métert, akkor nyilvánvaló, hogy nem olyan jó, mint az, aki 4 métert ugrott át, még akkor is, ha az utóbbi nem próbálkozott a 3 méteres ugrással.

nehezebb lesz, mivel nehezebb feladat az, amit 0,5-ös valószínűséggel megold. Abban az esetben, ha hibázik, ugyanezen elv alapján könnyebb kérdés következik, így igazodva a vizsgázó képességeihez, és eredményképp biztosítva a pontosabb felmérést. A mérésben két dolgot tudunk optimalizálni. Megállíthatjuk a tesztet, ha elértünk egy adott mérési hibát, vagyis ebben az esetben a pontosságot optimalizáljuk vagy megtehetjük ezt a feladatok számával is, megszabjuk, hány feladat után álljon le a mérés. Azonban minden alkalommal csak úgy tud működni egy ilyen rendszer, ha be vannak mérve a feladatok, viszont a feladatbemérés is hasonló problémákba ütközik. A CAT rendszerek igazán az 1980-as években kezdtek el terjedni, amikor a számítógépek technológiai fejlődése lehetővé tette a szélesebb körben való alkalmazásukat. A CAT számos előnnyel rendelkezik, ahogy már korábban tárgyalásra került. Fontos pozitívuma, hogy növeli a pontosságot, mivel a kérdéseket az egyéni képességekhez igazítja, és az eredmények azonnal elérhetők. A biztonság is javul, hiszen minden vizsgázó más kérdéseket kap, így a csalás esélye minimálisra csökken. A tesztelés hatékonyabbá válik, mivel a vizsgázóknak nem kell feleslegesen könnyű vagy nehéz kérdésekkel időt tölteniük (Thissen & Mislevy, 2000; Wainer, 2000; Wainer & Eignor, 2000). Az azonban vitatott, hogy ez a tesztelési rendszer milyen hatással is van a tanulókra. Egyes kutatások szerint növeli (Arvey és mtsai., 1990; Brüggemann és mtsai., 2024; Ling és mtsai., 2017), mások szerint csökkenti a motivációt (Frey és mtsai., 2009; Ortner és mtsai., 2014), de volt olyan felmérés is, ahol nem találtak semmilyen hatást (Martin & Lazendic, 2018). A CAT hatásai a negatív teljesítményérzelmekre, mint például a szorongás, harag, szégyen, unalom, reménytelenség és csalódottság változók terén hasonlóan kétértelmű empirikus eredmények léteznek: Ling és mtsai. (2017) negatív (alacsonyabb szorongás a CAT esetében, mint a lineáris tesztelésnél) és pozitív hatásokat is jelentettek, míg Kim (1995) erős pozitív hatásokról számolt be (Frey és mtsai., 2024). Ortner és mtsai. (2014) szerint a motivációt és az egyes viselkedési jellemzőket az adaptív teszteléssel kapcsolatban a tanuló saját teljesítményéről alkotott önértékelése határozza meg. Beszámolók alapján a tanulók hozzá vannak szokva a könnyebb, fixált nehézségű feladatokhoz, melyeket nagyobb valószínűséggel oldanak meg (Eggen & Verschoor, 2006). A CAT esetében azonban ehhez viszonyítva kisebb esélye van a jó megoldásoknak, ezért ez a tesztelési mód a hagyományoshoz képest nehezebbnek tűnhet a diákok számára, és/vagy a gyengébb képesség érzését keltheti egyesekben, ez főként a nagyon jó képességűek körében fordulhat elő. Azonban a CAT során a flow érzés minden gyereknél egyforma a teljesítménytől függetlenül, amely alátámasztja azt a feltevést, hogy valóban személyesen releváns volt a teszt a kitöltők számára (Ortner és mtsai., 2014). Más kutatásban pedig azt találták, hogy a CAT

csökkentette az állapotszorongást, viszont csak a tesztírók bizonyos csoportjainál (Hontangas és mtsai., 2000).

Rasch modell

Ahogy már korábban is említésre került, az IRT matematikai modellek nagy családjában vannak egydimenziós modellek, mint az 1-paraméteres, 2-paraméteres és 3-paraméteres logisztikus modellek, ahol az 1PL modell csak a nehézségi paramétert (b) használja, a 2PL modell hozzáadja a diszkriminációs paramétert (a), míg a 3PL modell tovább bővíti a modellt a tippelés paraméterével (c), így egyre pontosabban becsülik a feladatok megoldásának valószínűségét. Ezen felül vannak többdimenziós, hierarchikus válaszmodellek, és még számtalan másik, továbbfejlesztett, más elveket is alkalmazó modell. Az IRT területén egyre több kutató törekszik új modellek létrehozására és a meglévők módosítására. Az elmúlt két évtizedben az IRT modellek elterjedése drámai mértékben megnőtt. A vezető szakfolyóiratokban gyakran jelennek meg olyan cikkek, amelyek újításokat vagy teljesen új modelleket mutatnak be, gyakran alapvető változtatásokkal. Ez az új modellkészítő trend nagyrészt a publikációs nyomásnak köszönhető, mivel a kutatók új megközelítések kidolgozására összpontosítanak, hogy növeljék esélyüket a publikálásra. Az innováció ellenére minden modellnek vannak korlátai, és bizonyos feltételek mellett még a legkifinomultabbak is téves becsléseket adhatnak (van der Linden, 2016). Gyakran elhanyagolják az alapvetőbb modellek alapos értékelését és összehasonlítását, pedig ezeknek az egyszerűbb modelleknek a megértése elengedhetetlen ahhoz, hogy felmérjük teljesítményüket és gyakorlati alkalmazhatóságukat a valós helyzetekben (van der Linden, 2016).

Ezért ezek közül a kutatásban a legalapvetőbb, egydimenziós modellek családfájának tetején levő Rasch modellt fogjuk alkalmazni (von Davier, 2016). Egy dichotóm modellről van szó, mely felírható a következő képlet segítségével:

1. egyenlet Rasch modell

$$p_i(x_{ij} = 1 | \theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}$$

Ebben a képletben (1. egyenlet) p_i jelzi annak valószínűségét, hogy a „j”-dik személy helyesen válaszol az „i”-dik kérdésre ($x_{ij}=1$), ha θ_j ennek a személynek képessége, míg b_i az item nehézségi paramétere, amely megmutatja, hogy milyen képesség szükséges az adott item helyes

megválaszolásához. A modell szerint a helyes válasz valószínűsége attól függ, mennyire közel áll a személy vagy tanuló képessége a kérdés nehézségi szintjéhez. Ha a különbség a kettő között nulla, akkor 50% a valószínűsége annak, hogy a feladatot meg tudja oldani. Ha a tesztalanyunk az item „felett” van akkor ez a valószínűség növekszik és fordítva (Weiss, 2021).

Becslési módok

A Rasch modellt több becslési móddal is tudjuk becsülni, melyekből hármat használtunk fel a kutatásban.

Conditional Maximum Likelihood

A három becslési mód közül az első a Conditional Maximum Likelihood (továbbiakban: CML), melyet Rasch (1960) fejlesztett ki. A CML úgy becsüli a Rasch modellben a feladatok nehézségét, hogy nem igényel feltételezést a populáció eloszlásáról. Nem feltétel, hogy a diákok képességei normális eloszlást kövessenek, ehelyett a nehézségek valószínűségi függvényét a diákok nyers pontszámainak (a diák által megoldott feladatok számának) segítségével írja fel, ugyanazokat az item paraméter becsléseket nyújtja, függetlenül a populáció eloszlásának megválasztásától. Ezért robusztusabb, és általában előnyben részesített eljárás (van der Linden & Hambleton, 1997). Viszonylag kis minták is elegendőek megbízható becslések eléréséhez (Mair & Hatzinger, 2007). Ahhoz, hogy megkapjuk az itemek becslését egy lineáris egyenletrendszer megoldására van szükség. Kezdetben a számítási problémák túl nehéznek bizonyultak, a 60-nál több itemből álló tesztekre való alkalmazás nem volt megvalósítható. Ezeknek az optimális megoldását széleskörben kutatták (De Leeuw & Verhelst, 1986; Gustafsson, 1980; Liou & Price, 1994). Új, hatékonyabb algoritmusok jelentek meg, amelyekkel már több száz paramétert is lehet CML-módszerrel becsülni (van der Linden & Hambleton, 1997). Mivel ez a becslés a feladatok nehézségére vonatkozik, hátránya az, hogy a képesség paramétereit nem becsüli meg. Így, ha annak számítására is szükségünk van, akkor ezt egy második lépésben tehetjük meg (van der Linden, 2016).

