MAGYAR PEDAGÓGIA
103. évf. 4. szám 423–446. (2003)

AZ ISMERETEK ALKALMAZÁSÁNAK VIZSGÁLATA MODERN TESZTELMÉLETI (IRT) ESZKÖZÖKKEL

Molnár Gyöngyvér
Szegedi Tudományegyetem, Pedagógia Tanszék, MTA Képességkutató Csoport

A klasszikus tesztelméleti módszerekkel történő elemzéseknek Magyarországon jelentős múltja van. A számítástechnikai lehetőségek kiszélesedése, az egyre szélesebb körben is hozzáférhető programok, valamint az utóbbi évek nemzetközi vizsgálatainak elemzései rávilágítanak egy alapjaiban más módszerekkel, más alapokon nyugvó tesztelmélet fontosságára. A modern tesztelméleti eszközökkel végzett elemzésekből levonható következtetések a modern tesztelmélet valószínűségi tulajdonsága miatt nem fogalmazhatóak meg ugyanabban a determinisztikus szemléletmódban, mint a klasszikus tesztelméleti eszközökkel alátámasztott következtetések. A következő tanulmány egyik célja, hogy elindítson, illetve folytasson – hiszen nem ez az első magyar nyelvű modern tesztelmélettel és elemzésekkel foglalkozó írás – egy alapvetően új értékelési módszert és nyelvezetet.

Elméleti keret

A klasszikus tesztelmélethez képest a tesztelméletek újabb generációját képező modern (probabilisztikus, valószínűségi) tesztelmélet (Item Response Theory [IRT]) az itemek tulajdonságait valószínűségelméleti eszközökkel jellemzi (*Csapó*, 2000). A modern tesztelmélet kialakulását elősegítették a klasszikus tesztelmélettel kapcsolatban felmerült kritikák: a populációfüggőség és az ebből következő szórásfüggőség, skálafüggőség és a harmadik axióma kritikája (lásd részletesebben *Horváth*, 1997). A klasszikus tesztelméleti eszközökkel történő elemzések során nem lehet szétválasztani a populáció képességei okozta faktort és a teszt eredményeinek hatását, azaz nehéz megállapítani, hogy populációsajátosságról, vagy teszthibáról van-e szó. A modern tesztelmélet nem a klasszikus tesztelmélet egy továbbfejlesztett, vagy „jobb" változata, hanem alapvetően más matematikai eszközökre támaszkodó, statisztikai eljárásokat használó, modelleket felállító és függvényekkel dolgozó tesztelmélet.

Az egyes IRT (Item Response Theory) modellek különböző dimenziók mentén csoportosíthatók. Eltérhetnek egymástól abban, hogy milyen típusú összefüggést feltételeznek a helyes válasz valószínűsége és a válaszoló képessége között; a válaszok szintjén dichotóm, vagy nem dichotóm itemek elemzésére alkalmas-e a modell, illetve a legel-

terjedtebb mód a modellek itemparaméterek száma szerinti osztályozása. E tanulmányban részletesebben az utóbbi két csoportosítási móddal foglalkozunk. Más osztályozási módokról, illetve további modellekről lásd *Linden* és *Hambleton* (1997) könyvét.

Dichotóm adatok elemzésére alkalmas a Rasch modell (Rasch's simple logistic model) (*Rasch*, 1980). Alkalmazásáról lásd részletesebben *Bond* és *Fox* könyvét (2001). Nem dichotóm kódolású adatok elemzésére alkalmas *Masters* (1982) parciális kredit modellje (partial credit model). Például attitűd vizsgálatnál Likert skálán mért adatok elemzésére alkalmas *Andricht* (1978) rangskálás modellje (rating scale model). Röviden kitérnék a két modell közti különbségre. A rangskálás modellel elemzett adatbázis minden egyes itemének megegyező a skálaszerkezete. Ezzel szemben a parciális kredit modellben minden egyes itemnek akár teljesen különböző skálaszerkezete is lehet. Ez a tulajdonság megemeli a közelíthető szabad paraméterek számát (L-1)*(m-2) -re, ahol L: az itemek száma, m: a rangskálán lévő kategóriák száma (*Linacre*, 2000).

Mind dichotóm, mind nem dichotóm adatok elemzésére is alkalmasak az alábbi modellek. *Wilson* (1992) rendezett elosztási modellje (ordered partition model) külön tudja kezelni a kategória és az értékelés szintjét, azaz egy item esetében több kategória ugyanazt az értékelést kaphatja. Például a fogalmi megértés vizsgálatában a fogalommagyarázat négy választási lehetőséget tartalmaz. Egy tudományos magyarázatot, amire 2 pontot adunk, két részben korrekt, de minőségében különböző tévképzetet, amelyekre 1–1 pontot adunk és egy naív magyarázatot, ami 0 pontot ér. A modell ezeket az adatokat úgy kezeli és elemzi, mint egy négy kategóriás itemet, amelyiknek három különböző pontozása van. *Fischer* (1983) lineáris logisztikus teszt modellje (linear logistic test model) az egyszerű Rasch modell kiterjesztése. Az itemnehézségi paramétert több alapvető tényező lineáris kombinációjából határozza meg. *Linacre* (1994) sokoldalú modellje (multifaceted model) a válaszok elemzése során kezelni tudja azt, hogy egy nemcsak zárt kérdésekből álló feladatlapon (amit javítani kell) a tanulók eredményeit nemcsak a feladatok és a tanuló képességei, hanem a javító szigorúsága is befolyásolja. Ezáltal a kétoldalú mérést kiterjesztette háromoldalúra, amelyet a modell „sokoldalúságából" adódóan még tovább lehet bővíteni. A kiterjesztett egydimenziós modellek (generalised unidimensional models) (*Wu*, *Adams* és *Wilson*, 1998) lehetőséget adnak a fent említett modellek tetszőleges kombinációjának használatára, illetve saját modellek létrehozására.

A többdimenziós modern tesztelméleti modellek (multidimensional item response models) olyan itemek elemzésére is alkalmasak, amelyek több rejtett dimenziót tartalmaznak. *Adams*, *Wilson* és *Wang* (1997) nyomán a többdimenziós tesztek két fajtáját említeném meg: (1) az itemek közötti többdimenziós teszt (multidimensional between-item test), (2) az itemeken belüli többdimenziós teszt (multidimensional within-item test). Részletes leírásukat lásd *Wu*, *Adams* és *Wilson* (1998) könyvében.

Az itemparaméterek száma szerint a modellek három csoportját különíthetjük el: egy-paraméteres logisztikus modell – Rasch modell; két-paraméteres logisztikus modell; illetve három-paraméteres logisztikus modell.

Az egy-paraméteres logisztikus modellben (más néven Rasch modell) a személyparaméteren kívül egy paraméter, az itemnehézségi mutató szerepel. Ebben a modellben minden egyes item diszkriminációs indexe azonos, azaz az itemek karakterisztikus görbéi egymással párhuzamosan futnak.

A két-paraméteres logisztikus modell abban különbözik az egy-paraméteres logisztikus modelltől, hogy az itemnehézségi mutatón kívül az item diszkriminációs indexe is külön paraméterként szerepel. Az itemek karakterisztikus görbéi ebben a modellben nemcsak párhuzamosan futhatnak, hanem a különböző diszkriminációs indexű itemek karakterisztikus görbéi át is metszik egymást.

A három-paraméteres logisztikus modell figyelembe veszi a sikeres találgatás valószínűségét is, ennek következtében az itemek karakterisztikus görbéi különböző helyen metszik az ordináta tengelyt. A helyes válasz valószínűsége alacsony képességű személyeknél nem a nullához konvergál.

A bonyolult matematikai eszközökön alapuló számítások egy részét az OPLM (One-Parameter Logistic Model) program segítségével végeztük. A *Verhelst*, *Glas* és *Verstralen* (1995) által írt software egy érdekes modellen alapul, ami valahol a Rasch modell és a több-paraméteres modellek között helyezhető el. Bár a modell a közelítő eljárások során csak egy itemparamétert használ (nehézségi index), de a diszkriminációs indexeket előre meghatározott állandóként kezeli, amivel kiterjeszti az azonos diszkriminációs indexeket feltételező Rasch modell alkalmazhatóságát. Ennek következtében nem sorolható sem az egy-, sem a két-paraméteres logisztikus modellcsaládba sem. Az elkülönítésmutatókat jól kidolgozott eljárásokkal becsli. Az elemzések másik részéhez a Rasch modell mellett több modellel is dolgozó Quest és annak továbbfejlesztett változatát a ConQuest softwaret (*Wu*, *Adams* és *Wilson*, 1998) használtuk. Az eredmények közötti eltérések a modellek között fennálló eltérések következményei.

A feladatlapok kvantitatív adatelemzése során a változókat dichotóm változóként kezeltük. A helyes válasz 1, a helytelen 0 pontot ért. A második szintű feladatlapok hídfeladatai és a modern tesztelméleti eszközökkel számoló programcsomagok által lehetőség nyílt a három szinten előforduló összes feladat egy skálára hozására és az egyes itemek, tesztek valószínűségi alapokon nyugvó, populációfüggetlen értékelésére.

A felmérés módszerei

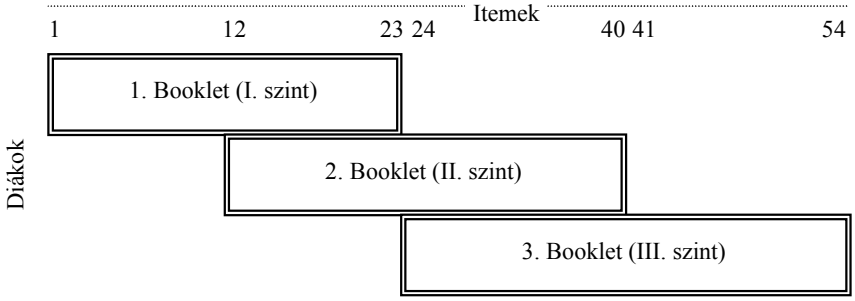
A felméréshez összeállított mintáról, a mérés lebonyolításáról, szerkezetéről és a feladatlapokról részletesebben lásd *Molnár* (2003). Jelen tanulmányban csak az értelmezéshez szükséges részletekre térünk ki.

