



TRENDSKÁLÁZÁSI MÓDSZEREK: A KORAI ÉS KÖZELMÚLTBELI IEA MATEMATIKAI ÉS TERMÉSZETTUDOMÁNYOS FELMÉRÉSEK ÖSSZEKAPCSOLÁSA

Majoros Erika

Umeå University, Svédország

A Tanulói Teljesítmények Vizsgálatának Nemzetközi Társasága (International Association for the Evaluation of Educational Achievement, IEA) által szervezett nemzetközi nagymintás felmérések egyike a Trends in International Mathematics and Science Study (TIMSS), amely 1995-től mér matematikai és természettudományos tudást negyedik és nyolcadik osztályban. A TIMSS-t négy IEA felmérés előzte meg, amelyek eredményit nem kapcsolták össze a TIMSS skálával. Jelen tanulmány lehetséges trendskálázási módszereket ismertet, és a megvalósult összekapcsolás illusztrálására a magyar diákok újra skálázott matematikai teljesítmény és motiváció, valamint a természettudományos teljesítmény eredményeit mutatja be.

Elméleti keret

Az IEA 1995 óta végez rendszerszintű felméréseket a matematikai és természettudományi tanulói teljesítmények vizsgálatára, mely felmérések eredményeit trendskálákon közli. Ezt megelőzően az IEA négy nemzetközi nagymintás felmérést végzett ezekben a tantárgyakban, azonban ezeknek a korai felméréseknek eredményeit nem kapcsolta össze az 1995-ban alapított trendskálával. Magyarország az összes korai felmérésben részt vett az IEA-hoz történt 1968-as csatlakozása után (B. Németh et al., 2012; Csíkos & Vidákovich, 2012).

Az 1995 előtt végzett IEA felméréseket első szakaszbeli, az azt követőket pedig második szakaszbeli felméréseknek nevezzük Gustafsson (2008) nyomán. Az első szakaszbeli rendszerszintű felméréseket az 1. táblázat tartalmazza. A két szakasz közötti összekapcsolás hiányát a korai felmérések közötti, a mérési eszközökben, a vizsgált populációkban és a mérési eljárásokban bekövetkezett változtatások indokolták (Martin & Kelly, 1996a). Az akkori technológiai és módszertani lehetőségek is korlátozhatták az összekapcsolás megvalósíthatóságát. Az első felmérés óta eltelt évtizedek során új technikai lehetőségek jelentek meg, így változtatások történtek például a tanulók és a tesztfeladatok mintavételével vagy az attitűdöket vizsgáló kérdések megfogalmazásával kapcsolatban. Ezek a döntések kihívások elé állítják az összehasonlíthatóságot, következésképpen az eredmények összekapcsolását. A modern technikai és módszertani fejlesztések azonban lehetővé teszik az ilyen kihívások kezelését.

1. táblázat

Az IEA által az első szakaszban szervezett nemzetközi matematikai és természettudományos felmérések

Felmérés	Adatgyűjtés ideje	Részt vevő oktatási rendszerek száma
Első Nemzetközi Matematikai Vizsgálat (First International Mathematics Study; FIMS)	1964	12
Első Nemzetközi Természettudományi Vizsgálat (First International Science Study; FISS)	1970–71	17
Második Nemzetközi Matematikai Vizsgálat (Second International Mathematics Study; SIMS)	1980–82	20
Második Nemzetközi Természettudományi Vizsgálat (Second International Science Study; SISS)	1983–84	24

Korábbi kutatások igazolják, hogy a két szakaszbeli IEA felmérések eredményeit össze lehet kapcsolni, különféle módszertani megközelítésekkel. Az egyik megközelítés az, hogy olyan tesztek kapcsolnak össze, amelyek megismételt feladatokat tartalmaznak, modern tesztelméleti (item response theory, IRT) modellek alkalmazásával. Afrassa (2005), Majoros és munkatársai (2021), valamint Strietholt és Rosén (2016) ezzel a megközelítéssel kapcsoltak össze matematika és olvasási teljesítményt vizsgáló felméréseket. Azonban ezek a matematikai összekapcsolási tanulmányok (Afrassa, 2005; Majoros et al., 2021) korlátozottak a TIMSS-skálával való összehasonlíthatóság és a bevont országok számának szempontjából. A nyolcadik évfolyamos matematika és természettudományos trendskála minden résztvevő országra kiterjedő kibővítése megvalósult Majoros (2023) munkájában. A negyedik évfolyamos természettudományos trendskálázás megvalósult Sosa Paredes és Andersson (2025) tanulmányában.

Egy másik lehetséges megközelítés különböző regionális, országos vagy nemzetközi teszt-eredmények összekapcsolására az, amelyben a tesztek nem tartalmaznak megismételt feladatokat, így a klasszikus tesztelméletre (classical test theory, CTT) támaszkodik. Ebben a megközelítésben az összekapcsolás szigorúbb feltételekre épül a képesséeloszlásra vonatkozóan (ld. pl. Chmielewski, 2019; Hanushek & Woessmann, 2012).

Az matematika- és természettudományos tanulás affektív aspektusainak trendmérése az IEA által a TIMSS 2011-es adminisztrációjával kezdődött. A 2011-es, a 2015-ös és a 2019-es háttérkérdőívek tartalmaznak megismételt kérdéseket, és eredményeiket trendskálákon jelentik, a kognitív eredményekhez hasonlóan (Martin et al., 2012, 2016; Yin & Fishbein, 2020). Ezen affektív skálák longitudinális kiterjesztését vizsgálja Majoros és munkatársai (2022) tanulmánya.

Összegezve megállapítható, hogy a tesztelméleti módszertani fejlődés és a nemzetközi nagymintás felmérések jelentőségének növekedése miatt érdemes feltárni a korai adatokban rejlő lehetőségeket. Jelen tanulmány kétféle módon járul hozzá a szakirodalomhoz. Egyrészt, a tárgyalt összekapcsolási technikák alkalmazhatók más olyan felmérésekre is, amelyekben az adminisztrációk között változások történtek. Másrészt, a bemutatott skálák felhasználhatók jövőbeni longitudinális tanulmányokhoz.

Rendszerszintű oktatási eredmények mérése

A mérés számok hozzárendelése tárgyak vagy események aspektusaihoz egy vagy másik szabály vagy egyezmény szerint (Stevens, 1968). Az oktatási mérés olyan mérést jelent, amely oktatási környezetben történik. Az oktatási rendszerek összetett és többszintű struktúrák, melyek tervezéséhez és irányításához az oktatási mérések szolgálnak visszajelzésül (Csíkos et al., 2020). Rendszerszinten a visszajelzést nemzetközi és/vagy nemzeti értékelések biztosítják. A nemzetközi és nagymintás felmérések a viszonylag hosszú visszajelzési idő miatt előnyösek lehetnek a közép- vagy hosszú távú tervezéshez. Ezen felmérésekben való részvétel különösen előnyös lehet azoknak az országoknak, amelyek nem tartanak fenn nemzeti értékeléseket.