Marginal Maximum Likelihood

A feladatok nehézségének és a tanulók képességének együttes becslése CML becsléssel (Andersen, 1972; Fischer & Parzer, 1991; Wright & Masters, 1982) statisztikailag sok helyzetben nem ad stabil eredményeket, a becslés minősége függ az itemek számától és nehézségétől is. Erre Darrell Bock & Lieberman (1970) a Marginal Maximum Likelihoodot (továbbiakban: MML) javasolta, kétértékű itemek esetére (van der Linden & Hambleton, 1997). Ez a becslési mód nem becsüli meg közvetlenül minden tanuló egyedi képességi szintjét, hanem kiszámolja a valószínűségét a megfigyelt válaszoknak a látens képesség összes lehetséges értékére való átlagolás (marginalizálás) alapján. Vagyis, ha figyelembe vesszük az összes lehetséges képességet, akkor mennyire valószínű egy válaszminta előfordulása, amennyiben szem előtt tartjuk az item paramétereit, nehézségeket. Az MML tehát megpróbálja megtalálni azokat a paramétereit, amik a meglévő válaszokat leginkább valószínűvé teszik (Liu & Chalmers, 2018; van der Linden, 2016). Az MML becslések konzisztensek és aszimptotikusan hatékonyak, vagyis a minta méret növelésével egyre pontosabbak lesznek a becslések (Andersen, 1973). Az MML becsléshez általában Newton-Raphson (NR) módszert használnak, ami egy iteratív módszer a függvény maximumának megtalálásához. Azonban, később elterjedt az Expectation-Maximization is, mely szintén használható, a korábbi módszer helyett. (Rabe-Hesketh & Skrondal, 2016).

Expectation-Maximization

Az utolsó becslési mód az Expectation-Maximization (továbbiakban: EM), mely két ismétlődő lépésben működik. Fontos megemlíteni, hogy míg a másik kettő becslési mód zárt formulát ad a paraméterek becslésére, amelyeket ezt követően valamilyen numerikus módszerrel oldunk meg, addig az EM esetén, az egész algoritmus numerikus approximációk egyre pontosodó sorát adja. Az E-lépés, amiben látens képességek várható értékeit becsüli meg a jelenlegi paraméterbecslések alapján. A második M-lépés vagy maximalizálási lépés az, ahol új paraméter becsléseket csinál, amelyek maximalizálják az E-lépés során számított teljes adathalmaz valószínűségének várható értékét. Ami lényegében annyit tesz, hogy megkeresi azokat a paraméter értékeket, amik a legjobban illeszkednek az adatra, beleszámítva a hiányzó adat várható értékét. Ez a két lépés iteratívan ismétlődik, ami mindaddig tart, amíg a paraméterbecslés nem konvergál. Sok előnyös tulajdonsága van, mint például, hogy sokféle modellhez alkalmazható, viszonylag robusztus a kiinduló paraméterekre, a kiinduló becslések

nem befolyásolják jelentősen a végső eredményt. Az iteratív módú becslésnek köszönhetően általában hatékonyabb, mint más becslési módszerek, különösen, ha magas a hiányzó adatok aránya. Viszont ennek a becslésnek is vannak hátulütői, mégpedig különösen nagy adathalmazok esetén lassan konvergálhat, sok cikluson keresztül futhat. Továbbá nem garantált az, hogy a legjobb megoldást - lehet csak a helyi, és nem a globális maximumot - találja meg (van der Linden, 2016; van der Linden & Hambleton, 1997).

Mind a három becslési módnak megvannak az előnyei, nincs egyetlen tökéletes módszer, mindegyik különböző feltételek mellett bizonyul a leghatékonyabbnak (Robitzsch, 2021). Nincs egyértelmű elemszám vagy mintaméret, ahol az egyik mód a legjobb választás, hisz függ a minta összetételétől is mit kapunk eredményként. Lord (1983) kutatásai alapján $N < 100 - 200$ mintanagyság esetében a Rasch modell pontosabb, mint a két paraméteres becslések. Zwinderman (1991) vizsgálatai szerint nagy mintanagyság és modellhez illeszkedő adatok esetén az MML hatékonyabb, modellhez nem illeszkedő adatok (pl. ferde eloszlás) esetén a CML robusztusabb. Mair & Hatzinger (2007) CML becslés esetén egy 100×30 -as adatmátrixon jó illeszkedést találtak. Kubinger és mtsai. (2012) rámutattak arra, hogy egy mintaméretnél sem tökéletesek a becslések, a CML torzított eredményeket adott szélsőséges itemek (nagyon nehezek vagy nagyon könnyűek) esetén, az MML becslésnél pedig a konvergencia problémája is felmerült nagyobb mintaméreteknél.

A következőkben bemutatjuk a kutatáshoz felhasznált adatok forrását, illetve megnézzük, hogy magukhoz viszonyítva a három becslési mód milyen mintaméret mellett adja a legjobb becslést.

Módszer

Az elemzéshez a 2017-es évi Országos kompetenciamérés adatait használtuk fel. A kompetenciamérés egy olyan országos felmérés, amely a 6., 8. és 10. évfolyamos tanulók szövegértési és matematikai képességeit méri fel, és melynek célja, hogy betekintést adjon az oktatási rendszer eredményességébe, és az esélyegyenlőség helyzetébe. Ez a mérés nem vizsga, nem kapnak rá érdemjegyet a diákok, a felmérésből származó eredmények az oktatási rendszer fejlesztését szolgálják. A felmérést az oktatásért felelős miniszter rendeli el a tanév rendjében. Az országban minden iskola 6., 8. és 10. évfolyamán tanuló összes diák részt vesz rajta, kivéve a mérés alól felmentetteket, például a sajátos nevelési igényű tanulókat, illetve azokat, akik kevesebb, mint egy éve részesülnek magyar nyelvű oktatásban vagy ideiglenes testi sérülés miatt nem tudnak részt venni a mérésben (Auxné Bánfi és mtsai., 2017). A méréshez kapcsolódó feladatokat, mint például maguknak a méréseknek a lebonyolítását, a kérdőívek elkészítését, az adatok rögzítését és jelentések elkészítését az Oktatási Hivatal végzi. A kompetenciamérés során a tanulók azonos feladatokat kapnak, melyeket négyszer 45 perces időkeretben kell megoldaniuk. Ezen feladatok egyik fő területe kétszer 45 percben a szövegértés, ahol a tanulóknak különböző szövegeket kell feldolgozniuk és válaszolni a tartalmukhoz kapcsolódó kérdésekre. Másik fő terület a matematika, ahol a mindennapi életből származó példákat kell megoldani a felmérést író gyerekeknek. Tehát elsősorban nem a tantervben meghatározott tananyag elsajátítását méri fel, inkább azt a képességet, hogy valódi megoldandó helyzetek kapcsán mennyire tudják alkalmazni őket (Auxné Bánfi és mtsai., 2017; Szabó és mtsai., 2018). A kutatáshoz felhasznált adatok a 2017-es Országos kompetenciamérésből származnak, melyre 2017. május 24-én tizenötödik alkalommal került sor (Szabó és mtsai., 2018). A mérésből származó adatok másodelemzésének etikai engedélyszáma: 321/2018/P.

Elemzési adatok leírása

Az elemzéshez a 10. osztályosok matematika eredményeit használtuk fel. A teszt sor összesen 56 ilyen feladatot tartalmazott és 84 957 tanuló töltötte ki, ebből 73 728 tanuló eredményei számítottak bele az országos átlagba, melyből 36 461 fiú és 37 267 lány. A matematika képességszint (mely 0-tól 7-ig terjed) mediánja és módusza is 4. Egyes feladatokra részpontokat is kaphattak a tanulók, az adatbázisban kétféle ilyen feladat szerepel az MN11601p és az MN10401p kódú. Ezeknél a feladatoknál, csak a teljesen megoldott feladatokat számítottuk megoldottnak, amennyiben csak részpontot kapott az adott tanuló, úgy már nem megoldott feladatként kódoltuk.

Szoftver és csomagok

Az elemzések az R 4.3.3 verziószámú programban lettek futtatva (R Core Team, 2021). Az R-hez különböző csomagokat fejlesztettek és fejlesztenek. A CML becslés az eRm csomagban található, ami Andersen (1972) által javasolt összegző algoritmust használja (Mair & Hatzinger, 2007). A mirt egy átfogó pszichometriai csomag R-ben, amely különféle modell alapú funkciókat kínál az IRT modellek illesztéséhez és elemzéséhez, beleértve az MML NR és EM alapú becsléseit (Liu & Chalmers, 2018) és az EM becslést is (Chalmers, 2012).