Vizsgálatunkat 2002 tavaszán 5337 tanuló részvételével három magyarországi nagyváros általános és középiskoláiban végeztük. A felmérés során a mérőeszközök kitöltésére egy teljes tanítási óra állt a diákok rendelkezésére. Az általános iskolákban a harmadikos évfolyamtól a végzős tanulókig minden évfolyam részt vett az adatfelvételben, a középiskolákban kilencedik évfolyamtól a tizenegyedik évfolyamig terjedt a résztvevők köre. Alsóbb osztályokban az olvasási képesség alacsony szintje miatt nem alkalmazhattuk tesztjeinket.

A mérőeszközök értékelése és itemanalízise a modern tesztelmélet alapján

Az adatok bevitele

Az adatok bevitele bookletek formájában történt. Az 1. ábra mutatja az egyes itemek, szintek és bookletek egymáshoz való viszonyát, illetve a második booklet, azaz a második szintű feladatsor itemeinek összekötő hídfunkcióját (*anchor item*) az első és a harmadik szint itemei között. Például az első és második szintű feladatlapot a 12-23 itemek kapcsolják össze, amelyek mindkét szinten azonosak. Ezeket az itemeket, amelyek legalább két bookletben megtalálhatóak, horgonyzott, azaz anchor itemeknek nevezzük.



1. ábra
A komplex problémamegoldó feladatlap-sorozat személy-item mátrixa
(Verhelst és mtsai, 1995 alapján)

A mérőeszközök megbízhatósága

A komplex problémamegoldó gondolkodás fejlettségét nem lehet homogén feladatokat tartalmazó tesztekkel vizsgálni. Ennek következtében a komplex problémamegoldó feladatlapok problémái nem egy egységes tudásterülettel foglalkoznak, megfogalmazásuk különbözik az iskolában megszokottól. Az életszerűséggel együtt járó komplexitásból adódóan a hagyományos tudás, vagy képességszintmérő teszteknél tapasztaltakhoz képest kevesebb itemet tartalmaznak, valamint mind tartalmilag, mind a feladattípusokat tekintve inhomogének. Ebből az inhomogenitásból következik, hogy a mérőeszközök megbízhatóságát jellemző, az egységes tudásterületet vizsgáló tudásszintmérő teszteknél elfogadott magasabb reliabilitásmutatóknál (0,9 feletti) alacsonyabb, de még az eredmények kvantitatív elemzésére megfelelő értékeket kapunk.

A modern és klasszikus tesztelméleti számításokra egyaránt alkalmas OPLM programcsomaggal a dichotóm skála helyett faktorsúlyok bevezetésével is elvégeztük az alapvető tesztelemzési számításokat. A pontozás finomításával magasabb reliabilitásmu-

tatókat kaptunk, azaz súlyozással pontosabban értékelhető a tanulók komplex problémamegoldó képessége. Az 1. táblázat mutatja az egyes szintek dichotóm kategóriákra, illetve súlyozott értékekre vonatkozó átlagát, szórását, Cronbach α-t és a súlyozott – súlyozatlan értékek közötti korrelációt.

1. táblázat. A komplex problémamegoldó feladatlapok átlaga, szórása és Cronbach α-ja dichotóm, illetve súlyozott értékek mellett

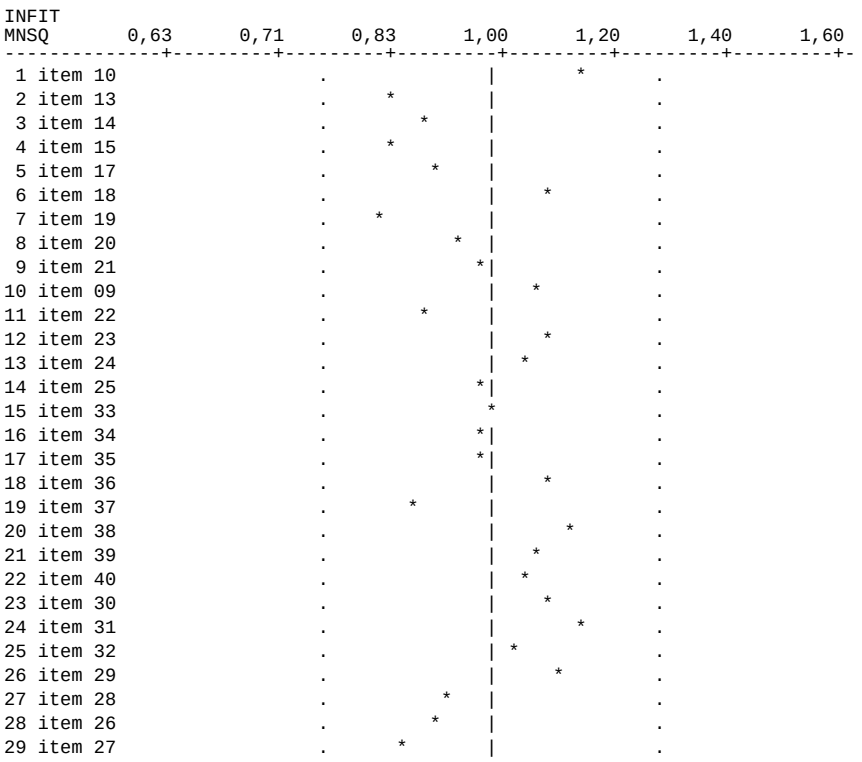
Szint	*I. szint (N=1660; itemszám=23)*		*II. szint (N=1597; itemszám=29)*		*III. szint (N=1729; itemszám=31)*	
Skála	*Dichotóm skála*	*Súlyozott értékek*	*Dichotóm skála*	*Súlyozott értékek*	*Dichotóm skála*	*Súlyozott értékek*
Átlag	10,790	40,854	13,926	48,011	13,890	41,006
Szórás	4,712	20,072	5,211	20,254	4,713	15,314
Alpha	0,814	0,827	0,828	0,834	0,766	0,797
r (súlyozott, súlyozatlan)	0,990		0,989		0,980	

Az itemek modell-illeszkedése, jelleggörbéinek megrajzolása és a diszkriminációs index jelentőssége

Az item modell-illeszkedése a modell által elvárt, előre jelzett és a valós teljesítmény közötti különbséget mutatja. Az itemek modell-illeszkedésének és a feladatok megoldásához szükséges képességszintek analízise során első lépésként a feladatlapokat szintenként külön elemeztük. Ezt követte a három szint feladatlapjainak egy tesztként való kezelése, továbbá a következő fejezetben az egy dimenziós modellből a feladatok matematikai és természettudományos irányultságát kihasználva a két dimenziós modellbe való áttérés.

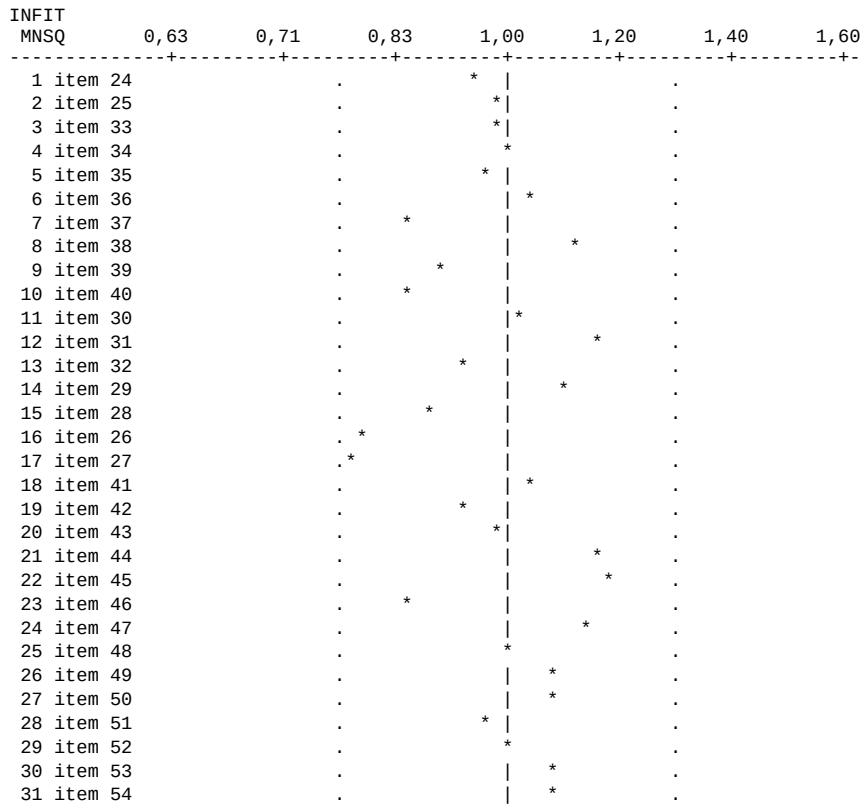
Az itemek modell-illeszkedését mutatja az infit paraméter. Az infit paraméterek kiszámolásához a Rasch modellel dolgozó Quest programot használtuk. A program az infit paraméterek átlagát automatikusan 1-nek veszi. Az egyes itemek annál jobban illeszkednek a modellhez, minél közelebb van az adott item infit paramétere – a megadott elfogadási sávon belül ($p<0,05$) – nullához. Általánosságban megfogalmazható, hogy a 0,70 és a 1,30 közötti értékek fogadhatóak el, az 1,30 felettiek nem, a 0,70 alattiak túlilleszkednek. Az egyszerűbb áttekintés kedvéért grafikusan ábrázoljuk a paraméterértékeket és az infit paraméterértékek elfogadható intervallumát. Mivel az első két szint itemeinek modell-illeszkedése nagyon hasonló, ezért kiemeltük a második szintű feladatsor feladatait, amelynek infit paramétereit a 2. ábra mutatja. Az ábrán az itemek az adatbázisban szereplő sorrendben szerepelnek egymás alatt. (Az 'item' felirat után található szám az item nevének utolsó két számjegyét jelenti, a felirat előtti sorszám pedig az adatbázisban elfoglalt helyét.)