Az IEA felmérésekben a méréseket teljesítménytesztek és háttérkérdőívek segítségével végzik. Ezen mérések a pszichometriában, azaz a pszichológiai tesztek tulajdonságainak értékelésével foglalkozó tudományában kifejlesztett elméleteket és módszertanokat alkalmazzák (Furr & Bacharach, 2014). Cronbach (1960) a pszichológiai tesztet két vagy több ember viselkedésének összehasonlítására szolgáló szisztematikus eljárásként határozta meg (p. 21). A pszichológiai tesztek fejlesztésének alapelvei az elmúlt évtizedekben drámaian megváltoztak (Embretson & Reise, 2000).

A valószínűséginek is nevezett modern tesztelmélet a pszichológiai értékelések fő elméleti alapjává vált, felváltva a CTT-ként emlegetett elveket. A CTT-keretrendszerben a módszereknek és a mérési eljárásoknak számos hiányosságuk van (Hambleton et al., 1991). A legfontosabb korlát az, hogy a tesztmegoldó jellemzői, mint például a képesség, valamint a teszttulajdonságok, mint például a nehézség, csak egymással összefüggésben értelmezhetők. Ezért a CTT keretében az eredmények tesztfüggőek, ami negatív következményekkel jár a tesztteredmények összehasonlíthatóságára nézve. Ez és a CTT egyéb korlátai arra késztették a pszichometrikusokat, hogy alternatív elméleteket és modelleket dolgozzanak ki a pszichológiai mérésre. A következőkben a látens, azaz nem közvetlenül megfigyelhető pszichológiai konstrukciók, mint például a matematikai képesség mérésére szolgáló két fő keretrendszert vázoljuk: a valószínűségi tesztelméletek és az ellenőrző faktorelemzés (confirmatory factor analysis, CFA) keretrendszereit.

A modern tesztelmélet alkalmazása

Az IRT egy modellalapú mérési keretrendszer (Embretson & Reise, 2000). Az IRT két alapvető gondolata, hogy a tesztmegoldó teljesítménye egy adott itemen látens tulajdonságok vagy képességek alapján előre jelezhető, és hogy az itemteljesítmény és a látens tulajdonság közötti kapcsolat kifejezhető az item karakterisztikus görbéje segítségével (ld. pl. Embretson & Reise, 2000; Hambleton et al., 1991; Molnár, 2006).

Hambleton és Jones (1993) rámutattak az IRT tesztelésre való alkalmazásának négy előnyére a klasszikus tesztelméleti megközelítésekkel szemben. Először is, az itemstatisztikák függetlenek attól a mintától, amelyből becsülték őket. Másodszor, a jártassági pontszámok függetlenek a teszt nehézségétől. Harmadszor, az IRT-modellek alapot biztosítanak a tesztitemek képességszintekhez való illesztéséhez. Végül, az IRT-modellek nem igényelnek szigorúan párhuzamos teszteket a megbízhatóság értékeléséhez. Az IRT további előnye, hogy lehetővé teszi az itemek és a tesztmegoldók mátrix mintavételét, amely mintavételi módszert gyakran használnak nagymintás felmérésekben.

A FIMS és a FISS kivételével az itt tárgyalt teljesítménytesztekben ún. mátrix mintavételi megközelítést alkalmaztak. Ez azt jelenti, hogy a tesztek összesen több itemet tartalmaznak, mint amennyit egyes tesztmegoldók kézhez kapnak. Lényegében a mátrix mintavétel során az

itemek mintáit diákok mintáin adminisztrálják (Mazzeo et al., 2006). Ezt a kettős mátrix mintavételt két ok motiválja. Először, a mért képességek és ismeretek terjedelme annak megfelelően nagy itemkészletet igényel. Másodsor, a diákok természettudományos és matematikai ismereteinek, valamint alkalmazásának tesztelése időigényes itemtípusokat igényel. A mátrix mintavétel előnye, hogy csökkenti az egyes diákok tesztelési terheit. Továbbá, a mátrix mintavételezéssel növelhető a teszt érvényessége, mivel a diákok nagyobb valószínűséggel töltik ki a tesztet a megadott időn belül. Összekapcsolási módszerek alkalmazásával összehasonlítható eredmények érhetők el különböző tesztformák használatával.

Ellenőrző faktorelemzés alkalmazása

A látens változók modellezésének egy másik megközelítése az ellenőrző faktoranalízis (Jöreskog, 1969). A CFA egy statisztikai technika a strukturális egyenletek modellezése (structural equation modeling, SEM) keretrendszerében. A CFA az indikátorként szolgáló megfigyelt változók és a látens változók, azaz faktorok közötti kapcsolatot vizsgálja. A faktor egy mögöttes konstrukció, amelyről feltételezik, hogy befolyásolja az indikátorokat, és figyelembe veszi a közöttük lévő korrelációkat. A kapcsolatok mintázatát faktorsúlyok fejezik ki. A faktorpontszámok faktorsúlyok segítségével számíthatók ki. A faktorpontszámok a tesztelők manifest pontszámát jelentik a látens dimenzió (Brown, 2015).

A CFA gyakran használt eszköz a skálafejlesztés során. Ezt a módszert gyakran alkalmazzák például a mérőeszközök pszichometriai tulajdonságainak értékelésére, konstrukciók validálására, a mérési invariancia csoportok közötti tesztelésére vagy a módszerhatások vizsgálatára (Brown, 2015).

A látens konstrukciókat felmérő kérdőíves adatok CFA módszerekkel skálázhatók. Az affektív itemek skálázására szolgáló CFA-modellezés alkalmazásának egyik példája az OECD által 2008 óta végzett Teaching and Learning International Survey (TALIS). Ezenkívül mérési invariancia tesztelést alkalmaztak többcsoportos ellenőrző faktoranalízissel (multiple-group confirmatory factor analysis, MGCFA) a résztvevő oktatási rendszerek összehasonlíthatóságának értékelésére (OECD, 2019).

Skálázás

Stevens (1968) négy típusba sorolja a skálákat: nominális (pl. sorrend nélküli kategóriák, mint a futballisták számozása), ordinális (pl. korlátozott számú kategória a Likert-skála szerinti sorrenddel, mint például az ásványok keménységi skálája), intervallum (folytonos és egyenlő távolságra lévő számok valódi nulla nélküli sora, mint pl. a Celsius- és Fahrenheit-fokok, naptári idő), és arányskálák (folytonos, egyenlő távolságra lévő számok valódi nullát tartalmazó sora, mint pl. a hosszúság). Kolen (2006) az oktatási mérési keretrendszerben a skálázást úgy határozta meg, mint számoknak vagy más rendezett mutatóknak egy oktatási teszten elért teljesítményhez történő hozzárendelése. A tesztitemeket jellemzően helyesnek, részben helyesnek vagy helytelennek pontozzák, és a nyers pontszám a helyes válaszok számán alapul.

Különböző típusú transzformációk léteznek a skálapontszámok nyers pontszámokból történő előállítására. A nemzetközi nagymintás felmérésekben a képességskálák előállítására az IRT-alapú skálázás a jellemző. Az IRT képességskálázási konvenciói olyan skálapontszámokat eredményeznek, amelyek átlaga 0 és szórása 1, és amelyek lineáris transzformációt igényelnek az értelmezéshez (Kolen, 2006). Például a TIMSS nemzetközi teljesítményskáláit az első adminisztrációkor 500 átlaggal és 100 szórással állították be (Martin & Kelly, 1996b).