Elemzés R-ben²

Az elemzés során először importáltuk az adatfájlt R-be, mely a tanulók feladatmegoldásait tartalmazza. Ezután kiválasztottuk a matematikához tartozó „MN” előtagú kóddal rendelkező feladatokat. Mivel nem volt minden feladat kétértékű, ezért a következő lépésben kivettük azokat a feladatokat, ahol több értéket is kaphattak a tanulók és átkódoltuk őket az egyszerűbb elemzés érdekében. Itt a 2-es értékeket 1-esre változtattuk, az 1-eseket 0-ra, míg a 0-kat változatlanul hagytuk, és az átkódolt oszlopokat visszahelyeztük az eredeti adatállományba. Ezután készítettünk ebből egy tisztított adatállományt, amiből kitöröltük azokat a sorokat, ahol csak hiányzó érték volt.

Következő lépésben kiszámoltuk az egyes itemek nehézségét, megnéztük az átlagnehézséget és kiválasztottuk azokat az itemeket, amik az átlaghoz képest $\pm 1,5$ egységen belül voltak. Ezekből az itemekből véletlenszerűen kiválasztottunk 20-at további elemzés céljából, ellenőriztük, hogy megfelelő varianciával rendelkeznek-e.

Létrehoztunk egy tesztelési adatállományt is, amihez véletlenszerűen kiválasztottunk 1000 tanulót a tisztított adathalmazból. Ezzel együtt a többi tanuló válaszai külön az eredeti, tisztított adathalmazban maradtak, így biztosítva, hogy az elemzésben dolgozhassunk a különböző adathalmazokkal.

Ezután a 20 kiválasztott itemhez sorban kisorsoltunk 500, 100, 50 és 20 tanulót az eredeti, tisztított adathalmazból. A méreteket úgy választottuk, hogy 20 fő megfelel egy kisebb osztálylétszámnak, 50 fő egy vidéki iskola évfolyamának, 100 fő egy nagyobb gimnázium évfolyamának, 500 fő pedig egy kisebb általános iskola létszámának vagy egy nagyobb

² Az elemzéshez írt szkript megtekinthető az alábbi linken:
https://osf.io/bzqmy/?view_only=65fcbe08e0e44e11aa51cb338e0f4f7e

egyetemi kar évfolyamának. Itt minden egyes feltételnél megnéztük, hogy a kiválasztott tanulók válaszainak van-e varianciája. Amennyiben a variancia nulla, akkor megnéztük, hogy minden válasz 1 vagy 0-e. Az első esetben egy 0-t tartalmazó megoldást, a második esetben egy 1-et tartalmazó megoldást adtunk hozzá a kiválasztott adatállományhoz, így megkaptuk a különböző méretű szűrt adatállományokat.

A következő lépésben a különböző mintaméreteken mind a három becslési módszerrel, egyesével, Rasch modellel megbecsültük az itemek nehézségét. A CML becsléssel alaphól könnyűséget (easiness) lehet számolni (Mair és mtsai., 2024), így ezeknek az értékeknek vettük az ellentétét, hogy a másik két módszerrel összehasonlítható nehézség értékeket kapjunk. A nehézségek becslése után kiszámoltuk a szűrt adatállományon és a teljes (tisztított) adatállományon becsült itemek nehézségének átlagos négyzetes eltérését. Ezáltal azt vizsgáltuk, hogy a teljes 80 ezer fős mintához képest milyen mértékben térnek el a kisebb adatbázisokon végzett elemzések item nehézségei.

Az átlagos négyzetes eltérés (mean squared error; MSE) statisztikai modellekben a hiba mértékét méri. Jelen esetben felmértük az átlagos négyzetes különbséget a teljes és a kisebb mintanagyságokon becsült értékek között. Ha a kisebb mintán végzett elemzések item nehézségei megegyeznek a teljes adat item nehézségeivel, akkor az MSE értéke 0 (Robeson & Willmott, 2023)

Utolsó lépésben a prediktív pontosság került kiszámításra, ami azt mutatja meg, hogy az adott becslési mód mennyire képes megjósolni olyan diákok válaszmintázatait, akik nem szerepeltek abban a mintában, amely segítségével az itemeket kalibráltuk (Meloun & Militký, 2011; Rezvani & Wang, 2023). Kiszámoltuk a szűrt adatállományon az itemek nehézségét. Majd ezen nehézségek segítségével a tesztelési adatállományban levő tanulók képesség paramétereit (théta) EAP (expected a posteriori) módszerrel megbecsültük. A becslés pontosságát a becsült és a valós válaszok közötti átlagos négyzetes hiba (MSE) kiszámításával néztük meg. Azonban ezt csak az MML és EM becslések esetében számoltuk ki, mivel az eRm csomagban, mellyel a CML becslést végeztük, nem lehetséges a théta értékek becslése új, ismeretlen adatokon.

Eredmények

Az elemzés során kapott eredmények közül először az egyes becslési módszerekkel, különböző mintaméreteken kapott nehézségeket nézzük meg. A könnyebb értelmezhetőség érdekében az adatokat a három becslési mód által kapott nehézségek átlaga szerint sorba rendeztük.

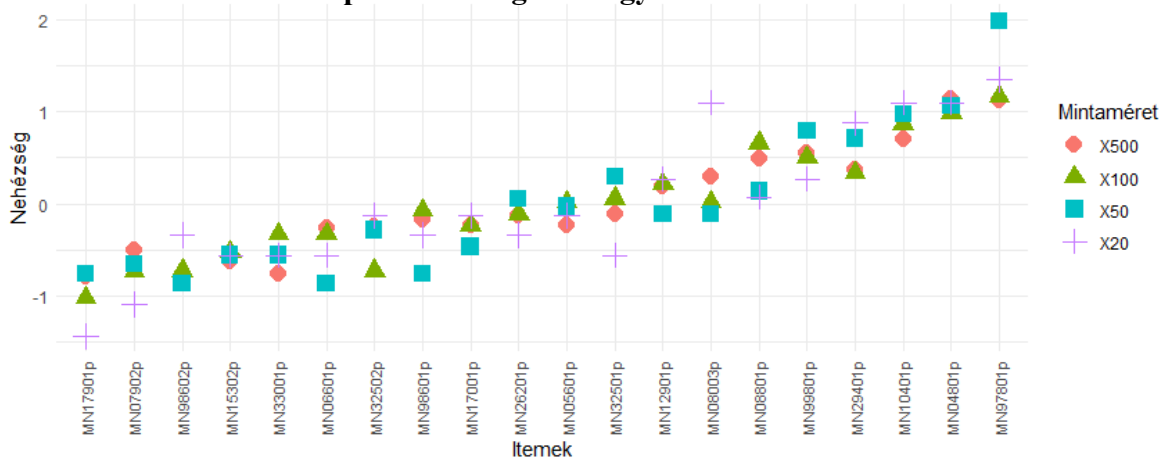
1. táblázat

CML becslési móddal kapott nehézségek az egyes mintaméreteken

Itemek	500 fős minta	100 fős minta	50 fős minta	20 fős minta
MN17901p	-0,787	-1,024	-0,758	-1,437
MN07902p	-0,497	-0,728	-0,656	-1,092
MN98602p	-0,864	-0,728	-0,865	-0,335
MN15302p	-0,623	-0,521	-0,558	-0,560
MN33001p	-0,755	-0,331	-0,558	-0,560
MN06601p	-0,261	-0,331	-0,865	-0,560
MN32502p	-0,244	-0,728	-0,284	-0,125
MN98601p	-0,167	-0,069	-0,758	-0,335
MN17001p	-0,227	-0,241	-0,464	-0,125
MN26201p	-0,124	-0,111	0,054	-0,335
MN05601p	-0,227	0,015	-0,029	-0,125
MN32501p	-0,107	0,056	0,299	-0,560
MN12901p	0,189	0,218	-0,112	0,272
MN08003p	0,301	0,015	-0,112	1,101
MN08801p	0,487	0,657	0,136	0,076
MN99801p	0,552	0,497	0,796	0,272
MN29401p	0,374	0,338	0,711	0,878
MN10401p	0,708	0,861	0,973	1,101
MN04801p	1,140	0,989	1,065	1,101
MN97801p	1,131	1,166	1,987	1,348

2. ábra

CML becslési móddal kapott nehézségek az egyes mintaméreteken



Az item nehézségek után az MSE értéket is megtekintjük az egyes becslési módoknál.