Ha az itemeket illeszkedés szerint sorba rendeznénk, akkor mind első, mind második szinten a modellhez legjobban illeszkedő item a 19-es, a legkevésbé illeszkedő a 10-es item lenne. Mindkét szinten a megengedett sávon belül vannak az itemek infit paraméterei, ezért az első és második szintű tesztről elmondható, hogy minden egyes iteme jól illeszkedik az adott szintű feladatlap feladataiból álló modellhez.



2. ábra
A második szintű komplex problémamegoldó feladatlap itemeinek modell-illeszkedése

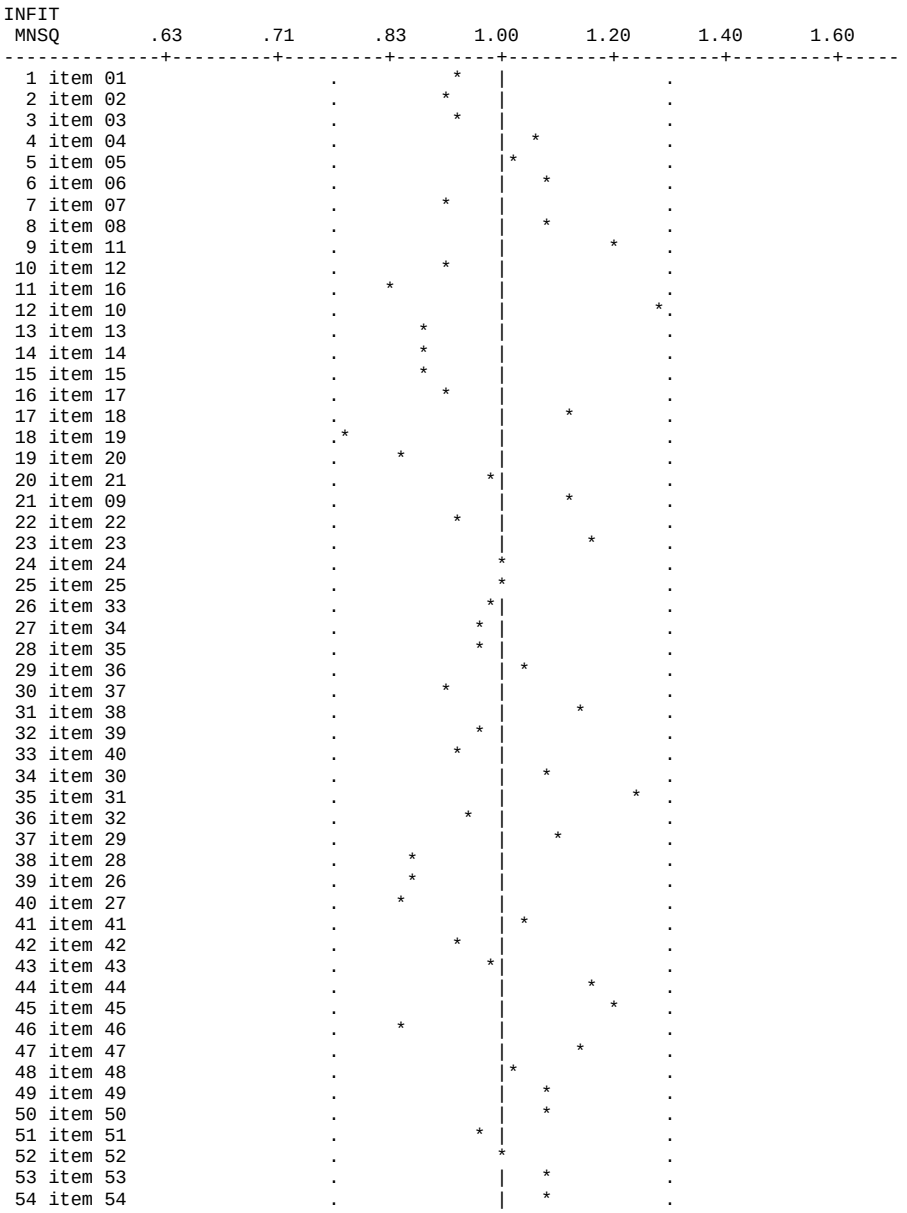
A középiskolások komplex problémamegoldó gondolkodásának fejlettségi szintjét vizsgáló harmadik szintű feladatlap itemeiről is hasonló megállapítás tehető, mint az első és második szint itemeiről. A 3. ábra mutatja a harmadik szintű feladatlap itemeinek infit paraméterértékeit. Az itemek közül a 27-es item modell-illeszkedése a legjobb. Második szinten ugyanezen itemnél a 19., 13. és 14. item modell-illeszkedése erősebb (ezek az itemek harmadik szinten nem fordulnak elő). Nincs 1,2 feletti infit paraméterérték, azaz a harmadik szintű komplex problémamegoldó feladatlap itemei jól illeszkednek a modellhez. A feladatlap esetleges továbbfejlesztése során a modell-illeszkedés szempontjából egyik itemet sem kellene kihagyni a tesztből.



3. ábra
A harmadik szintű komplex problémamegoldó feladatlap itemeinek modell-illeszkedése

Miután külön-külön elemeztük a három komplex problémamegoldó feladatlap itemeinek modell-illeszkedését, és megállapítottuk, hogy a tesztek itemei tesztenként jó modell-illeszkedésűek, megnézzük, hogy hogyan viselkednek az itemek infit paraméterei a három feladatlapot egy tesztként kezelve. A Rasch modell lehetőséget teremt a három teszt együttes elemzésére, egy tesztként való kezelésére, a feladatok egy skálára hozására. A második szintű feladatlap hídfunkcióját kihasználva közös modellben elemezhetjük a tesztek itemeit. A 4. ábra egy modellben mutatja a három feladatlapon szereplő 54 különböző item modell-illeszkedését. Az infit paraméterek alapján a 16 és 19-es itemek modell-illeszkedése a legjobb, és a 10-es és 31-es itemeké a leggyengébb. Ezt erősíti az itemek jelleggörbéjének lefutása és diszkriminációs indexe is (lásd később). Amint a fejezet későbbi részében látni fogjuk, a 16-os és 19-es itemek diszkriminációs indexe a legmagasabb, 6-os, azaz ezek az itemek különítik el legjobban a diákokat egymástól, míg a 10-es és 31-es itemek elkülönítésmutatója 1-es. Ezek túl könnyűnek bizonyultak, a diákok legnagyobb része sikeresen oldotta meg ezt a két feladatot. Az 54 itemet egy

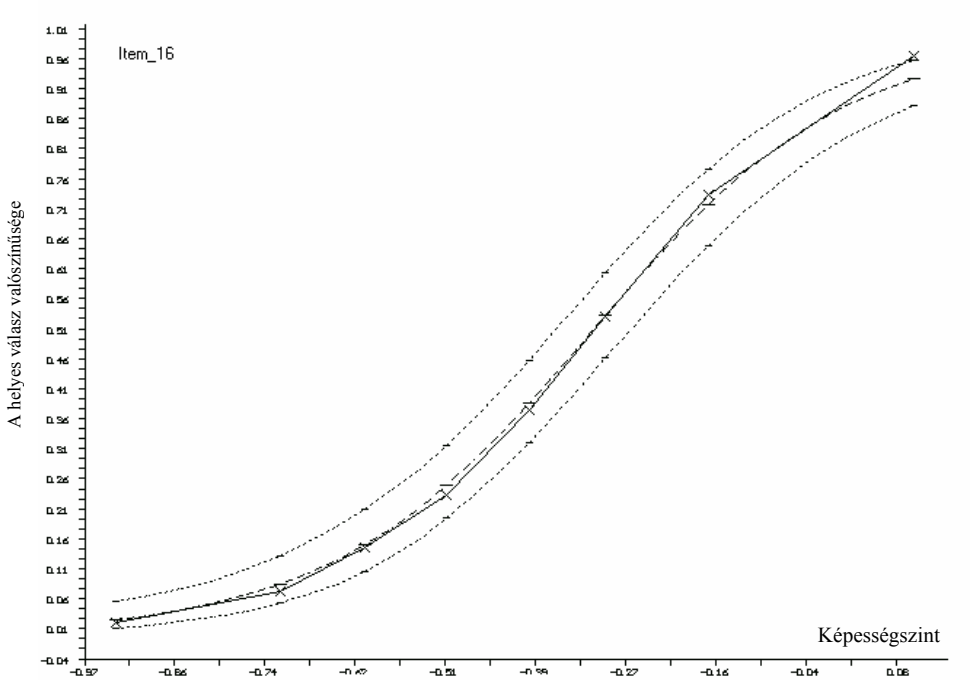
tesztként kezelve megállapítható, hogy minden egyes item illeszkedik a modellhez, az elemzéseknél és a teszt továbbfejlesztésénél egyiket sem szükséges elhagyni.