A skálázás során a valószínűsíthető érték (plausible value, PV) elnevezésű módszertant (ld. pl. Beaton & Johnson, 1992; Mislevy, 1991; Mislevy et al., 1992; Rutkowski et al., 2010) alkalmazzák a populáció szintű képességbecslések generálására egy mátrix mintavételezésű tesztervből. A PV módszertan szerint a tanulói teljesítményt hiányzó értéként kezelik. Minden tanuló válaszol a tesztelemelek egy részhalmazára, és ezeket a válaszokat használják fel a tanulói képességeloszlás generálására a populációban. A populáció képességeloszlásának becsléséhez Rubin (1987) többszörös imputációs megközelítésének mérési modellen alapuló kiterjesztését alkalmazzák. A becsült képességeloszlásból véletlenszerű PV értékeket rendelnek minden tesztmegoldóhoz. A modern nemzetközi nagymintás felmérésekben alkalmazott PV-módszertan eltér az egyéni jártassági pontszámokra összpontosító más megközelítésektől (L. Rutkowski et al., 2010; Mazzeo et al., 2006). A PV-megközelítés lehetővé teszi a nagy itembankok használatát a pontosabb jártassági becslés érdekében, de a mátrixmintavétel miatt a jelentett pontszámokat csoportszinten kell kezelni. Az ilyen típusú értékeléseket csoportos pontszámértékeléseknek nevezik (Mazzeo et al., 2006). A tanulmányi eredmények mérése nemzetközi kontextusban sajátos kihívásokkal jár. A következő szakaszban ezeket a kihívásokat vázoljuk fel.

Mérési invariancia és itemtorzítás

A tanulói eredmények több oktatási rendszert magában foglaló és összehasonlító értékelésénél figyelembe kell venni a kulturális különbségeket, például a mérendő konstrukciók jelentését illetően. A mérési invariancia vagy ekvivalencia egy mérés azon tulajdonságára utal, amely arra vonatkozik, hogy a pontszámok ugyanazt a jelentést hordozzák-e különböző körülmények között, pl. az értékelés időpontja, az adminisztrációs módszerek, országok vagy populációk között (Kline, 2016; Meade & Lautenschlager, 2004). Ahogy Kline (2016) rámutat, az irodalomban számos címke létezik ugyanarra az invarianciatípusra. A legtöbb terminológia a Meredith (1993) által megállapított invarianciaszintek hierarchiájából származik. Ezek az invarianciaszintek a paraméterekre vonatkozó csoportok közötti korlátozások hierarchiáját képviselik (Wu & Estabrook, 2016).

A faktoriális invariancia hierarchiája négy szintből áll: konfigurációs (mintázati), gyenge (metrikus), erős (skaláris) és szigorú (reziduális) invariancia (Kline, 2016; Meredith, 1993; Meredith & Teresi, 2006; Putnick & Bornstein, 2016). A konfigurációs invariancia a legkevésbé korlátozó szint, és azt jelenti, hogy a konstrukciók alapvető szerveződése, azaz a faktortöltések mintázata azonos a csoportok, pl. országok között. A gyenge faktoriális invariancia a konfigurációs invarianciára épít, és azt jelenti, hogy a faktortöltések egyenértékűek a csoportok között. Az erős faktoriális invariancia azt jelenti, hogy a faktorátlagok egyenértékűek a csoportok között. A hierarchia tetején található szigorú faktoriális invariancia azt jelenti, hogy a reziduális értékek varianciái egyenértékűek a csoportok között.

A mérési invariancia, vagyis az itemtorzítás hiányának fogalmát eredetileg az IRT kontextusában fogalmazták meg (Meredith & Teresi, 2006). A mérési invariancia – a szigorú terminológia szerint a faktoriális invariancia – és az itemtorzítás hiánya közötti kapcsolatról elmondható, hogy a két fogalom különálló, de szorosan összefügg egymással (Meredith & Teresi, 2006). Az itemtorzítás hiányát Meredith és Teresi (2006) az itemparaméterek csoportok közötti invarianciájaként definiálja. Amikor ez az invariancia nem teljesül, különböző itemműködés (differential item functioning, DIF) következik be.

Skálák összekapcsolása

Az előző szakaszok a nemzetközi nagymintás értékelések keresztmetszeti aspektusával foglalkoztak. A második fázisú IEA felmérések egyik fő célja a kognitív eredmények rendszerszerű trendmérésének fenntartása, és újabban bizonyos háttérváltozók trendskálázása is. Ebből a célból a felméréseket összekapcsolják annak érdekében, hogy az eredmények ugyanazon a trendskálán legyenek értelmezhetőek. Ebben a szakaszban a keresztmetszeti mérési pontok összekapcsolásának keretrendszerét vázoljuk fel.

Az nemzetközi nagymintás felmérésekben a trendskálák felépítése két vagy több teszt összekapcsolásán alapul. Az összekapcsolás az a folyamat, amelynek során a tartalmukban és/vagy nehézségükben eltérő tesztek eredményeit statisztikai módszerekkel összemérhetővé teszik (Holland & Dorans, 2006; Linn, 1993; Mislevy, 1992). Az összekapcsolás legerősebb változata a kiegyenlítés (equating). Mazzeo és von Davier (2013) a skálák összekapcsolásának folyamatát úgy definiálják, mint felmérések sorozata által előállított eredmények olyan skálájának elérését, amely hosszútávon stabil és összehasonlítható jelentést tart fenn.

Két vagy több értékelés összekapcsolása az IRT-keretrendszerben magában foglalja a különböző tesztadminisztrációkból származó itemparaméter-bebecslések egy közös skálára helyezését. A Hambleton és munkatársai (1991) által meghatározott négy összekapcsolási design az egycsoportos, az ekvivalens csoportokon alapuló, a közös személyeken alapuló és a horgonyteszt módszertan.

Egycsoportos design esetén ugyanaz a hallgatói csoport tölti ki az összekapcsolandó teszteket. Ez a design nem feltétlenül praktikus, mert a tesztelési idő túl hosszú az egyének számára, ezért a fáradtság befolyásolhatja az eredményeket. Ezenkívül a több teszt kitöltése által okozott gyakorlási hatás akadályozhatja a paraméterbecslés pontosságát.

Az ekvivalens csoportokon alapuló design alkalmazása során véletlenszerűen kiválasztott és egyenértékű hallgatói csoportokkal veszik fel az összekapcsolandó teszteket. Ez a design praktikusabb, mint az első, mert elkerüli a fáradtság és a gyakorlás hatásait.

A közös személyeken alapuló design azt jelenti, hogy az összekapcsolandó teszteket megoldó csoportokban van egy közös tesztmegoldó halmaz. Ez a design hasonló kihívásokat jelent a közös csoport számára, mint az egycsoportos terv, ezért kevésbé tekinthető praktikusnak.