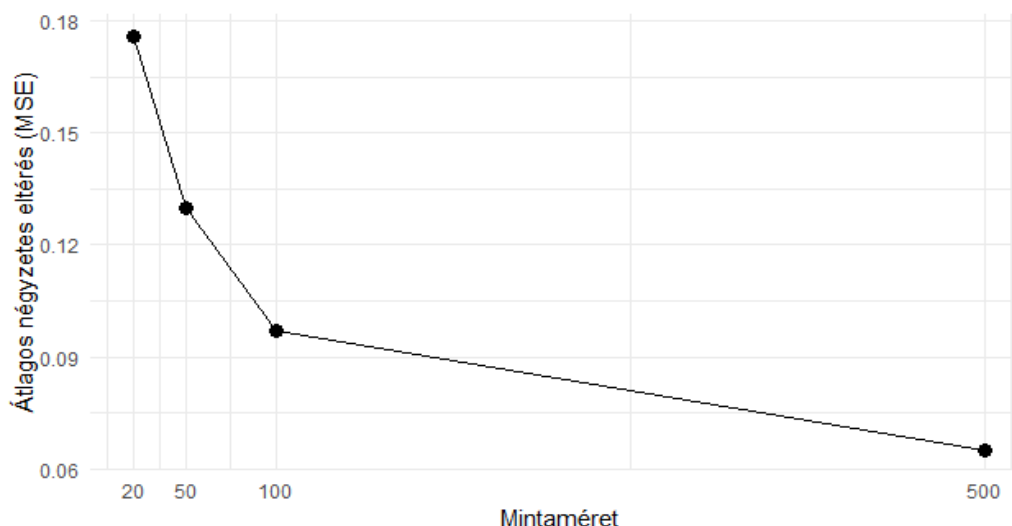
2. táblázat

CML becslési mód eredményeinek átlagos négyzetes eltérése

Mintaméret	Átlagos négyzetes eltérés
500	0,065
100	0,097
50	0,130
20	0,176

3. ábra

CML becslési mód eredményeinek átlagos négyzetes eltérése



A 2. ábra jól mutatja, hogy az egyes feltételek mellett hogyan viselkedik a CML becslési módszer. Az 500 fős mintán adott nehézségbecslések viszonylag kissé térnek el az átlagtól, aránylag stabil becsléseket kapunk ennél a mintaméretnél. Ez a 100 fős mintánál is elmondható, bár egyes itemeknél már látható picit nagyobb eltérés is, mint a nagyobb mintánál. Az 50 és 20 fős mintaméreteknél már nagyobb a variabilitás, kevésbé tűnik megbízhatónak a becslés. Illetve a szélsőséges, nagyon könnyű vagy nehéz itemek esetében is megfigyelhető, hogy a kisebb méreteken tett becslések jobban szóródnak (lásd még: 1. táblázat). Ezek a megállapítások összhangban vannak a 3. ábrán látható átlagos négyzetes eltérés eredményeivel. Látható, hogy a becslés hibájának mérete csökken a mintaméret növekedésével. 20 fő esetében ez 0,176, 50 főnél 0,130, 100 főnél 0,097, 500 főnél 0,065. 20 és 50 fő között nagyobb mértékű a változás, mint 100 és 500 fő között, itt már kisebb a hiba mértékének csökkenése (lásd még: 2. táblázat).

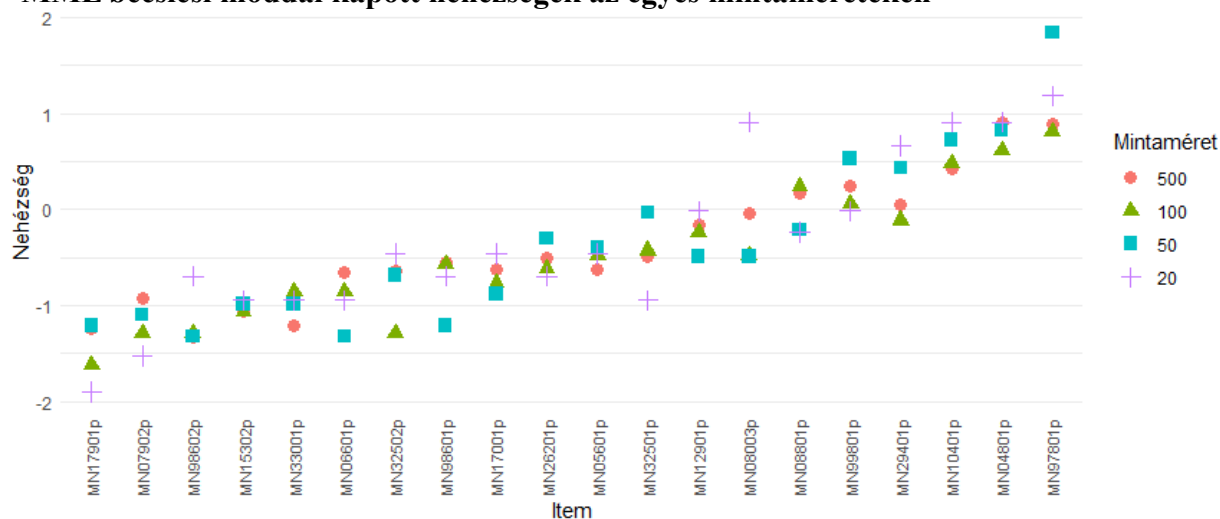
3. táblázat

MML becslési móddal kapott nehézségek az egyes mintaméreteken

Itemek	500 fős minta	100 fős minta	50 fős minta	20 fős minta
MN17901p	-1,239	-1,611	-1,205	-1,894
MN07902p	-0,920	-1,291	-1,093	-1,524
MN98602p	-1,322	-1,291	-1,322	-0,693
MN15302p	-1,059	-1,065	-0,985	-0,942
MN33001p	-1,204	-0,856	-0,985	-0,942
MN06601p	-0,661	-0,856	-1,322	-0,942
MN32502p	-0,641	-1,291	-0,681	-0,460
MN98601p	-0,555	-0,565	-1,205	-0,693
MN17001p	-0,622	-0,757	-0,880	-0,460
MN26201p	-0,508	-0,612	-0,304	-0,693
MN05601p	-0,622	-0,472	-0,396	-0,460
MN32501p	-0,490	-0,426	-0,030	-0,942
MN12901p	-0,160	-0,246	-0,490	-0,014
MN08003p	-0,034	-0,472	-0,490	0,912
MN08801p	0,173	0,245	-0,212	-0,235
MN99801p	0,245	0,066	0,527	-0,014
MN29401p	0,046	-0,112	0,431	0,664
MN10401p	0,420	0,474	0,725	0,912
MN04801p	0,901	0,616	0,827	0,912
MN97801p	0,890	0,813	1,844	1,184

4. ábra

MML becslési móddal kapott nehézségek az egyes mintaméreteken



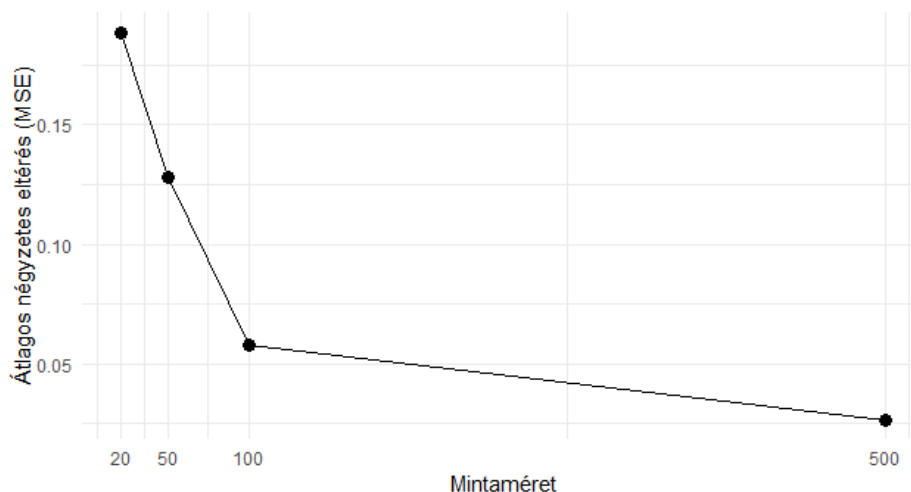
4. táblázat

MML becslési mód eredményeinek átlagos négyzetes eltérése

Mintaméret	Átlagos négyzetes eltérés
500	0,026
100	0,058
50	0,128
20	0,188

5. ábra

MML becslési mód eredményeinek átlagos négyzetes eltérése



Az MML becslési mód esetében is hasonló tendencia figyelhető meg a nehézség becslése és a mintaméret között. A 4. ábrán látszik, hogy a nagyobb mintaméreteken becsült nehézségek kisebb mértékben szóródnak az átlag körül, mint a kisebb méretek esetében (lásd még: 3. táblázat). Ezt szintén igazolják az átlagos négyzetes eltérés eredményei, melyek a méretek növekedésével egyre kisebb hibát mutatnak a becslésben. 20 fő esetén ez 0,188, 50 fő esetében 0,128, 100 főnél 0,058, 500 főnél pedig 0,026. Ahogy a CML-nél, itt is megfigyelhető, hogy a 100-as minta előtt meredekebb a hibák mértékének csökkenése, ami 100 és 500 fő között kisebb már (lásd 4. táblázat és 5. ábra).