4. ábra
A komplex problémamegoldó feladatlap-sorozat itemeinek modell-illeszkedése

Az itemek modell-illeszkedését mutatja az item jelleggörbe is. Az itemek modell-illeszkedésének részletesebb elemzéséhez az OPLM program grafikus modulját használtuk. Az elemzés során minden item jelleggörbéjét megrajzoltuk, itt csak a legjellegzetesebbeket vizsgáljuk meg. A görbék kiválasztását lefutásuk mellett diszkriminációs indexük határozta meg. Mint korábban említettük, a valószínűségi tesztelmélet ezen mutatója azt jelzi, hogy az adott item mennyire tudja jól elkülöníteni egymástól a jó, illetve rossz képességű tanulókat. Minél magasabb ez a paraméter, annál jobban differenciálja a diákokat az adott item. Az OPLM a diszkriminációs indexeket úgy alakítja ki, hogy az értékek mértani közepe egy előre megadott szám, alapértelmezésben 3 legyen. Az 54 item diszkriminációs indexei 1 és 6 között helyezkednek el, illetve az itemek többségének jelleggörbéje végig a modell által megengedett hibasávon belül fut. A grafikonokról többek között leolvasható, milyen képességszint szükséges az item adott valószínűséggel történő megoldásához, továbbá mely képességcsoportúak oldják meg nagyobb valószínűséggel az adott itemet, valamint mennyire különíti el egymástól az adott item a jó és rossz képességű diákokat.

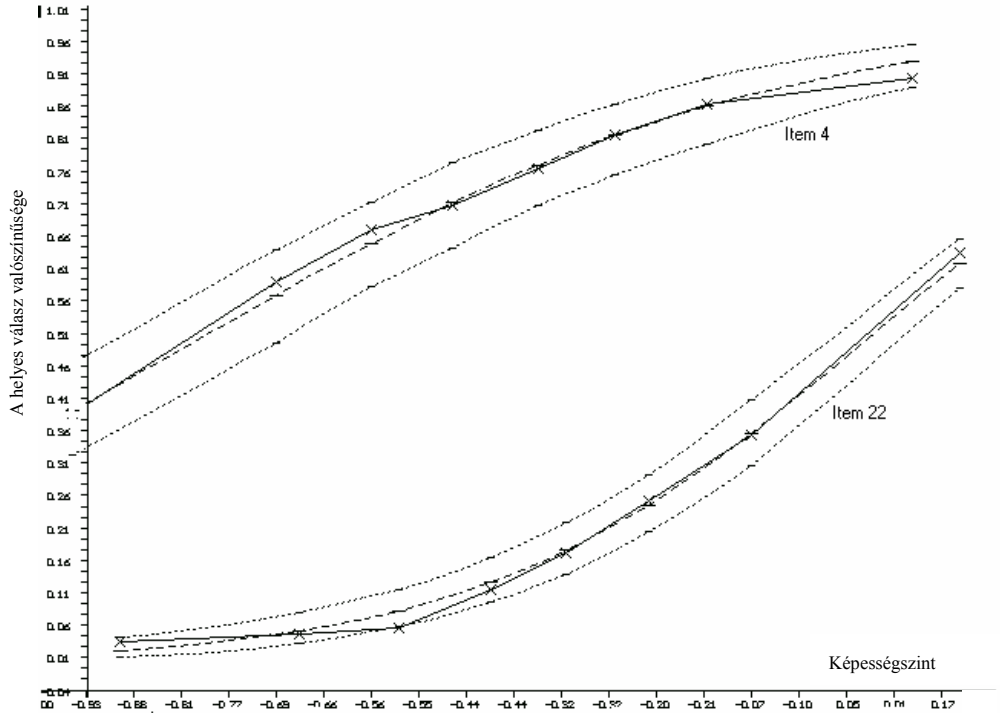
A legmagasabb, 6-os diszkriminációs indexet kapott itemek mindegyike (16. és 19. item) alacsonyabb mutatójú volt, amikor a tesztsorozatot három különálló tesztként kezeltük. Ezen itemek modell-illeszkedése (mint az 5. ábra is mutatja) jó; továbbá élesen elkülöníti egymástól a magasabb és alacsonyabb képességű diákokat.



5. ábra
A 16. item jelleggörbéje
(Annak eldöntése, hogy 6 liter, vagy 66 dl kólát éri meg jobban venni 1080 Ft-ért.)

Az 5-ös diszkriminációs indexű itemek még szintén jó modellilleszkedésűek. Ezen itemek görbéinek lefutása már nem illeszkedik pontosan a modellgörbére, de eltérései a modell hibasávján belül futnak, illetve még kirajzolódik a teljes logisztikus görbe. Ennek következtében végig lehet kísérni, hogy milyen képességszint alatt nő exponenciálisan a megoldás valószínűsége, illetve milyen képességszint után (inflexiós pont) kezd csökkenni a helyes megoldás valószínűségének növekedése.

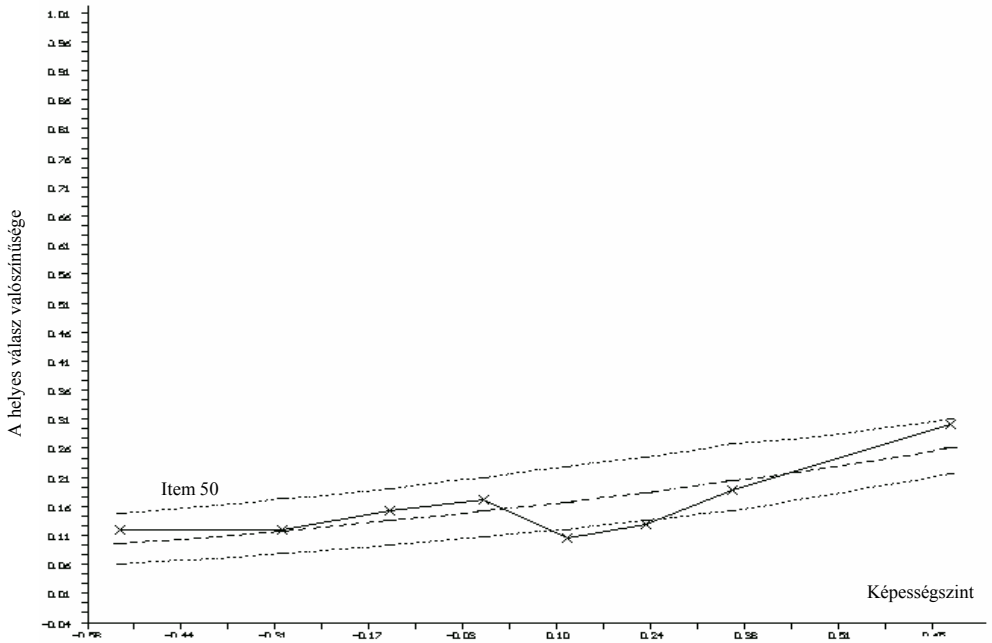
A diákokat a korábbiakhoz képest már kevésbé különítik el a 4-es, illetve 3-as (közepes) diszkrimináns indexű itemek. A könnyebb feladatokat a diákok nagy része – jobb képességűek – meg tudták oldani, ezért még mindig volt differenciáló ereje a képességszint függvényében. A nehéz feladatokat csak a magasabb képességszintű, illetve tudásszintű diákok oldották meg sikeresen. Ezt a két szélsőséges helyzetet szemléltetjük a 6. ábrán, ahol egymásra vetítettük a 4-es (könnyebb) és a 22-es (nehezebb) feladatok jelleggörbéjét. Az alacsony tudásszinttel is megoldható 4-es feladat jelleggörbéje a logisztikus görbe inflexiós pont utáni, felsőbb szakaszára jellemző lefutású, míg a magasabb tudás, illetve képességszinttel megoldható 22-es feladat jelleggörbéje a logisztikus görbe alsóbb, inflexiós pont előtti szakaszához illeszthető.



6. ábra

A 4. item (A leghosszabb utat kellett kiválasztani.), illetve 22. item (28 gombóc fagyiból hány gombóc fagyit evett meg a történetben szereplő gyerek, ha a bátyja már csak fele annyit, apukája kétszer annyit evett meg mint ő) jelleggörbéje

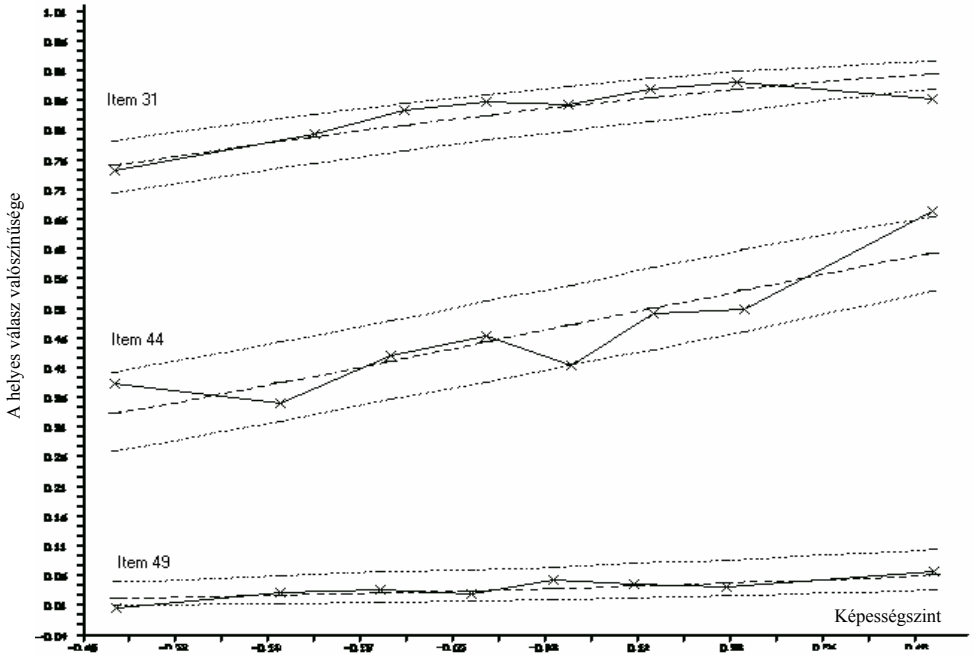
A 2-es diszkriminációs indexet kapott itemek közül az 50. itemet (11.000 méter magasságból ledobva egy kólásüveget hány percig tartana, amíg Földet ér?) a közepes képességű diákok kisebb valószínűséggel oldják meg helyesen, mint a rosszabb, illetve jobb képességűek. Ezen a nehéz feladaton a rosszabb képességűek a vártnál jobban teljesítenek, míg a jobbak az elvárt teljesítményt mutatják, aminek következtében közel azonos valószínűséggel oldják meg ezt az itemet a rosszabb és a jobb képességű diákok (7. ábra).