A horgonyteszt design megoldja a fent említett hiányosságokat. A hallgatók különböző csoportjai töltik ki az összekapcsolandó teszteket. Minden teszt tartalmaz egy csoport közös itemet, amelyek összességét horgonytesztnak nevezzük. Ez a design nagyon népszerű a megvalósíthatósága miatt.

Az IEA a teljesítményteszteket különböző időpontokban (a TIMSS esetén négyévente) veszi fel, és az egymást követő adminisztrációk tartalmaznak megismételt tesztitemeket. Ezen horgonytesztet alkalmazó designon belül számos lehetséges paraméterbecslési módszer alkalmazható az értékelések összekapcsolására.

A különböző tesztekben származó válaszadatokon alapuló paraméterbecslés elvégezhető együtt vagy külön-külön. Az az eljárás, amelynek során a különböző tesztformákból származó válaszadatokat összevonják a paraméterbecsléshez, egyidejű kalibrálásnak nevezzük. Ez a módszer kisebb standard hibákat biztosít, kevesebb feltételezést tartalmaz, mint más IRT-eljárások, és jó összekapcsolás érhető el kevés közös itemmel (Wingersky & Lord, 1984).

Külön paraméterbecslés esetén a tesztek összekapcsolásának egyik módja, hogy a horgonytétel paramétereit az egyik teszten becsült értékekhez rögzítik, amikor a többi teszt tételét kalibrálják. Ezt a folyamatot Kolen és Brennan (2014) fixparaméteres kalibrációnak nevezi. Alternatív megoldásként a paraméterek külön-külön is becsülhetők a horgonytesztre, és skálatranszformáció alkalmazható úgy, hogy a közös tételparaméterek eloszlása megegyezzen.

A Dorans és munkatársai (2011) által leírt legjobb kiegyenlítési gyakorlatok néhány aspektusa a skálák összekapcsolására is vonatkozik. Először is, mivel a gyűjtött adatok mennyisége jelentős hatással van az eredmények hasznosságára, a nagy reprezentatív minták kedvezőek. Az ilyen adatok biztosítják, hogy a statisztikai bizonytalanság (standard hiba) sokkal kisebb legyen, mint az eredmények egyéb variációs forrásai. Másodszor, a horgonyitemek viselkedését a különböző adminisztrációkban értékelni kell. A feltételezés az, hogy az itemek viselkedése hasonló az azonos évfolyamú diákcsoportok között, és ezt DIF-módszerekkel tesztelik. Végül, a horgonyteszten elért teljesítménynek szorosan korrelálnia kell a teljes teszteredménnyel.

Skálák összekapcsolása a TIMSS felmérésekben

A TIMSS-felmérésekben alkalmazott skálázási és összekapcsolási eljárásokat az IEA által készített különféle technikai jelentések részletesen ismertetik (ld. pl. Martin et al., 2020; Martin & Kelly, 1996b). Jelen tanulmány kétféle oktatási eredményre összpontosít. Először a kognitív eredményekre, azaz a tanulók nyolcadik évfolyamos matematika és természettudományos teljesítményére. Másodszor affektív eredményekre, pontosabban, hogy a diákok mennyire motiváltak a matematika tanulására. Az alábbiakban a mért konstrukció típusa alapján foglaljuk össze az eljárásokat.

Kognitív teljesítmény: matematika és természettudomány

A skálázási eljárás három fő lépésből áll: itemparaméter-becslés, személypontosítás és populációmodellezés. Az első lépést IRT egyidejű kalibrációval végzik, azaz az egyes TIMSS-ciklusok itemparamétereit az aktuális és az előző értékelés adatai alapján becsülik meg. Az első lépésben a háromparaméteres logisztikus modellt (3PL) alkalmazzák a dichotóm feleletválasztós itemekhez, a kétparaméteres logisztikus modellt (2PL) a dichotóm szabad választású itemekhez, és az általánosított parciális kreditmodellt (generalized partial credit model, GPCM) a politóm szabad választású itemekhez, azaz azokhoz, amelyek maximális pontszáma egynél több. A korai ciklusokban a skálázást olyan mintákon végezték, amelyek azonos méretű, véletlenszerűen kiválasztott almintákból álltak a részt vevő országok mindegyikéből (ld. pl. Adams et al., 1996). Újabban a teljes mintákat használják és úgy súlyozzák, hogy minden ország egyenlően járuljon hozzá az itemkalibrációhoz (ld. pl. Foy et al., 2020). Ez utóbbi megközelítés azt jelenti, hogy több információt használnak fel az itemkalibrációhoz, ami pontosabb paraméterbecsléseket biztosít, de számítási szempontból is igényesebb.

A személypontosítási eljárás során öt PV-t rendelnek minden tesztmegoldóhoz az IRT modellalapú skálapontszám-eloszlásból. A tanulói pontszámok becslésének megbízhatóságának növelése érdekében a TIMSS skálázás harmadik lépése egy populációmodellezésnek vagy kondicionálásnak nevezett folyamat. Ez azt jelenti, hogy az összes rendelkezésre álló háttér-adatot bevonják egy főkomponens-elemzésbe, és az adatok varianciájának 90%-át képviselő komponenseket minden oktatási rendszer regressziós modelljébe foglalják.

Miután a skálázás befejeződött, az újonnan létrehozott skálát összekapcsolják a TIMSS trendskálával. Ez két lépésben valósul meg. Az első lépés az új értékelési adatok látens képességeloszlásának lineáris transzformációja, hogy megfeleljen az előző ciklus eloszlásának. Az eloszlások egyeztetéséhez a közös országok, azaz a két ciklus trendországainak adatait használják fel. Ezek a trendországok egyenlően járulnak hozzá a transzformációs állandók kiszámításához. A második lépés ennek a lineáris transzformációnak az alkalmazása az újonnan

gyűjtött felmérési adatok egészére. Ezen lépések befejezése után az átalakított PV-k kerülnek a trendskálára (ld. pl. Foy et al., 2020).

Tanulási motiváció és más attitűdskálák

Az 1995 és 2007 között végzett TIMSS felmérésekben ún. indexeket képeztek bizonyos háttérkérdőív-skálához. A legújabb TIMSS háttérkérdőíves skálák (2011-től kezdődően) IRT-skálázással készülnek, a Rasch parciális kreditmodell (partial credit model, PCM) alkalmazásával. Az affektív konstrukciók longitudinális skálázása először a TIMSS 2011 és a TIMSS 2015 összekapcsolásával jelent meg (Martin et al., 2016). Jelenleg bizonyos háttérkérdőíves skálák, amelyek a 2011-es, 2015-ös és 2019-es TIMSS-ciklusokban számos horgonyitemet tartalmaztak, egy kétlépéses transzformációs folyamaton keresztül kapcsolódnak össze. Az első transzformáció a TIMSS 2019 logit skála pontszámait a TIMSS 2015 skálájára helyezi az átlag/szigma módszer (Kolen & Brennan, 2014; Marco, 1977) alkalmazásával. Ezeket a skálákat a TIMSS 2019 és a TIMSS 2015 adatok külön kalibrálásával becsülik meg. A második lépés a TIMSS 2015 logit pontszámok transzformálása a trendskálára (átlag 10, szórás 2).