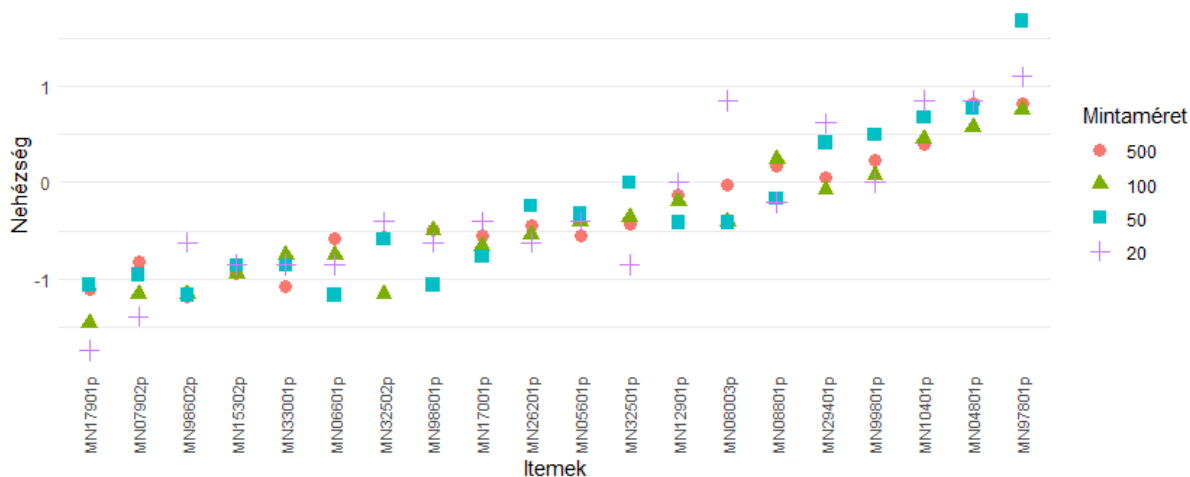
5. táblázat

EM becslési móddal kapott nehézségek az egyes mintaméreteken

Itemek	500 fős minta	100 fős minta	50 fős minta	20 fős minta
MN17901p	-1,106	-1,450	-1,059	-1,735
MN07902p	-0,815	-1,153	-0,956	-1,386
MN98602p	-1,183	-1,153	-1,167	-0,619
MN15302p	-0,941	-0,944	-0,858	-0,847
MN33001p	-1,074	-0,754	-0,858	-0,847
MN06601p	-0,580	-0,754	-1,167	-0,847
MN32502p	-0,562	-1,153	-0,583	-0,405
MN98601p	-0,485	-0,490	-1,059	-0,619
MN17001p	-0,545	-0,663	-0,764	-0,405
MN26201p	-0,442	-0,532	-0,244	-0,619
MN05601p	-0,545	-0,405	-0,327	-0,405
MN32501p	-0,425	-0,364	0,000	-0,847
MN12901p	-0,129	-0,201	-0,411	1e-06
MN08003p	-0,016	-0,405	-0,411	0,847
MN08801p	0,170	0,241	-0,163	-0,201
MN29401p	0,056	-0,080	0,411	0,619
MN99801p	0,235	0,080	0,496	1e-06
MN10401p	0,392	0,447	0,672	0,847
MN04801p	0,825	0,575	0,764	0,847
MN97801p	0,815	0,754	1,677	1,099

6. ábra

EM becslési móddal kapott nehézségek az egyes mintaméreteken



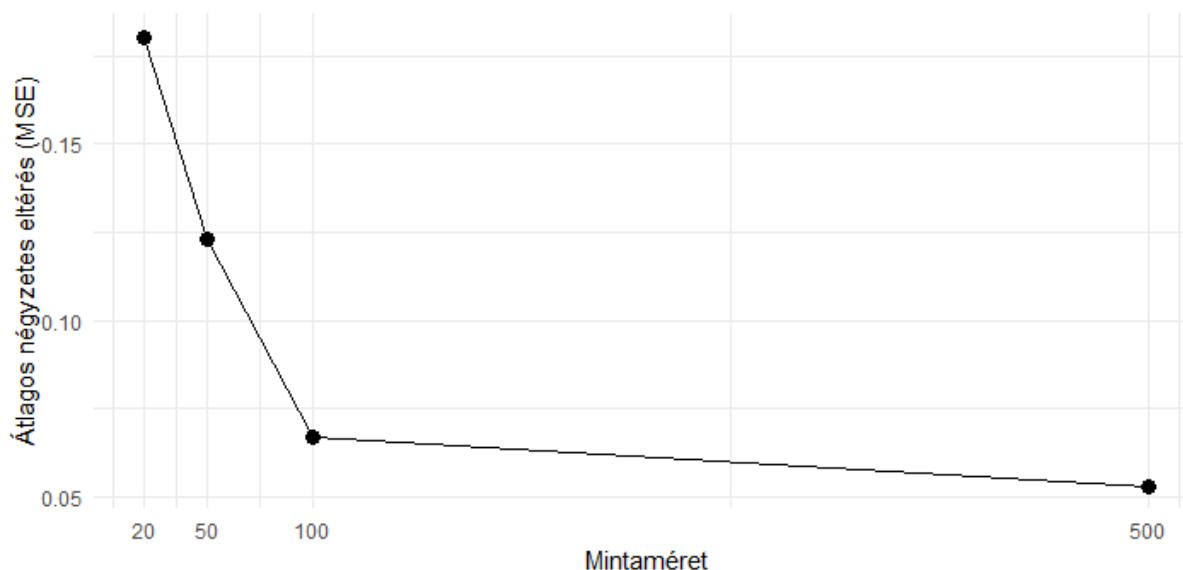
6. táblázat

EM becslési mód eredményeinek átlagos négyzetes eltérése

Mintaméret	Átlagos négyzetes eltérés
500	0,053
100	0,067
50	0,123
20	0,180

7. ábra

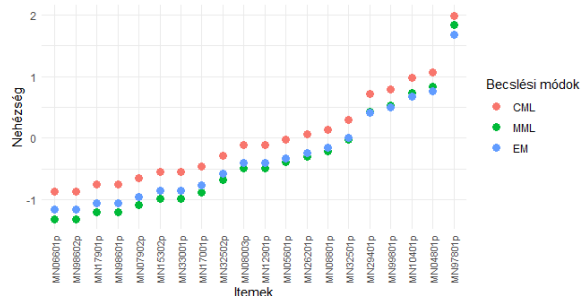
EM becslési mód eredményeinek átlagos négyzetes eltérése



A korábbi becslési módokhoz hasonlóan, az EM becslési módnál is az itemek nehézségének értékénél a kisebb méreteknél nagyobb variabilitás tapasztalható, míg a nagyobbak esetében ez a szóródás kisebb (lásd: 5. táblázat és 6. ábra), melyet az átlagos négyzetes eltérés eredményei is alátámasztanak. 20 főnél 0,180, 50 főnél 0,123, 100 főnél 0,067, 500 fős mintán pedig 0,053 (lásd: 6. táblázat és 7. ábra).