7. ábra
Az 50. item jelleggörbéje

Az 1-es diszkriminációs indexet kapott itemeknél a program nem tudott képességcsoportokat képezni, ezeket az itemeket csaknem ugyanolyan valószínűséggel oldják meg a gyenge, mint a magas tudásszintű tanulók. Ezeknél az itemeknél mutatkozó jelenségre a klasszikus tesztelméleti elemzések során már az alacsony elkülönítésmutatókkal együtt járó magas, vagy alacsony átlagok, illetve közepes átlagok közepes szórással is utaltak. A lapos karakterisztikus görbéjű itemek vagy nagyon nehezek (49. item), vagy nagyon könnyűek voltak (31. item). A nehéz problémákat rejtő feladatokkal valószínű, hogy még semmilyen formában nem találkoztak a tanulók (10.000 m magasságban a repülőn miért a légkondicionálót és nem a fűtést kapcsolták be, amikor a kinti hőmérséklet –35 fok?), a diákokat képesség szerint nem differenciáló könnyű kérdéseket pedig hétköznapi ismereteik alapján (pl. reklámokból) is meg tudták oldani (Jó-e a pH 5.5 a bőrnek?). A 8. ábrán egymásra vetítettük az említett eseteket, kiegészítve a középső „cikk-

cakkos" lefutású grafikonnal. Utóbbi a találgatással megoldott item karakterisztikus görbéjét szemlélteti, amit okozhatott az előzetes ismeretek hiánya, vagy a feladat nem egyértelmű megfogalmazása.

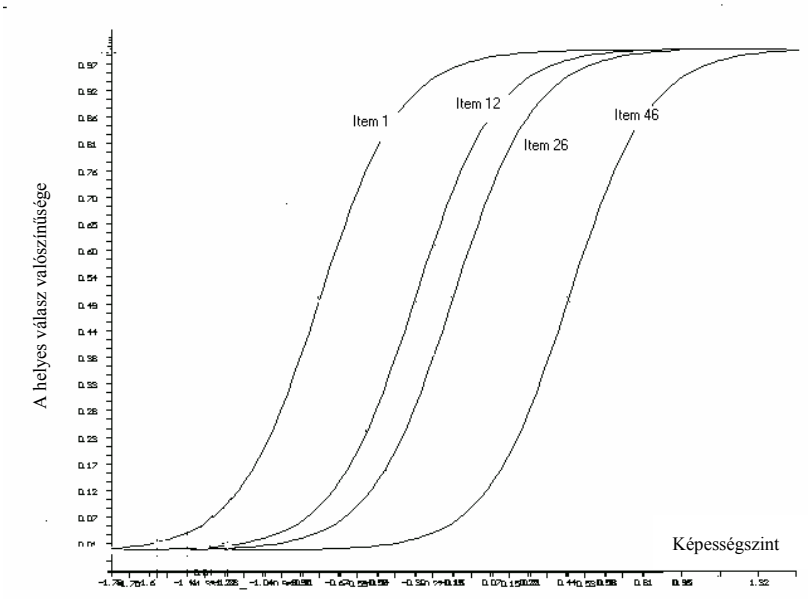


8. ábra
A 31., 44. és 49. item jelleggörbéi

Az itemek modellilleszkedésének diszkriminációs index függvényében történő elemzése után megnézzük, hogyan alakulnak az azonos diszkriminációs indexű itemek egymásra vetített karakterisztikus görbéi. Szemléltetésül az 5-ös diszkriminációs indexű itemek közül kiválasztottunk egy csak első- (1. item), első és második- (12. item), második- és harmadik- (26. item), illetve csak harmadik szinten (46. item) előforduló itemet. A kiválasztott itemeknek a szintek növekedésével egyre nehezebbeknek kell lenniük. Az itemek jelleggörbéjét egymásra vetítve (9. ábra) egymással párhuzamos karakterisztikus görbéket kapunk, ami azt jelenti, hogy az azonos diszkriminációs indexű itemeknél valóban csak az itemek nehézségi fokában van különbség, a többi jellemzőjük megegyezik. Minél inkább pozitív irányba tolódik a görbe, annál nehezebb az adott feladat, mivel annál magasabb képességszint szükséges sikeres megoldásához.

Az itemanalízis során hasonló jelenségeket tapasztaltunk, mint a korábbi mérésekben, elemzésekben. A jól diszkrimináló itemek kevésbé térnek el az iskolában megszokott feladatoktól, a megtanult ismeretek felidézését, alkalmazását kérik. Minél életszerűbb, minél több háttérismeretet igényel egy feladat, minél több zavaró információ áll a diákok rendelkezésére, annál kevésbé sikeresek még a magasabb tudásszintű tanulók is a helyes vá-

lasz megadásában. Az értékelés során azonban nem szabad figyelmen kívül hagyni, hogy nagy valószínűséggel hasonló stílusú feladatlappal még sohasem találkoztak a diákok, ezért az újdonság ereje is meghatározó lehetett a teljesítmények alakulásában.



9. ábra
Azonos diszkriminációs indexű itemek karakterisztikus görbéi

A teljesítmények elemzése

A teljesítmények eloszlása szintenként

A teljesítmények eloszlását a klasszikus tesztelméleti ábrázolásokhoz képest újabb dimenzióban ábrázolja a 10., 11., és 12. (lásd következő fejezet) ábra. (Az elemzéseket a ConQuest programcsomaggal végeztük.) Az ábrák ugyanazon a számegyenesen mutatják a megfelelő szintű feladatlap itemeinek itemnehézségi index szerinti eloszlását és a feladatlapot kitöltő diákok képességszint szerinti eloszlását. Az ábrák bal oldalán látható a diákok, jobb oldalán az itemek képességszint alapján történt elhelyezése – minta-, illetve itemtérképe (*map of persons ability/ item's difficulty map*). A két oldalt összevetve megállapítható, hogy az adott feladatlap nehézsége mennyire felel meg a kijelölt korosztály (minta) komplex problémamegoldó fejlettségi szintjének, illetve útmutatót ad a feladatlapok esetleges továbbfejlesztéséhez: melyik itemet lehetne elhagyni a feladatlapról, mert túl nehéz, vagy túl könnyű, illetve milyen nehézségű feladatokat kellene még tar-

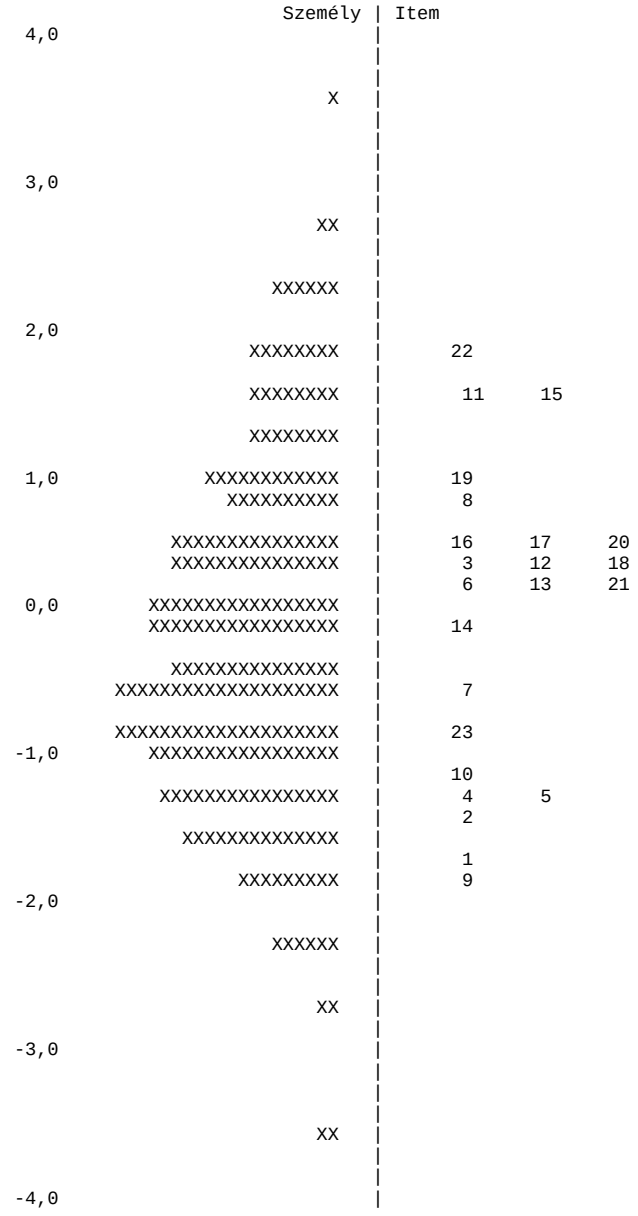
talmaznia a tesztnek, hogy a teszt megoldásához szükséges képességszint-intervallum egybeessen a diákok problémamegoldó képességének fejlettségi szintjével. Általánosságban megfogalmazható, hogy elméletileg, ha egy személy képességparamétere magasabb, mint az item nehézségi indexe, akkor az adott item helyes megválaszolásának valószínűsége több mint 50 százalék, azaz a személy képességparamétere az az itemparaméter lesz, amely itemet az adott személy 50 százalékos valószínűséggel old meg. A számegyenes negatív képességszint-értékei nem negatív szintű képességet jelentenek, hanem átlag alatti képességszintet, mert a program az itemnehézség átlagát automatikusan nullának veszi.