Módszer

Jelen tanulmány a trendskálázási módszerek illusztrálására a magyar diákok újra skálázott matematikai teljesítmény és motiváció, valamint a természettudományos teljesítmény eredményeit mutatja be. Az empirikus munka a 13 éveseket (FIMS és SIMS), a 14 éveseket (FISS és SISS) és a nyolcadikos diákokat (TIMSS ciklusok) reprezentáló populációk adatain alapul.

Az első fázisú vizsgálatok adatait a későbbi felmérések adataihoz képest eltérően dolgozták fel. A korai felmérések adatait és dokumentációját frissített formátumokban a Center for Comparative Analysis of Educational Achievement (COMPEAT) projekt¹ tette elérhetővé. A TIMSS ciklusokhoz tartozó adatokat és dokumentációt az IEA Study Data Repository-ről² töltöttük le.

Eljárások

Az összekapcsolásról részletesen lásd Majoros (2023) és Majoros és munkatársai (2022). Jelen tanulmányban csak az értelmezéshez szükséges részletekre térünk ki. Az eljárások során alkalmazott eszközök SPSS 25 (IBM Corp., 2017) és különböző R (R Core Team, 2025) csomagok voltak.

Kognitív trendskálák

Az itemparaméterek egyidejű kalibrálását végeztük el a FIMS és SIMS itemválaszokon, melyen belül a horgonyteszt itemparamétereit az IEA által jelentett TIMSS 1995-ös értékekhez rögzítettük. Ezeket az itemparamétereket az 1999-es felmérésben végzett újra skálázási eljárás után publikálták (Martin et al., 2020). Ezután a tanulók képességeit külön becsültük meg a FIMS és a SIMS esetében, öt PV-t rendelve a diákokhoz. A tanulók képességbecsléseinek a

¹ <https://www.gu.se/en/center-for-comparative-analysis-of-educational-achievement-compeat>

² <https://www.iea.nl/data-tools/repository/timss>

TIMSS trendskálán való elhelyezéséhez a jelentett TIMSS 1995-ös skálázáshoz használt eredeti transzformációs állandókat alkalmaztuk. A természettudományos teljesítményskála felállításakor ugyanezt az eljárást alkalmaztuk.

Az országos szintű átlagok kiszámításakor az egyénenként becsült öt PV-t rendszerszinten aggregáltuk Rubin (1987) szabályait követve az adathalmazok közötti összevonásra vonatkozóan, amelyeket a Gonzalez (2014) által felvázolt többszörös imputációval értünk el. Ez az eljárás biztosítja, hogy mind a mintavételi zajt, mind a hiányzó adatokból adódó bizonytalanságot figyelembe vegyük az átlagok standard hibáinak kiszámításakor. Az átlagokat szenátusi súlyokkal (senate weight) súlyoztuk, amelyek összege minden ország adatai esetében 500. A TIMSS adminisztrációk átlageredményeit ugyanezzel az eljárással számoltuk ki.

Motivációs trendskálák

A motivációs skálák megalkotásakor egy ún. piaci kosár megközelítést alkalmaztunk. A piaci kosár megközelítés feltételezi, hogy az értékelésbe vagy felmérésbe bevont itemek meghatározzák a mért konstrukciót. Ebben az esetben a feltételezés az volt, hogy a különböző adminisztrációkban szereplő kérdőív itemek meghatározzák az egyes mérendő konstrukciókat, és reprezentatív itemek piaci kosarának tekinthetők. Az összes adminisztrációt egyidejűleg tekintve a hiányzó válaszok a kérdőívek ciklusok közötti változásainak következményei.

Országokonként egy mérési modellt alkalmaztunk a hiányzó válaszokat pótló, valószínűsíthető válaszok generálására, a Zwitser és munkatársai (2017) által javasolt eljárást követve. A mérési modellt minden országra külön illesztettük, ezzel figyelembe véve az országok közötti lehetséges különbségeket. A mérési modell a GPCM modell volt a TIMSS eljárással való összhang érdekében. A mérési modellek illesztésével becsült itemparaméterek felhasználásával a hiányzó válaszokat válaszdónként ötször imputáltuk. Ezután kiszámítottuk az egyéni összpontszámokat, így diákonként öt PV-t becsülve. Végül egy 5 átlagú és 1 szórású skálává transzformáltuk ezeket a pontszámokat, és a fent leírt aggregálást alkalmaztuk az országos átlagok kiszámításához.

Eredmények és értelmezésük

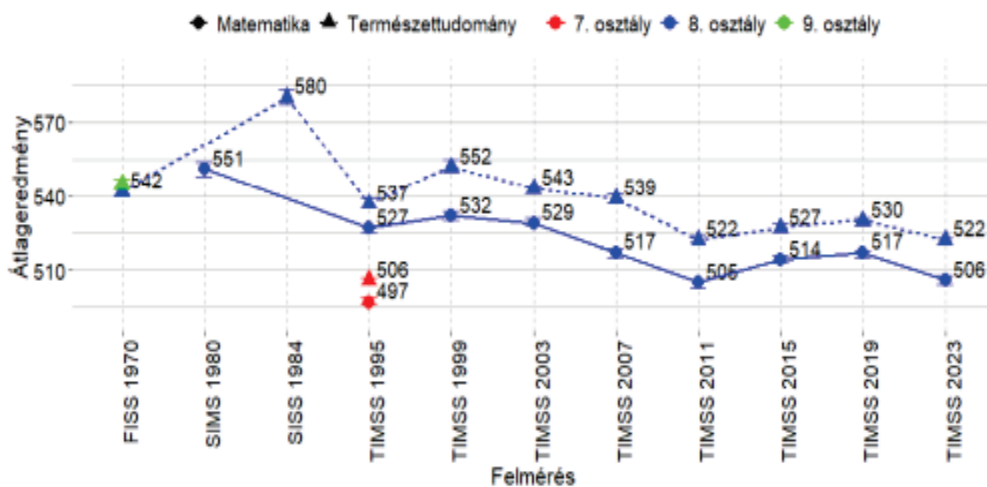
Kognitív trendek

A nyolcadik osztályos matematikai és természettudományos átlagok alakulását 1970 és 2023 között az 1. ábra mutatja be, iskolai évfolyam szerint. A matematikai korai eredményeket, tehát a SIMS mérésen elért eredményeket részletesebben tárgyalta Csíkos és Vidákovich (2012). Az összekapcsolt skálára helyezett eredmény világosan kiemelkedik a későbbi értékektől (551). A korai természettudományos eredményeket (FISS, SISS) részletesebben lásd B. Németh és munkatársai (2012) munkájában. Összességében elmondható, hogy a magyar diákok kiemelkedően jól teljesítettek a korai kognitív teszteken. A természettudományos eredmények a TIMSS felmérések kezdete óta magasabbak, mint a matematikai eredmények.