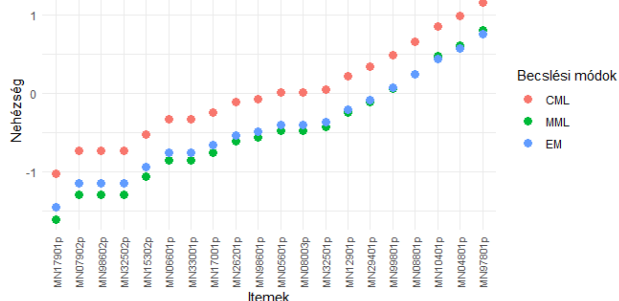
8. ábra

A három becslési mód által számolt item nehézségek 20 fős mintán



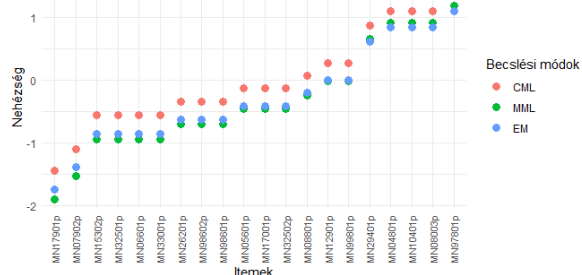
9. ábra

A három becslési mód által számolt item nehézségek 50 fős mintán



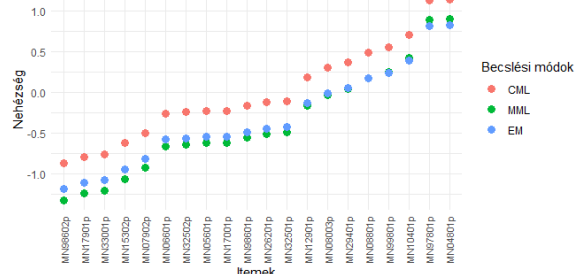
10. ábra

A három becslési mód által számolt item nehézségek 100 fős mintán



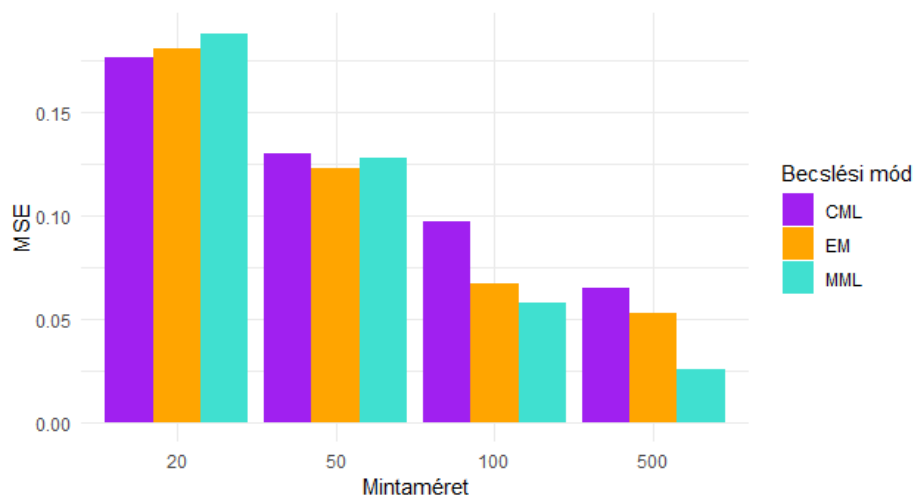
11. ábra

A három becslési mód által számolt item nehézségek 500 fős mintán



12. ábra

Mindhárom becslési mód átlagos négyzetes eltérései az egyes mintaméretekben



A 8.-11. ábrákon látszik, hogy a nehézségek értékei az egyes becslési módoknál ugyanolyan trendet követnek a különböző mintaméreteken, azonban a CML becslés a másik kettőhöz képest magasabb nehézséget ad az egyes itemeknek. A 12. ábrán látható, hogy a 20 fős mintán a CML becslés kisebb hibával dolgozik, a nagyobb mintaméreteken azonban az MML és EM becslések MSE értéke kisebb. A 20 és 50 fős mintaméreteken az EM ezen eredmények alapján jobbnak bizonyul, mint az MML, az 50 fős mintán ezen becslés hibája a legkisebb. A nagyobb mintákon ez megfordul 100 főtől kezdődően már az MML becslési mód bizonyul jobbnak, mint az EM, mely kisebb hibát mutat, mint a CML.

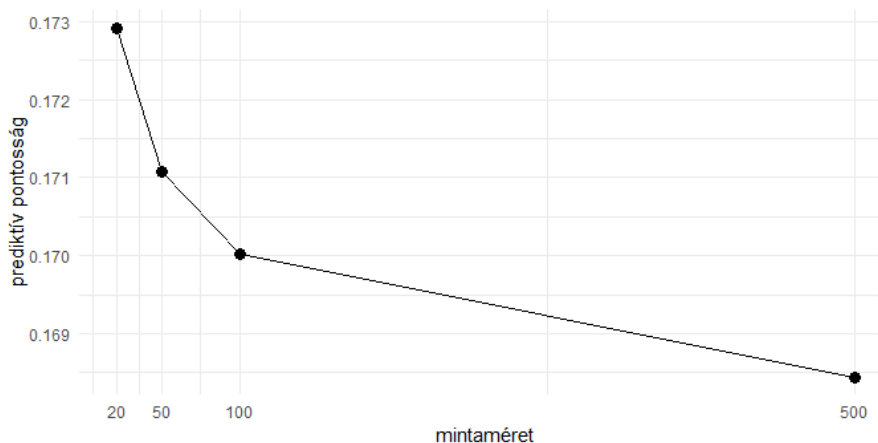
7. táblázat

MML becslési mód prediktív pontossága az egyes mintaméreteken

Mintaméret	Prediktív pontosság
500	0,168
100	0,170
50	0,171
20	0,173

13. ábra

MML becslési mód prediktív pontossága az egyes mintaméreteken



Utolsóként a kapott prediktív pontosság értékeket mutatjuk be, itt csak az MML becslést elemezzük, az EM becsléssel szintén hasonló eredményeket kaptunk. MML becslésnél 20 fő esetében 0,173, 50 fő esetében 0,171, 100 főnél 0,170, 500 fős mintánál pedig 0,168 ez az érték. Itt a legkisebb csökkenés a 100 és az 50 fős minták között van. A két kisebb (20,50), illetve a két nagyobb (100,500) mintaméret között egyforma a pontosság mértékének változása (lásd: 7. táblázat és 13. ábra).

Diszkusszió

Kutatásunk során megvizsgáltuk a Rasch modell CML, MML és EM becsléseinek eredményeit különböző feltételek mellett. Az eredményekből is látszik, hogy nincsen egy tökéletes módszer, amely mindig a legjobb becslést adja, mindegyik becslési módnak vannak hibaértékei, melyek változnak különböző mintaméreteken. A legkisebb 20 fős mintán, ami egy kisebb osztály létszámának felel meg a CML becslési mód a legmegbízhatóbb, azonban a hibák aránya csökken a mintaméret növekedésével. Továbbá az összehasonlító ábrák alapján nehezebbnek becsüli a feladatokat, ami különösen azért lehet gond, mivel a tanuló képességét így alul becsülhetjük, illetve a teszt készítése során is tévesen rangsorolhatjuk be az egyes feladatokat, ami így torzíthatja a teszt eredményét. Emellett mind a három becslési mód kis méreten a szélsőséges itemeket nem megbízhatóan becsüli, ami szintén az előbb említett problémákat veti fel. Az MML és EM becslések hibája nagyobb méreteknél kisebb, mint a CML-nél. A prediktív pontosságot is számításba véve, 100 és 500 fős mintán is az MML becslési móddal lehet a legpontosabb bejósolást adni a vizsgált feltételek között. A feladatok nehézségének becslésekor, mind a négy mintanagyság esetében, a három becslési mód értékei egyszerre változtak, hasonló trendet követtek, ahogy az egyes feladatok nehézségi szintje nőtt, úgy arányosan a módszerek becslései is növekedtek, kisebb eltérésekkel. A kapott eredményeink alapján a kisebb mintákon 20 főig a CML becslés bizonyul a legmegbízhatóbbnak, ennél nagyobb mintaméreteken a másik két becslést ajánlott használni. 50 főnél az EM bizonyult a legkisebb hibával becslőnek, ezután viszont az MML bizonyult jobbnak, mind a 100 mind az 500 fős méreteken, így 100 főtől ezt a becslési módot érdemes használni.