A 10. ábrán a minta képességszint szerinti eloszlásában minden egyes 'x' hét tanulót képvisel, az itemek száma pedig az item nevét (utolsó két számjegyét) jelenti. A minta eloszlása jól közelíti a normál eloszlást, van néhány kiemelkedő és néhány alacsonyabb képességű diák is, akiket a teszt már kevésbé differenciál. A magasabb képességű diákok komplex problémamegoldó képessége magasabb, mint a feladatlapon az itemek 50%-os valószínűséggel történő megoldásához szükséges képességszint, a legalacsonyabb képességűek pedig nem érik el azt a képességszintet, ami a feladatlap problémáinak 50%-os valószínűséggel történő megoldásához szükséges. A feladatlap esetleges továbbfejlesztésének szemszögéből nézve ez annyit jelent, hogy az adott populáció képességszintjének teljes lefedéséhez néhány nehezebb és néhány könnyebb itemmel bővíthető a feladatlap. Összességében, a szignifikanciaszint határain belül elmondható, hogy a harmadik, negyedik és ötödik osztályosok szintjének megfelelő az első szintű feladatlap.

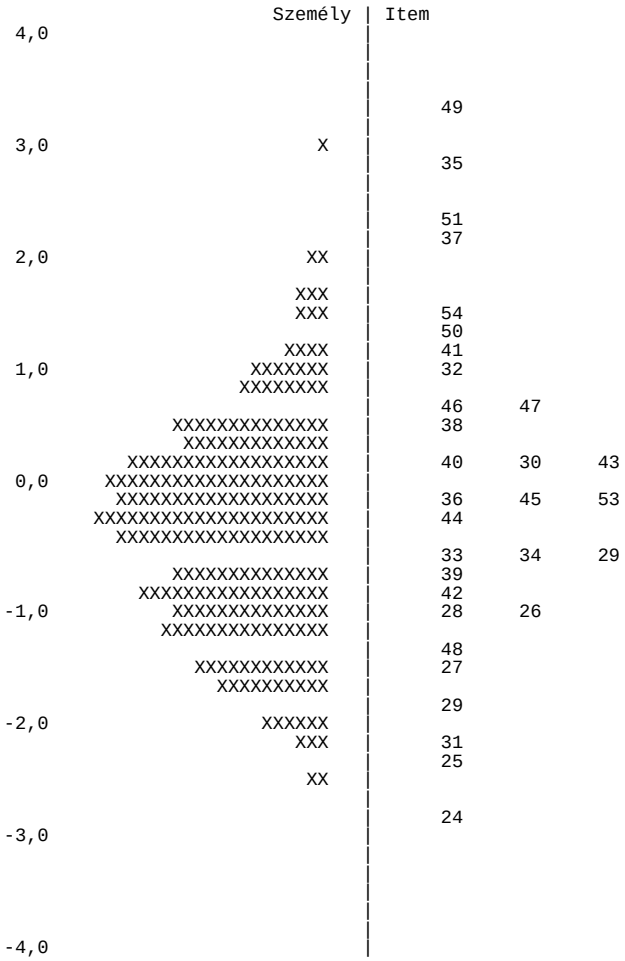
A második szintű feladatlap itemtérképéről és a teljesítmények eloszlásáról hasonló megállapítások tehetőek, ezért ebben a tanulmányban nem ábrázoltuk a feladatlap itemeinek minta- és itemtérképét. A minta eloszlása közelíti a normál eloszlást. A feladatok megoldásához szükséges képességszint a szignifikancia határain belül megegyezik a populáció képességeloszlásával. Egy item (35) logit értéke magasabb, mint az összes diák képességparamétere, ami azt jelenti, hogy annak valószínűsége, hogy ezt az itemet az adott populációban valaki megoldja, kisebb, mint 50 százalék. Összességében az első szinten elmondottakhoz hasonló következtetést vonhatunk le: a második szintű feladatlap az adott populáció képességszintjének megfelelő.

A harmadik szintű komplex problémamegoldó feladatlap itemeinek megoldásához szükséges képességszintet és a diákok komplex problémamegoldó képességszintjét egy egyenesen ábrázolja a 11. ábra. Az ábrán minden egyes 'x' hét tanulót reprezentál. Négy item (49, 35, 51, 37) 50 százalékos valószínűséggel történő megoldásához szükséges magasabb komplex problémamegoldó képességszint, amivel a mintában csak hét tanuló rendelkezik. A 24-es item logit értéke minden személy képességparamétere alatt van, azaz annak valószínűsége, hogy mindenki megoldja ezt a problémát, nagyobb, mint 50 százalék továbbá annak valószínűsége, hogy az átlagos képességűek (logit érték=0) megoldják ezt a problémát közel 100 százalék. (A középiskolás diákok 80 százaléka oldotta meg helyesen ezt a problémát.) A 49-es item logit értéke a képességparaméter-értékek felett van, azaz elméletileg annak a valószínűsége, hogy valaki megoldja ezt a feladatot kisebb 50 százaléknál, sőt már a magasabb képességszintű diákoknál (logit érték=2) is kisebb mint 25 százalék. (A középiskolások 4 százaléka helyesen oldotta meg ezt a problémát.)

A harmadik szintű feladatlap megoldásához szükséges képességszint egybeesik az adott populáció fejlettségi szintjével, jól differenciálja a feladatlap a diákokat.



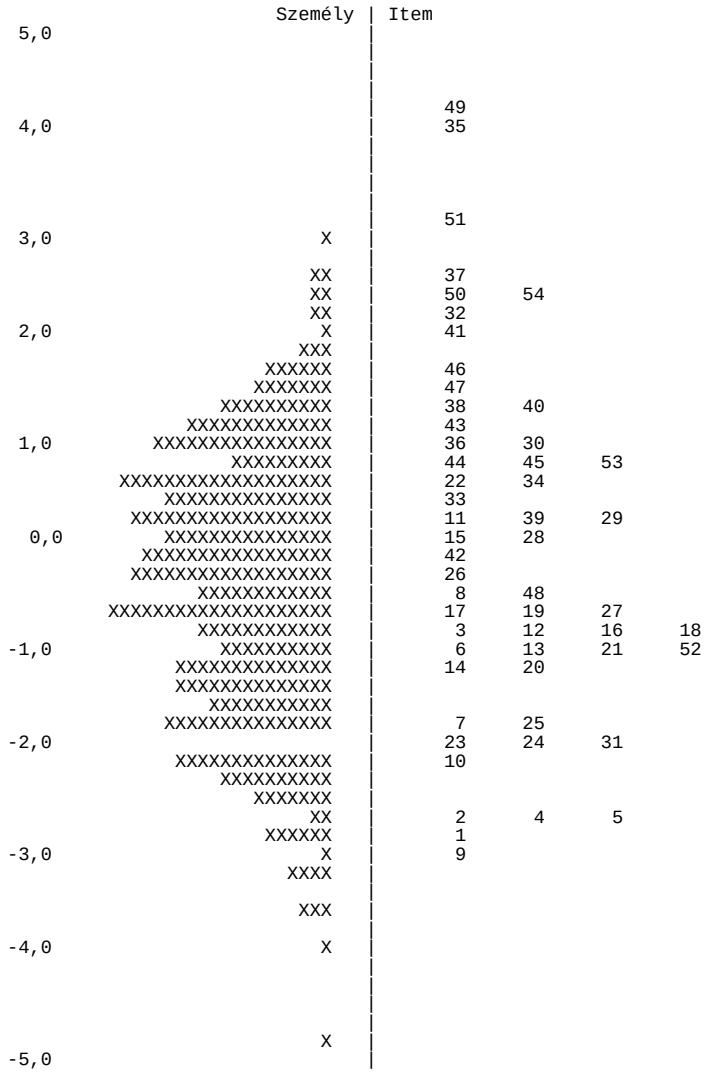
10. ábra
Az első szintű feladatlap minta- és itemtérképe (Minden egyes 'x' hét tanulót képvisel.)



11. ábra
A harmadik szintű feladatlap minta- és itemtérképe (Minden egyes 'x' hét tanulót képvisel.)

A teljesítmények eloszlása egy skálára transzformálva

Amint a korábbi elemzések során, a teljesítmények eloszlásának vizsgálata során is a feladatlapok külön-külön történő elemzése után megnézzük, hogyan viselkedik a három feladatlap, ha egy tesztként kezeljük azokat (12. ábra). Ebben az esetben a program (ConQuest) a felmérésben részt vett 9–17 éves korosztályt egy populációnak kezeli, egy számegyenesen ábrázolja a különböző fejlettségű diákokat (Az ábrán minden 'x' 15 tanulót képvisel.) Az elemzés során a program ugyanerre a számegyenesre transzformálja a három feladatlap itemeit, de most egy tesztként kezelve azokat, kihasználva a második szint összekötő funkcióját.



12. ábra
A három feladatlap egy tesztként elemezve (Minden egyes 'x' tizenöt tanulót képvisel.)

A 12. ábra megerősíti a feladatlaponkénti elemzések eredményét: az itemek e korosztály vizsgálatára összességében megfelelő nehézségűek, sőt még magasabb képességű diákok mérésére is alkalmas – a 49., 35. és 51. item nehézségi indexe magasabb, mint a legmagasabb képességparaméter-érték. A több mint 5000 diákból 135 tanuló komplex problémamegoldó gondolkodásának fejlettségi szintje nem éri el a feladatlapok problémáinak 50 százalékos valószínűséggel történő megoldásához szükséges fejlettségi szintet, e csekély részminta pontosabb differenciálásához a teszt-sorozat továbbfejlesztése

során még néhány könnyebb itemmel ki lehet egészíteni az első szintű feladatsort. A feladatsor itemeinek szintenkénti nehezedését bizonyítja, hogy a képességskála magasabb tartományában magasabb számok (itemnevek) szerepelnek, ami az egyre magasabb szintű feladatlapban való előfordulásra utal. A képességskála alacsonyabb tartományában lévő alacsonyabb számok azt jelzik, hogy ténylegesen az első szintű feladatlap itemei a legkönnyebbek.