A hosszútávú trendek alakulásának elemzése összetett feladat, mely során figyelembe kell venni többek között a rendszerszintű reformokat, a tantervekben bekövetkezett változásokat, illetve a mérőeszközök illeszkedésének mértékét a magyar tantervhez viszonyítva. Az IEA-vizsgálatok elméleti háttere a részt vevő országok tantervei, valamint a helyi szakértők által megjelölt kézikönyvek elemzésén alapul (B. Németh et al., 2012), azonban a részt vevő országok számának növekedése értelemszerűen hatással van erre az illeszkedésre.

1. ábra

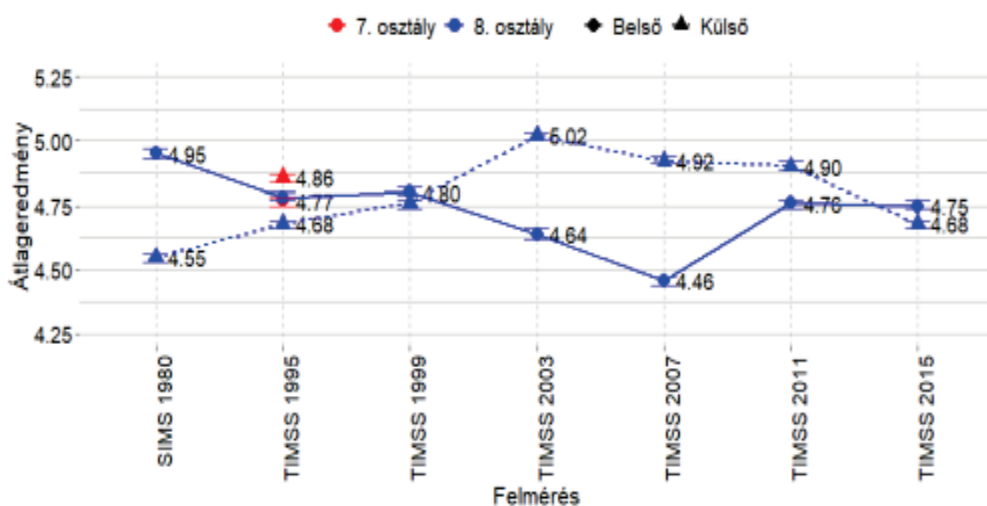
A magyar kognitív trendek 1970 és 2023 között

**A matematikatanulási motivációk trendjei**

A nyolcadikos matematikatanulási motivációs átlagok alakulását 1980 és 2015 között a 2. ábra mutatja be, iskolai évfolyam szerint. Eccles és Wigfield (2002) modelljét követve a motiváció forrása alapján különböztettük meg és soroltuk be a tanulói kérdőívek itemeit. Az instrumentális okokat (pl. jutalom, szülői elismerés) érintő kérdéseket a külső (extrinsic) motiváció, a matematikatanulás élvezetét érintő kérdéseket pedig a belső (intrinsic) motiváció fogalmához társítottuk.

2. ábra

A magyar matematikatanulási motivációs trendek 1980 és 2015 között



Ami különösen érdekes az átlageredmények áttekintésekor, hogy a belső motivációs trend hasonló a kognitív eredmények alakulásának trendjéhez, míg a külső motivációs trend közel ellenkezőleg alakult 2015-ig. A fent említett okokból itt nem vállalkozunk a trendek elemzésére.

Korlátok

A skálák használatakor figyelembe kell venni a mintavételi különbségeket. Például Strietholt és munkatársai (2013) kidolgoztak egy korrekciós modellt az országok és az IEA olvasási felmérések közötti összehasonlíthatóság javítására életkor és iskolázottság tekintetében. Az időbeli és országonkénti különbségek figyelembevételének egy másik lehetséges megközelítése az életkor és az évfolyam magyarázó változóként való kezelése.

Számos tényező befolyásolja az összekapcsolás minőségét. Ezek a tényezők magukban foglalják az értékelések közötti hasonlóság mértékét, a konstrukciók stabilitását tartalmi, jelentésbeli és kontextusbeli szempontból, valamint a horgonyitemek számát és viselkedését. Bizonyos fokú bizonytalanság lehet ezen szempontok értékelése során, mivel az összekapcsolás régi adatokat tartalmaz.

A longitudinális teljesítményskálák egyik lehetséges korlátja az országon belüli összehasonlíthatóságban rejlik a felmérések célpopulációi miatt. Az IEA mintavételi stratégiájának változását a homogén minták lehető legjobb közelítésével kezelték. Az új skálapontszámokat használó további elemzésekben az életkor és az évfolyam kontrollváltozóként kezelhető.

A TIMSS 1995-höz kapcsolódó horgonytesztekben található közös elemek száma további aggodalomra ad okot. Az affektív skálákban a horgonyitemek aránya jóval kisebb, mint a kognitív skálák esetében. Azonban az alkalmazott egyidejű kalibrációs eljárás biztosítja a legjobb megközelítést erre az esetre, ahogy arra Wingersky és Lord (1984) rámutat.

Egy másik korlátozás a hiányzó válaszok különböző típusainak kódolására és kezelésére vonatkozik a teljesítménytesztekben. Az első szakaszbeli tesztekben a hiányzó válaszok „el nem ért” típusát nem különböztették meg, ellentétben a TIMSS skálázási eljárással. Azonban lehetséges lenne ezt a megkülönböztetést elvégezni, és feltárni az eredményekre gyakorolt hatást.

A motivációs felméréseket a közelmúltig nem trendmérésre tervezték. Idővel módosítások történtek, pl. a válaszlehetőségek száma 1995-től megváltozott ötről négyre. A középső, neutrális opció kezelésének módja a SIMS-ben korlátozza a skálákat. Ezenkívül az alkalmazott összekapcsolási eljárások nem vették figyelembe a motivációs pontszám eloszlásának időbeli különbségeit, mivel longitudinálisan egycsoportos modellezést alkalmaztak.

Az affektív trendskálák egy másik korlátja, hogy az átlagok standard hibái alábecsültek. Ennek az az oka, hogy a TIMSS-ben alkalmazott rétegzett, többlépcsős mintavételi terv miatt a becslések standard hibáinak kiszámításához használt eljárásban feltételezett egyszerű véletlenszerű mintavétel nem alkalmazható (Rutkowski et al., 2010).

Végül az egyik legnagyobb kihívást jelentő fennmaradó kérdés az, hogy a kérdések megfogalmazásának változásai befolyásolják-e a motivációs indikátorok közötti belső kapcsolatot (pl. faktorszerkezet). A kérdőív itemek száma szinte minden időpontban változott, ami megnehezíti ennek a kérdésnek a vizsgálatát.

A skálák lehetséges alkalmazásai

Az itt ismertetett skálák lehetőséget teremtenek hosszú távú változások összehasonlító vagy országokon belüli vizsgálatára. Az első szakaszbeli felmérések kognitív skálái nyilvánosan elérhetők a COMPEAT weboldalon³, a skálák összekapcsolásának dokumentációjával együtt. A rendszerszintű változások időt vesznek igénybe, ezért az iskolareformok értékeléséhez hosszú távú adatokra van szükség. Továbbá, az ok-okozati kutatási kérdések megválaszolására szolgáló hatékony statisztikai megközelítések alkalmazhatók a rendszerszintű longitudinális adatokra.