A kutatás előnyéhez tartozik, hogy valós adatokon alapuló szimulációs eljárást használtunk, ami növeli a reliabilitást. Ez azért fontos, mert a legtöbb szimulációs vizsgálatban idealizált adatgeneráló folyamatot feltételezve teljesen mesterséges adatokon tesztelik a feladatokat. A mesterséges helyzeteknél pedig felmerül az eredmények általánosíthatóságának problémája, hiszen a legtöbb esetben feltételezzük, hogy a tesztelt modell illeszkedik a szimulált mintára (Kleijnen, 1995; Sargent, 2010). A legfontosabb mutató, melyet vizsgáltunk a prediktív pontosság, vagyis, hogy mennyire képes a modell előrejelezni a diákok teljesítményét minél kevesebb adat alapján. Míg a prediktív hiba esetén a teljes és a szűrt adatállományokon végzett becsléseket hasonlítottuk össze, itt a szűrt adaton végzett nehézségbecslések alapján becsültük meg a tesztelési adatállományban lévő, vagyis új tanulóknak a képességét. Mivel tekinthetjük

úgy, hogy a diákok képessége állandó, képességeik a valóságban nem változnak különböző tesztelések során, így a modell becslésének pontosságának függvényében az előre jelzés pontosságát kell nézni.

Limitációnak tekinthető viszont, hogy átkódoltuk azokat a feladatokat, amelyikében részpontokat is kaphattak a tanulók, hisz itt nem vesszük azt figyelembe, hogy lehet jó a tanuló gondolatmenete a feladatmegoldás során, meg tudná oldani a feladatot, csak figyelmetlenségből elír valamit. Ezt a Raschnál bonyolultabb modellekkel lehet kezelni (például: Nominal Response Model) (Samejima, 2016). Így azonban ezeket a feladatokat is megoldatlannak tekintettük, mely esetlegesen befolyásolhatja a becslés kimenetelét. Továbbá problémát jelenthet, hogy alacsony elemszám esetén, kevés a lehetséges válaszmintázat, ezzel megnő annak az esélye, különösen nagyon nehéz, vagy könnyű feladatok esetén, hogy ezekre a feladatokra csak 1-es vagy csak 0-s válaszok jelennek meg, így a minta kevésbé reprezentatív lesz, csökkenhet a modell pontossága. Itt megjelenhetnek a becslés során a nagyon kiugró, mintaméret alapján nagy szóródást mutató nehézségbecslések. Nagyobb elemszám esetén többféle válasz, képességszint és nehézségű feladat kerülhet be, mely javíthatja a szélsőséges feladatok bemérésének és az előrejelzésnek a pontosságát.

Egyes szófordulatok és kifejezések magyarra fordításához az alábbi eszközöket használtam: OpenAI. (2024). *ChatGPT* (4.0) [Large language model]. <https://chat.openai.com>
Google AI. (2024). *Gemini 1.5 Flash* [Large language model]. <https://gemini.google.com/app>

Irodalomjegyzék

- Andersen, E. B. (1972). The Numerical Solution of a Set of Conditional Estimation Equations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 34(1), 42–54. <https://doi.org/10.1111/j.2517-6161.1972.tb00887.x>
- Andersen, E. B. (1973). A goodness of fit test for the rasch model. *Psychometrika*, 38(1), 123–140. <https://doi.org/10.1007/BF02291180>
- Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology*, 43(4), 695–716. <https://doi.org/10.1111/j.1744-6570.1990.tb00679.x>
- Auxné Bánfi, I., Balázsi, I., Balkányi, P., Balogh, V. K., Gyapay, J., Lak, Á. R., Ostorics, L. I., Palincsár, I., Rábainé Szabó, A., Rózsa, C., Szabó, Á., Szabó, L. D., Szepesi, I., Szipőcsné Krolopp, J., & Vadász, C. (2017). *Országos kompetenciamérés technikai leírás*. Oktatási Hivatal.
https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/orszmer2012/OKM_Technikaileiras.pdf (2024.08.30.)
- Bowman, M. L. (1989). Testing individual differences in ancient China. *American Psychologist*, 44(3), 576–578. <https://doi.org/10.1037/0003-066X.44.3.576.b>
- Breakspear, S. (2012). *The Policy Impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance*. 71. <https://doi.org/10.1787/5k9fdqffr28-en>
- Brüggemann, T., Ludewig, U., Lorenz, R., & McElvany, N. (2024). Effects of Test Mode and Medium on Elementary School Students' Test Experience. *European Journal of Psychological Assessment*, 40(4), 282–289. <https://doi.org/10.1027/1015-5759/a000767>
- Cai, L., Choi, K., Hansen, M., & Harrell, L. (2016). Item Response Theory. *Annual Review of Statistics and Its Application*, 3(1), 297–321. <https://doi.org/10.1146/annurev-statistics-041715-033702>
- Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6). <https://doi.org/10.18637/jss.v048.i06>

- Darrell Bock, R., & Lieberman, M. (1970). Fitting a response model form dichotomously scored items. *Psychometrika*, 35(2), 179–197. <https://doi.org/10.1007/BF02291262>
- De Leeuw, J., & Verhelst, N. (1986). Maximum Likelihood Estimation in Generalized Rasch Models. *Journal of Educational Statistics*, 11(3), 183–196. <https://doi.org/10.3102/10769986011003183>
- DuBois, P. H. (1970). Early Psychological Testing. In *History of Psychological Testing* (1970. kiad.). Allyn and Bacon Inc.
- Eggen, T. J. H. M., & Verschoor, A. J. (2006). Optimal Testing With Easy or Difficult Items in Computerized Adaptive Testing. *Applied Psychological Measurement*, 30(5), 379–393. <https://doi.org/10.1177/0146621606288890>
- Embretson, S. E., & Reise, S. P. (2013). *Item Response Theory* (1st ed.). Psychology Press. <https://doi.org/10.4324/9781410605269>
- Fischer, G. H., & Parzer, P. (1991). LRSM: Parameter Estimation for the linear Rating Scale Model. *Applied Psychological Measurement*, 15(2), 138–138. <https://doi.org/10.1177/014662169101500203>
- Frey, A., Hartig, J., & Moosbrugger, H. (2009). Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Tests. *Diagnostica*, 55(1), 20–28. <https://doi.org/10.1026/0012-1924.55.1.20>
- Frey, A., Liu, T., Fink, A., & König, C. (2024). Meta-Analysis of the Effects of Computerized Adaptive Testing on the Motivation and Emotion of Examinees. *European Journal of Psychological Assessment*, 1015-5759/a000821. <https://doi.org/10.1027/1015-5759/a000821>
- Gregory, R. J. (2015a). Ability Testing: Group Tests and Controversies. In *Psychological testing: History, principles, and applications* (Seventh edition, global edition, o. 227). Pearson.
- Gregory, R. J. (2015b). *Psychological testing: History, principles, and applications* (Seventh edition, global edition). Pearson.
- Gustafsson, J. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, 33(2), 205–233. <https://doi.org/10.1111/j.2044-8317.1980.tb00609.x>

- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10(4), 255–282. <https://doi.org/10.1007/BF02288892>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer Netherlands. <https://doi.org/10.1007/978-94-017-1988-9>
- Hidegkuti, I., & Balázs, K. (2015). Tesztelmélet. In *Pszichológiai Módszertani Tanulmányok* (o. 65–95). Debreceni Egyetemi Kiadó. <https://oszkdk.oszk.hu/storage/00/02/38/36/dd/1/39642692.pdf> (2024.09.21.)
- Hontangas, P., Ponsoda, V., Olea, J., & Wise, S. L. (2000). The Choice of Item Difficulty in Self-Adapted Testing. *European Journal of Psychological Assessment*, 16(1), 3–12. <https://doi.org/10.1027//1015-5759.16.1.3>
- Kim, J. (1995). *The effects of computerized adaptive testing and examinee test motivation on algebra skills, test anxiety, and attitudes* [Doktori disszertáció, University of Alabama]. ProQuest. <https://www.proquest.com/docview/304160034> (2024.09.08.)
- Kleijnen, J. P. C. (1995). Verification and validation of simulation models. *European Journal of Operational Research*, 82(1), 145–162. [https://doi.org/10.1016/0377-2217\(94\)00016-6](https://doi.org/10.1016/0377-2217(94)00016-6)
- Kubinger, K. D., Steinfield, J., Reif, M., & Yanagida, T. (2012). Biased (conditional) parameter estimation of a Rasch model calibrated item pool administered according to a branched testing design. *Psychological Test and Assessment Modeling*, 54(4), 450–460. https://www.psychologie-aktuell.com/fileadmin/download/ptam/4-2012_20121224/08_Kubinger.pdf (2024.10.04.)
- Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a Computerized Adaptive Test More Motivating Than a Fixed-Item Test? *Applied Psychological Measurement*, 41(7), 495–511. <https://doi.org/10.1177/0146621617707556>
- Liou, L. W., & Price, T. D. (1994). Speciation By Reinforcement Of Premating Isolation. *Evolution*, 48(5), 1451–1459. <https://doi.org/10.1111/j.1558-5646.1994.tb02187.x>
- Liu, C.-W., & Chalmers, R. P. (2018). Fitting item response unfolding models to Likert-scale data using mirt in R. *PLOS ONE*, 13(5), e0196292. <https://doi.org/10.1371/journal.pone.0196292>
- Lord, F. M. (1983). Small N Justifies Rasch Model. In *New horizons in testing: Latent trait test theory and computerized adaptive testing* (o. 51–61). Academic Press.