13. ábra
A három feladatlap egy tesztként elemezve, külön dimenzióban
(Minden egyes 'x' 23 tanulót képvisel.)

A következő ábra, hasonlóan a 12. ábrához a három szinten nyújtott teljesítményeket és képességszinteket egy skálán ábrázolja, de a 13. ábrán – bár egymással párhuzamos,

összehasonlítható képességskálákon, de – már külön dimenzióban ábrázoljuk a diákok matematikai, illetve természettudományos ismereteinek alkalmazási képességének fejlettségét. Az első dimenzió a matematikai ismeretek alkalmazási képességének dimenziója, a második dimenzió a természettudományos ismeretek alkalmazási képességének dimenziója. Az eredmények alapján a matematikai jellegű problémák megoldásához szükséges képességszint szélesebb skálán mozog, mint a természettudományos ismeretek alkalmazásának képességszintjei. Utóbbi területen homogénebb a diákok teljesítménye, nincsenek kiugróan rossz, se jól teljesítő diákok. Ha továbbfejlesztenénk a feladatlap-sorozatot, akkor a gyengébbek jobb differenciálása érdekében matematikai jellegű problémákkal kellene kiegészíteni azt, továbbá a 49-es természettudományos feladatot el lehetne hagyni, mert szignifikánsan magasabb képességszint szükséges 50 százalékos valószínűséggel történő megoldásához, mint a legmagasabb képességszintű diák képességszintje.

A teljesítmények elemzése, az eredmények egy skálára hozása

A klasszikus tesztelméleti számítások csak arra adnak lehetőséget, hogy a teljesítményeket külön-külön tesztenként nézve hasonlítsuk össze. Ennek az a következménye, hogy az azonos feladatlapokat kitöltők eredményei egymással összehasonlíthatóak, de a más szintűekével nem. Ezért a fejlődés ábrázolásánál nem köthetők össze a különböző szinteken mutatott fejlődési görbék. A modern tesztelméleti számítások az összemérendő tesztek pontértékeinek azonos skálára hozásával lehetővé teszik ennek kiküszöbölését.

Az egyeztetés alapját a különböző tesztekben lévő azonos itemek, átfedések, közös tesztrészletek szolgáltatják. Az anchor itemekkel összekötött, különböző életkorúak által megírt, különböző nehézségű tesztek itemeinek egy skálára hozása, azaz a különböző szintű feladatlapon nyújtott teljesítmények összehasonlíthatósága, a vertikális egyeztetés az IRT egy fontos alkalmazási területévé vált (*Horváth*, 1997). Gyakorlati jelentőssége számottevő, mert ezáltal lehetővé vált a különböző tesztekkel mért azonos tulajdonság összehasonlítása, és a tesztek nehézségi szintjében lévő eltérések kiegyensúlyozása.

Jelen felmérésben az anchor itemeket a második szint itemei adták (1. ábra), a számolásokat elvégeztük mind az OPLM, mind a Quest programcsomaggal is. Az OPLM program az egy skálára hozás folyamatában először a diszkriminációs indexek segítségével súlyozza az egyes itemek értékeit, és meghatározza az itemekhez tartozó még szabad itemparaméterek kiinduló értékét. A következő lépésben az anchor itemek itemparamétereit rögzítve kiszámoltatjuk a többi item paraméterét is, amelyek már a kötött itemekhez viszonyított mutatószámok. A komplex feladatlapsorozat 54 itemének paraméterét az 1. táblázat mutatja. A táblázatban minden item mellett két paraméterérték szerepel. Az egyiket az OPLM, a másikat a Rasch modellel dolgozó Quest programmal számoltuk ki. A paraméterértékek közötti különbség oka a két modell között fennálló eltérésben rejlik. Ha grafikonon ábrázolnánk az értékeket, a Rasch modell alapján számolt függvény egy nyújtott transzformáltja az OPLM által adott paraméterértékekből álló függvénynek. Előbbi az itemparaméter-értékeket szélesebb skálán helyezi el. Korábban már utaltunk rá, hogy a feladatlapok jó modellilleszkedését mutatja az itemparaméterek nagyságának alakulása. A szintek előrehaladtával egyre magasabb indexekkel találkozhatunk, a tesz-

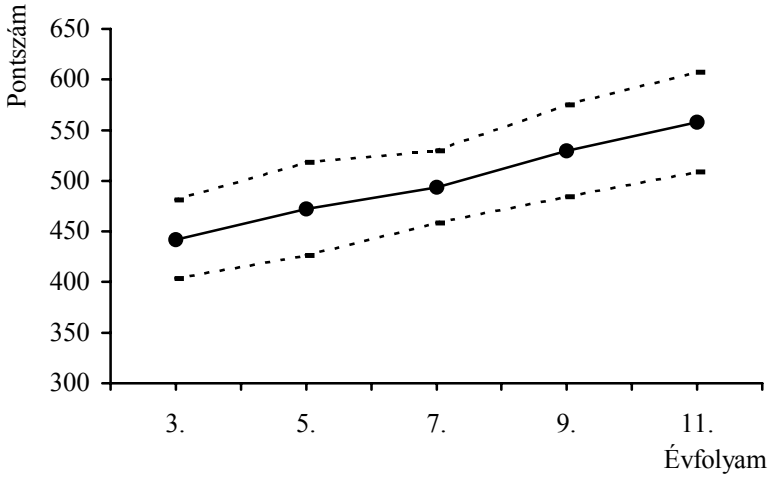
ten elért eredmények fényében egyre nagyobb jelentőségűekké válnak a magasabb szintű feladatok, egyre magasabb képességszint szükséges a megoldásukhoz.

A súlyozott itemek segítségével minden egyes személyre meghatározható az egyes feladatlapokon elért súlyozott összpontszám. Ez a mutató azt jelzi, hogy ha egy adott képességparaméterű személyt ugyanazon feltételek között sokszor tesztelünk, akkor milyen szinten fog teljesíteni. Erre a helyi keretek miatt részletesen nem térünk ki.

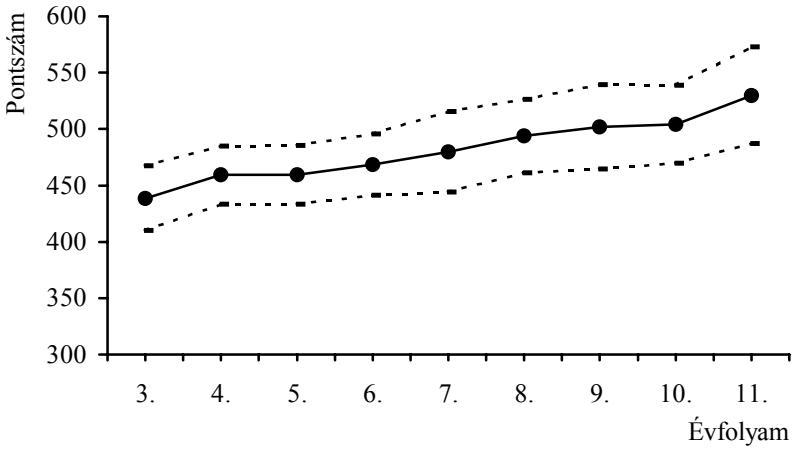
2. táblázat. A komplex problémamegoldó feladatlap-sorozat itemeinek itemparaméterei

Item	*Itemparaméter*		*Item*	*Itemparaméter*		*Item*	*Itemparaméter*	
	OPLM	*Quest*		*OPLM*	*Quest*		*OPLM*	*Quest*
K1101	-0,79	-2,89	K1220	-0,42	-1,20	K2329	0,02	0,29
K1102	-0,74	-2,65	K1221	-0,39	-1,08	K2328	-0,06	0,09
K1103	-0,37	-0,89	K1209	-1,23	-2,97	K2326	-0,15	-0,25
K1104	-0,83	-2,58	K1222	0,03	0,60	K2327	-0,26	-0,69
K1105	-0,75	-2,55	K1223	-0,77	-1,90	K3341	1,46	2,11
K1106	-0,35	-1,01	K2324	-0,74	-1,96	K3342	-0,11	-0,02
K1107	-0,55	-1,74	K2325	-0,65	-1,69	K3343	0,34	1,25
K1108	-0,19	-0,48	K2333	0,04	0,42	K3344	0,25	0,75
K1111	0,24	0,26	K2334	0,10	0,57	K3345	0,30	0,81
K1112	-0,38	-0,90	K2335	1,29	4,05	K3346	0,44	1,71
K1116	-0,36	-0,77	K2336	0,31	0,94	K3347	1,03	1,63
K1210	-0,83	-2,04	K2337	0,66	2,64	K3348	-0,42	-0,53
K1213	-0,39	-1,08	K2338	1,02	1,42	K3349	1,89	4,26
K1214	-0,41	-1,16	K2339	0,02	0,34	K3350	1,72	2,40
K1215	-0,12	0,11	K2340	0,48	1,30	K3351	1,40	3,27
K1217	-0,30	-0,66	K2330	0,33	0,98	K3352	-0,45	-1,02
K1218	-0,31	-0,77	K2331	-1,73	-1,85	K3353	0,25	0,89
K1219	-0,28	-0,56	K2332	0,67	2,27	K3354	1,04	2,52

Az itemparaméterek egy skálára hozása után a következő lépcsőfok a személyparaméterek meghatározása volt, ami minden egyes személyhez hozzárendel egy képességszintet. Ezek évfolyamonkénti átlagát lineáris transzformációval eltoljuk úgy, hogy az átlag a nemzetközi mérésekben is használatos 500 pont körül ingadozzon. Ezzel összehasonlíthatóvá vált a különböző szintű feladatlapokat kitöltő diákok teljesítménye. A komplex problémamegoldó képesség fejlődésének mértékét a 14. ábrán, az explicit matematika feladatlapon nyújtott teljesítmények alakulását a 15. ábrán, az explicit természettudományos feladatokon elért eredmények alapján számolt képességszintek alakulását a 16. ábrán ábrázoljuk. Mindegyik ábrán feltűntettük a szórás mértékét is.