Mindazonáltal megfigyelési adatokból érvényes ok-okozati következtetések levonása kihívást jelent (ld. pl. Allardt, 1990; Rutkowski & Delandshere, 2016). A megfigyelési adatok ebben az összefüggésben olyan adatokat jelentenek, amelyeket a tanulók kezelésének vagy csoportosításának ellenőrzése nélkül gyűjtöttek, mint az IEA felmérések. Számos kutató tett javaslatokat az ilyen adatokat alkalmazó eljárásokra (ld. pl. Gustafsson, 2008; Gustafsson & Nilsen, 2022; Robinson, 2013; Schlotter et al., 2014).

Ilyen kutatási terület például az oktatáspolitikák társadalmi-gazdasági teljesítményszakadéokra gyakorolt hatása. Strietholt és munkatársai (2019) áttekintették a nemzetközi összehasonlító szakirodalmat, és megállapították, hogy a meglévő kutatások nagy része leíró jellegű, és csupán keresztmetszeti adatokon alapuló korrelációkat becsl. Az első szakaszbeli IEA-felmérések társadalmi-gazdasági háttérmutatóinak feltérképezésére irányuló további kutatások a teljesítményskálákkal kombinálva potenciálisan hozzájárulhatnak ehhez a vizsgálati irányhoz. A longitudinális skálák kihasználásának egy másik lehetséges területe a globális oktatási reformmozgalommal kapcsolatos kérdések (Fuller & Stevenson, 2019; Sahlberg, 2016). Ilyen kapcsolódó jelenségek például a privatizáció, a szabad iskolaválasztás, az iskolai verseny vagy a tanárképzés (Tóth & Sipos, 2014).

Irodalom

- Adams, R. J., Wu, M., & Macaskill, G. (1996). Scaling methodology and procedures for the mathematics and science scales. In M. O. Martin & D. L. Kelly (Eds.), *Third international mathematics and science study technical report* (Vol. 2, pp. 111–146). TIMSS & PIRLS International Study Center, Boston College.
- Afrassa, T. M. (2005). Monitoring mathematics achievement over time: A secondary analysis of FIMS, SIMS and TIMS: A Rasch analysis. In C. Alagumalai, D. David, & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars* (pp. 61–77). Springer. https://doi.org/10.1007/1-4020-3076-2_4
- Allardt, E. (1990). Challenges for comparative social research. *Acta Sociologica*, 33(3), 183–193. <https://doi.org/10.1177/000169939003300302>
- B. Németh, M., Korom, E., & Nagy, L. (2012). A természettudományos tudás nemzetközi és hazai vizsgálata. In B. Csapó (Ed.), *Mérlegen a magyar iskola* (pp. 131–190). Nemzeti Tankönyvkiadó.
- Beaton, A. E., & Johnson, E. G. (1992). Overview of the scaling methodology used in the national assessment. *Journal of Educational Measurement*, 29(2), 163–175. <https://doi.org/10.1111/j.1745-3984.1992.tb00372.x>
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (Second edition). The Guilford Press.
- Chmielewski, A. K. (2019). The global increase in the socioeconomic achievement gap, 1964 to 2015. *American Sociological Review*, 84(3), 517–544. <https://doi.org/10.1177/0003122419847165>

³ <https://www.gu.se/en/center-for-comparative-analysis-of-educational-achievement-compeat/mathematics-and-science>

- Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). Harper & Row.
- Csikós, Cs., Pásztor, A., Rausch, A., & Sztányi, J. (2020). A matematikai nevelés kutatásának aktuális irányzatai. *Magyar Tudomány*, 181(1), 24–33. <https://doi.org/10.1556/2065.181.2020.1.3>
- Csikós, Cs., & Vidákovich, T. (2012). A matematikatudás alakulása az empirikus vizsgálatok tükrében. In B. Csapó (Ed.), *Mérlegen a magyar iskola* (pp. 83–130). Nemzeti Tankönyvkiadó.
- Dorans, N. J., Moses, T. P., & Eignor, D. R. (2011). Equating test scores: Toward best practices. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 21–42). Springer. https://doi.org/10.1007/978-0-387-98138-3_2
- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53, 109–132. <https://doi.org/10.1146/annurev.psych.53.100901.135153>
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Psychology Press.
- Foy, P., Fishbein, B., von Davier, M., & Yin, L. (2020). Implementing the TIMSS 2019 scaling methodology. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report* (pp. 12.1–12.146). TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
- Fuller, K., & Stevenson, H. (2019). Global education reform: Understanding the movement. *Educational Review*, 71(1), 1–4. <https://doi.org/10.1080/00131911.2019.1532718>
- Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics: An introduction* (2nd ed.). SAGE Publications.
- Gonzalez, E. J. (2014). Calculating standard errors of sample statistics when using international large-scale assessment data. In R. Strietholt, W. Bos, J.-E. Gustafsson, & M. Rosén (Eds.), *Educational policy evaluation through international comparative assessments* (pp. 59–73). Waxmann Verlag.
- Gustafsson, J.-E. (2008). Effects of international comparative studies on educational quality on the quality of educational research. *European Educational Research Journal*, 7(1), 1–17. <https://doi.org/10.2304/eerj.2008.7.1.1>
- Gustafsson, J.-E., & Nilsen, T. (2022). Methods of causal analysis with ILSA data. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International handbook of comparative large-scale studies in education* (pp. 1–28). Springer International Publishing. https://doi.org/10.1007/978-3-030-38298-8_56-1
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. SAGE Publications.
- Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth*, 17(4), 267–321. <https://doi.org/10.1007/s10887-012-9081-x>
- Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (pp. 187–220). Praeger Publishers.
- IBM Corp. (2017). *IBM SPSS Statistics for Windows*. IBM Corp.
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2), 183–202. <https://doi.org/10.1007/bf02289343>
- Kline, R. B. (2016). *Principles and practice of structural equation modeling*. Guilford Press.
- Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (pp. 155–186). Praeger Publishers.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3. ed.). Springer.
- Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 6(1), 83–102. https://doi.org/10.1207/s15324818ame0601_5
- Majoros, E. (2023). Linking the first- and second-phase IEA studies on mathematics and science. *Large-Scale Assessments in Education*, 11(1), 14. <https://doi.org/10.1186/s40536-023-00162-y>
- Majoros, E., Christiansen, A., & Cuellar, E. (2022). Motivation towards mathematics from 1980 to 2015: Exploring the feasibility of trend scaling. *Studies in Educational Evaluation*, 74, 101174. <https://doi.org/10.1016/j.stueduc.2022.101174>