- Lord, F. M., Novick, M. R., & Birnbaum, A. (2008). The Construction of True and Error Scores. In *Statistical Theories of Mental Test Scores* (Nachdr. der Ausg. Reading, Mass. [u.a.], 1968, o. 37–38). Information Age Publ. <https://doi.org/10.1016/B978-0-12-742780-5.50011-1>
- Mair, P., & Hatzinger, R. (2007). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science*, 49(1), 26–43. https://www.psychologie-aktuell.com/fileadmin/download/PsychologyScience/ps_1_2007_26-43.pdf (2024.10.04.)
- Mair, P., Rusch, T., Hatzinger, R., Maier, M. J., & Debelak, R. (2024). Extended Rasch Modeling. *The Comprehensive R Archive Network*. <https://cran.r-project.org/web/packages/eRm/eRm.pdf> (2024.10.12.)
- Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology*, 110(1), 27–45. <https://doi.org/10.1037/edu0000205>
- McCall, W. A. (1939). *Measurement*. Macmillan.
- Meloun, M., & Militký, J. (2011). Statistical analysis of multivariate data. In *Statistical Data Analysis* (o. 151–403). Elsevier. <https://doi.org/10.1533/9780857097200.151>
- Nagybányai Nagy, O. (2006). A pszichológiai tesztek reliabilitása. In *A pszichológiai mérés alapjai: Elmélet, módszer és gyakorlati alkalmazás* (o. 105). <https://mek.oszk.hu/05500/05536/>
- OECD. (2012). *PISA 2009 Technical Report*. OECD. <https://doi.org/10.1787/9789264167872-en>
- Ortner, T. M., Weißkopf, E., & Koch, T. (2014). I Will Probably Fail: Higher Ability Students' Motivational Experiences During Adaptive Achievement Testing. *European Journal of Psychological Assessment*, 30(1), 48–56. <https://doi.org/10.1027/1015-5759/a000168>
- Petra, T. Z. H. T., & Aziz, M. J. A. (2020). Investigating reliability and validity of student performance assessment in Higher Education using Rasch Model. *Journal of Physics: Conference Series*, 1529(4), 042088. <https://doi.org/10.1088/1742-6596/1529/4/042088>
- R Core Team. (2021). *R: A language and environment for statistical computing* [Software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rabe-Hesketh, S., & Skrondal, A. (2016). Generalized Linear Latent and Mixed Modeling. In *Handbook of item response theory*. CRC Press. <https://doi.org/10.1201/9781315374512>

- Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.
- Rezvani, S., & Wang, X. (2023). A broad review on class imbalance learning techniques. *Applied Soft Computing*, 143, 110415. <https://doi.org/10.1016/j.asoc.2023.110415>
- Robeson, S. M., & Willmott, C. J. (2023). Decomposition of the mean absolute error (MAE) into systematic and unsystematic components. *PLOS ONE*, 18(2), e0279774. <https://doi.org/10.1371/journal.pone.0279774>
- Robitzsch, A. (2021). A Comprehensive Simulation Study of Estimation Methods for the Rasch Model. *Stats*, 4(4), 814–836. <https://doi.org/10.3390/stats4040048>
- Rózsa, S., Nagybányai Nagy, O., & Oláh, A. (2016). A pszichológiai mérés történeti gyökerei és a mérés problémája a pszichológiában. In *A pszichológiai mérés alapjai: Elmélet, módszer és gyakorlati alkalmazás*. <https://mek.oszk.hu/05500/05536/>
- Samejima, F. (2016). Graded Response Models. In *Handbook of item response theory* (o. 95–107). CRC Press. <https://doi.org/10.1201/9781315374512>
- Sargent, R. G. (2010). Verification and validation of simulation models. *Proceedings of the 2010 Winter Simulation Conference*, 166–183. <https://doi.org/10.1109/WSC.2010.5679166>
- Szabó, L. D., Szepesi, I., Takácsné Kárász, J., & Vadász, C. (2018). *Országos kompetenciamérés 2017 Országos jelentés* [Országos jelentés]. Oktatási Hivatal. https://okm.kir.hu/fit2/pdf/OKM_2017_Orszagos_jelentes.pdf (2024.08.05.)
- Thissen, D., & Mislevy, R. J. (2000). Testing Algorithms. In *Computerized adaptive testing: A primer* (2nd ed, o. 101–133). Lawrence Erlbaum Associates.
- Thompson, N. (2024). What is Item Response Theory (IRT)? *Assessment Systems*. <https://assess.com/what-is-item-response-theory/>
- Thorndike, E. L. (1918). Individual differences. *Psychological Bulletin*, 15(5), 148–159. <https://doi.org/10.1037/h0070314>
- Tompa, K. (2001). A matematika érettségi eredményeinek elemzése. *Iskolakultúra*, 11(9), 108–114. https://www.epa.hu/00000/00011/00052/pdf/iskolakultura_EPA00011_2001_09_108-114.pdf (2024.09.06)

- Tóth, E. (2015). Az országos kompetenciamérés hatása a tanítási munkára pedagógusinterjúk alapján. *Magyar Pedagógia*, 115(2), 115–138. <https://doi.org/10.17670/MPed.2015.2.115>
- Ul Hassan, M., & Miller, F. (2019). Optimal Item Calibration for Computerized Achievement Tests. *Psychometrika*, 84(4), 1101–1128. <https://doi.org/10.1007/s11336-019-09673-6>
- van der Linden, W. J. (Szerk.). (2016). *Handbook of item response theory*. CRC Press. <https://doi.org/10.1201/9781315374512>
- van der Linden, W. J., & Hambleton, R. K. (Szerk.). (1997). *Handbook of Modern Item Response Theory*. Springer New York. <https://doi.org/10.1007/978-1-4757-2691-6>
- von Davier, M. (2016). Rasch Model. In *Handbook of item response theory* (o. 31–50). CRC Press. <https://doi.org/10.1201/9781315374512>
- Wainer, H. (2000). Introduction and History. In *Computerized Adaptive Testing: A Primer* (2nd ed). Lawrence Erlbaum Associates.
- Wainer, H., & Dorans, N. J. (2000). *Computerized adaptive testing: A primer* (2nd ed). Lawrence Erlbaum Associates.
- Wainer, H., & Eignor, D. (2000). Caveats, Pitfalls, and Unexpected Consequences of Implementing Large-Scale Computerized Testing. In *Computerized adaptive testing: A primer* (2nd ed., o. 271–297). Lawrence Erlbaum Associates 200.
- Wainer, H., & Mislevy, R. J. (2000). Item Response Theory, Item Calibration, and Proficiency Estimation. In *Computerized adaptive testing: A primer* (2nd ed). Lawrence Erlbaum Associates.
- Weiss, D. (2021). The Rasch model. *Assessment Systems*. <https://assess.com/the-rasch-model/>
- 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>
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Mesa Pr. <https://research.acer.edu.au/measurement/2/> (2024.10.04.)
- Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*, 29(1), 18. <https://doi.org/10.1186/s41155-016-0040-x>

Zheng, Y. (2014). *New methods of online calibration for item bank replenishment* [Doktori disszertáció, University of Illinois]. <http://hdl.handle.net/2142/50699> (2024.09.15.)

Zwinderman, A. H. (1991). *Studies of estimating and testing Rasch models*. Nijmegen Institute for Cognition Research and Information Technology. <https://www.semanticscholar.org/paper/Studies-of-estimating-and-testing-Rasch-models-Zwinderman/ac0a317fcc2854318367042bcd6d7190af5b1fac> (2024.10.04.)