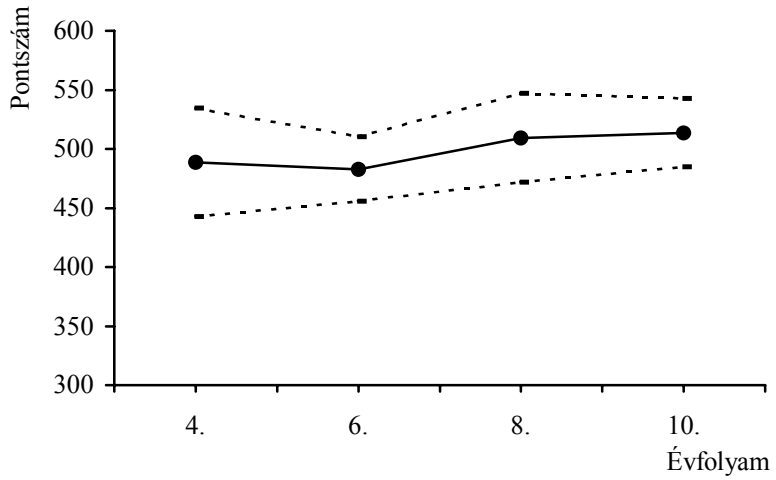
14. ábra
A komplex problémamegoldó képesség fejlettségi szintjei a különböző évfolyamokon



15. ábra
A matematika teszten mutatott teljesítmények évfolyamonkénti bontásban

Az eddigi elemzésekhez képest új információt az ötödikes és hatodikos, illetve a nyolcadikos és kilencedikesek közötti fejlődés kimutatása jelenti, továbbá a szintek egymással való összehasonlíthatósága. A komplex problémamegoldást életszerű helyzetekben vizsgáló feladatokon egyre magasabb eredményeket érnek el a diákok, problémamegoldó képességük egyre fejlettebbé válik. Még a leggyengébb tizenegyedikes évfo-

lyamosok is elérik a nyolcadik osztályosok átlagos szintjét. A matematika területén lehetünk a legintenzívebb fejlődésnek tanúi, bár itt a szórás mértéke is nagyobb, mint a komplex problémamegoldás esetében. A természettudományos feladatokon negyedik évfolyamon van olyan diák, akinek tudásszintje a középiskolásokéval vetekszik. Ez valószínű a média, az Internet, valamint a gyerek ismeretterjesztő könyvek hatása, ahonnan a diákok szinte korlát nélkül jutnak hozzá a legkülönfélébb természettudományos ismeretekhez. Összességében ezen a területen tapasztaljuk a legkisebb fejlődést, az átlagos fejlődés mértéke kisebb, mint a szórások nagysága.



16. ábra
A természettudományos teszt eredményei évfolyamonkénti bontásban

Az elemzés alapján megfogalmazható következtetések

A klasszikus tesztelméleti számítások csak az azonos tesztet megoldók eredményeinek összehasonlítására adnak lehetőséget, ezért a fejlődés ábrázolásánál nem köthetők össze a különböző szinteken mutatott fejlődési görbék, például jelen esetben nem hasonlíthatóak össze az első és második szinten teljesítő diákok képességszintjei. Ezzel szemben a modern tesztelméleti számítások lehetővé teszik horgony itemek felhasználásával a különböző szintű feladatlapon elért eredmények azonos skálára hozását. Ennek gyakorlati jelentőssége számottevő, mert ezáltal lehetővé válik a különböző tesztekkel mért azonos tulajdonság összehasonlítása, és a tesztek nehézségi szintjében lévő eltérések kiegyensúlyozása. A valószínűségi tesztelemzés egy másik előnye, hogy közvetlenül összehasonlíthatóvá válik az item nehézsége és a diákok képességi szintje. A komplex problémamegoldó feladatlap-sorozat tekintetében a Rasch modell segítségével történő elemzés eredménye azt mutatja, hogy a feladatlap-sorozat problémái az érintett korosztály képességszintjének megfelelőek. A komplex problémamegoldást életszerű helyzetekben vizsgáló feladatokon egyre magasabb eredményeket érnek el a diákok, probléma-

megoldó képességük egyre fejlettebbé válik. A matematika területén lehetünk a legintenzívebb fejlődésnek tanúi. A 9–17 éves diákok matematikai természetű problémamegoldó képessége szélesebb skálán mozog, több kiugróan magas és több kiugróan alacsony képességszintű diák van, mint a természettudományos ismeretek alkalmazásának területén, ahol az átlagos fejlődés mértéke kisebb, mint a szórások nagysága. Ettől eltekintve negyedik évfolyamon van néhány olyan diák, akinek természettudományos tudásszintje a középiskolásokéval vetekszik. Ez valószínű a média, az Internet, valamint az egyre nagyobb számban megjelenő színes, képes ismeretterjesztő könyvek hatása lehet, ahonnan a diákok szinte korlát nélkül jutnak hozzá a legkülönfélébb természettudományos ismeretekhez.

A tanulmányban bemutatott vizsgálat a T 030555 számú OTKA kutatási program, illetve a SZTE-MTA Képességkutató Csoport keretében készült.

Irodalom

Adams, R. J., Wilson, M. R. és Wang, W. C. (1997): The multidimensional random coefficients multinominal logit. *Applied Psychological Measurement*, **21**. 1–24.

Andricht, D. A. (1978): A rating formulation for ordered response categories. *Psychometrika*, **43**. 561–573.

Bond, T. és Fox, C. M. (2001): *Applying The Rasch Model. Fundamental Measurement in the Human Sciences*. Lawrence Erlbaum Associates, Publishers, Hillsdale, New Jersey.

Csapó Benő (2000): Tudásszintmérő tesztek. In: Falus Iván (szerk.): *Bevezetés a pedagógiai kutatás módszereibe*. Műszaki Tankönyvkiadó, Budapest. 277–316.

Fischer, G.H. (1983): Logistic latent trait models with linear constraints. *Psychometrika*, **48**. 3–26.

Horváth György (1997): *A modern tesztmodellek alkalmazása*. Akadémiai Kiadó, Budapest.

Linacre, J. M. (1994): *Many-faced Rasch Measurement*. MESA press, Chicago.

Linacre, J. M. (2000): Comparing "Partial Credit" and "Rating Scale" Models. *Rasch Measurement Transactions*, **14**. 3. sz. (http://www.rasch.org/rmt/rmt143k.htm, 2004. február 21.)

Linden, W. V. D. és Hambleton, R. K. (1997, szerk.): *Handbook of Modern Item Response Theory*. Springer Verlag. (http://www.assess.com/Books/b-46616.htm, 2004. február 21.)

Masters, G. N. (1982): A Rasch model for partial credit scoring. *Psychometrika*, **47**. 149–174.

Molnár Gyöngyvér (2003): A komplex problémamegoldó képesség fejlettségét jelző tényezők. *Magyar Pedagógia*, **103**. 1. sz. 81–102.

Rasch, G. (1980): Probabilistic models for some intelligence and attainment tests. University of Chicago Press, Chicago, Idézi: Bond, T. és Fox, C. M. (2001): *Applying The Rasch Model. Fundamental Measurement in the Human Sciences*. Lawrence Erlbaum Associates, Publishers, Hillsdale, New Jersey.

Verhelst, N. D., Glas, C. A. W. és Verstralen, H. H. F. M. (1995): *One–Parameter Logistic Model OPLM*. CITO, Arnhem.

Wilson, M. R. (1992): The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement*, **16**. 309–325.

Wu, M., Adams, R. J. és Wilson, M. R. (1998): *ACER ConQuest. Generalised item Response Modelling Software*. ACER Press, Australia.

Molnár Gyöngyvér

ABSTRACT

GYÖNGYVÉR MOLNÁR: ASSESSMENT OF KNOWLEDGE APPLICATION WITH IRT

This paper reports the results of an examination of the relationship between two ways of assessing students' knowledge. More than 5000 Hungarian students (9 to 17-year-olds) were assessed in 2002 regarding their performance on reading, mathematics literacy and science tests as well as their application of the same knowledge in complex problem solving tasks. The test included multiple-choice, short answer, and extended response items. Two IRT programs were used to analyse the results, OPLM (*Verhelst*, *Glas* and *Verstralen*, 1995) and ConQuest (*Wu*, *Adams* and *Wilson*, 1997). This paper compares the results and the benefits of these alternatives. The model applied to the survey is a generalised form of the Rasch model. This is a mixed coefficients model where items are described by a fixed set of unknown parameters, while student outcome levels (the latent variable) are random effects. For each item parameter, the ConQuest fit mean square statistic index provided an indication of the compatibility between the model and the data. For each student, the model describes the probability of obtaining different item scores. Figures are included to show the distribution of Rasch-estimated item difficulties. The student achievement distribution is located parallel to the item difficulty distribution. This implies that, on average, the students in the study had an ability level that was adequate for a 50 percent chance of solving an average item correctly. The accumulation of comparisons across cases yields an item-fit statistic. Each of the domains was scaled separately to examine the targeting of the tests. Trend indicators show how results change over time. The outcomes draw a profile of useful knowledge and skills among 9- to 17-year-olds.

Magyar Pedagógia, **103**. Number 4. 423–446. (2003)

Levelezési cím / Address for correspondence: Molnár Gyöngyvér, Szegedi Tudományegyetem, Pedagógiai Tanszék, MTA Képességkutató Csoport, H–6722 Szeged, Petőfi S. sgt. 30–34.