- Majoros, E., Rosén, M., Johansson, S., & Gustafsson, J.-E. (2021). Measures of long-term trends in mathematics: Linking large-scale assessments over 50 years. *Educational Assessment, Evaluation and Accountability*, 33(1), 71–103. <https://doi.org/10.1007/s11092-021-09353-z>
- Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 14(2), 139–160. <https://doi.org/10.1111/j.1745-3984.1977.tb00033.x>
- Martin, M. O., & Kelly, D. L. (Eds.). (1996a). *Third international mathematics and science study technical report: Design and development* (Vol. 1). TIMSS & PIRLS International Study Center, Boston College.
- Martin, M. O., & Kelly, D. L. (Eds.). (1996b). *Third international mathematics and science study technical report: Implementation and analysis* (Vol. 2). TIMSS & PIRLS International Study Center, Boston College.
- Martin, M. O., Mullis, I. V. S., Foy, P., Arora, & Alka (2012). Creating and interpreting the TIMSS and PIRLS 2011 context questionnaire scales. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS 2011*. TIMSS & PIRLS International Study Center, Boston College. <https://timssandpirls.bc.edu/methods/t-context-q-scales.html>
- Martin, M. O., Mullis, I. V. S., Hooper, M., Yin, L., Foy, P., & Palazzo, L. (2016). Creating and interpreting the TIMSS 2015 context questionnaire scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 15.1–15.312). TIMSS & PIRLS International Study Center, Boston College.
- Martin, M. O., von Davier, M., & Mullis, I. V. S. (Eds.). (2020). *Methods and procedures: TIMSS 2019 technical report*. TIMSS & PIRLS International Study Center, Boston College.
- Mazzeo, J., Lazer, S., & Zieky, M. J. (2006). Monitoring educational progress with group-score assessments. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 681–699). Praeger Publishers.
- Mazzeo, J., & von Davier, M. (2013). Linking scales in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 230–258). CRC Press. <https://doi.org/10.1201/b16061-15>
- Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, 7(4), 361–388. <https://doi.org/10.1177/1094428104268027>
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525–543. <https://doi.org/10.1007/BF02294825>
- Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, 44(11), S69–S77. <https://doi.org/10.1097/01.mlr.0000245438.73837.89>
- Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56(2), 177–196. <https://doi.org/10.1007/BF02294457>
- Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Educational Testing Service, Policy Information Center. <https://eric.ed.gov/?id=ED353302>
- Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29(2), 133–161. <https://doi.org/10.1111/j.1745-3984.1992.tb00371.x>
- Molnár, Gy. (2006). A Rasch-modell alkalmazása a társadalomtudományi kutatásokban. *Iskolakultúra*, 16(12), 99–113.
- OECD. (2019). *TALIS 2018 technical report*. OECD Publishing. http://www.oecd.org/education/talis/TALIS_2018_Technical_Report.pdf
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71–90. <https://doi.org/10.1016/j.dr.2016.06.004>
- R Core Team. (2025). *R: A language and environment for statistical computing*. <https://www.R-project.org/>

- Robinson, J. P. (2013). Causal inference and comparative analysis with large-scale assessment data. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 521–545). CRC Press.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, Inc. <https://doi.org/10.1002/9780470316696>
- Rutkowski, D., & Delandshere, G. (2016). Causal inferences with large scale assessment data: Using a validity framework. *Large-Scale Assessments in Education*, 4(1), Article 6. <https://doi.org/10.1186/s40536-016-0019-1>
- Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, 39(2), 142–151. <https://doi.org/10.3102/0013189x10363170>
- Sahlberg, P. (2016). The global educational reform movement and its impact on schooling. In K. E. Mundy, A. Green, B. Lingard, & A. Verger (Eds.), *The handbook of global education policy* (pp. 128–144). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781118468005.ch7>
- Schlotter, M., Schwerdt, G., & Woessmann, L. (2014). Econometric methods for causal evaluation of educational policies and practices: A non-technical guide. In R. Strietholt, W. Bos, J.-E. Gustafsson, & M. Rosén (Eds.), *Educational policy evaluation through international comparative assessments* (pp. 95–126). Waxmann.
- Sosa Paredes, Y. K., & Andersson, B. (2025). Linking the first, second and third IEA science studies. *Studies in Educational Evaluation*, 87, 101529. <https://doi.org/10.1016/j.stueduc.2025.101529>
- Stevens, S. S. (1968). Measurement, statistics, and the schemapiric view. *Science*, 161(3844), 849–856. <https://doi.org/10.1126/science.161.3844.849>
- Strietholt, R., Gustafsson, J.-E., Hogrebe, N., Rolfe, V., Rosén, M., Steinmann, I., & Hansen, K. Y. (2019). The impact of education policies on socioeconomic inequality in student achievement: A review of comparative studies. In L. Volante, S. Schnepf, J. Jerrim, & D. Klinger, *Socioeconomic inequality and student outcomes. Education Policy & Social Inequality* (Vol. 4, pp. 17–38). Springer Singapore. https://doi.org/10.1007/978-981-13-9863-6_2
- Strietholt, R., & Rosén, M. (2016). Linking large-scale reading assessments: Measuring international trends over 40 years. *Measurement: Interdisciplinary Research and Perspectives*, 14(1), 1–26. <https://doi.org/10.1080/15366367.2015.1112711>
- Strietholt, R., Rosén, M., & Bos, W. (2013). A correction model for differences in the sample compositions: The degree of comparability as a function of age and schooling. *Large-Scale Assessments in Education*, 1(1), Article 1. <https://doi.org/10.1186/2196-0739-1-1>
- Tóth, E., & Sipos, J. (2014). A rendszerszintű mérések hatása a pedagógusok tanítási gyakorlatára: Egy tanári kérdőíves vizsgálat tanulságai. *Iskolakultúra*, 24(10), 38–53.
- Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement*, 8(3), 347–364. <https://doi.org/10.1177/014662168400800312>
- Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika*, 81(4), 1014–1045. <https://doi.org/10.1007/s11336-016-9506-0>
- Yin, L., & Fishbein, B. (2020). Creating and interpreting the TIMSS 2019 context questionnaire scales. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures* (pp. 16.1–16.331). TIMSS & PIRLS International Study Center, Boston College.
- Zwitser, R. J., Glaser, S. S. F., & Maris, G. (2017). Monitoring countries in a changing world: A new look at DIF in international surveys. *Psychometrika*, 82(1), 210–232. <https://doi.org/10.1007/s11336-016-9543-8>

ABSTRACT

TREND SCALING METHODS: LINKING RECENT AND OLDER IEA STUDIES ON MATHEMATICS AND SCIENCE

Erika Majoros

Keywords: linking scales, international large-scale assessments, mathematics achievement, science achievement, mathematics motivation, TIMSS

One of the international large-scale assessments administered by the International Association for the Evaluation of Educational Achievement (IEA) is the Trends in International Mathematics and Science Study (TIMSS), which has been measuring mathematics and science achievement in grades four and eight since 1995. TIMSS was preceded by four IEA surveys, the results of which were not linked to the TIMSS scale. This study describes possible trend scaling methods and, to illustrate the achieved linkage, presents the rescaled mathematics and science performance and motivation for mathematics of Hungarian students.

Magyar Pedagógia, 126(2). 83–98. (2026)
<https://doi.org/10.14232/mped.2026.2.83>

Majoros Erika:  <https://orcid.org/0000-0003-4417-5016>
Umeå University, Svédország
SE-901 87 Umeå, Svédország
erika.majoros@umu.se